THE 10TH ANNUAL

DOE Joint Genome Institute

Genomics of Energy & Environment Meeting

March 24 - 26, 2015 • Walnut Creek, California

Abstracts

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Tenth Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy

Office of Science

March 24-26, 2015

Walnut Creek Marriott

Walnut Creek, California

Contents

Table of Contents

Speaker Presentations

Abstracts alphabetical by speaker

Enabling recombinatorial fungal metabolic engineering via genome editing and DNA synthesis

William G. Alexander* (walexander@wisc.edu)[1,2], Drew T. Doering[2], Nikolay Rovinskiy[1,2], Mariana Lopes[2,3], David Peris[1,2], Jin Kang[1,2], Angela Tarver[4], Miranda Harmon-Smith[4], Jan-Fang Cheng[4], Samuel Deutsch[4], Jeff S. Piotrowski[1], Trey K. Sato[1], Audrey P. Gasch[1,2], Chris Todd Hittinger[1,2]

[1]DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI; [2]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI; [3]Universidade Federal de Minas Gerais, Belo Horizonte, Brasil; [4]Joint Genome Institute, Walnut Creek, CA.

Fungal organisms, in particular *Saccharomyces cerevisiae*, have been engineered previously to produce valuable fuel and commodity compounds. For example, metabolic networks enabling use of novel carbon sources, such as xylose, have also been installed to varying degrees of success. Here we describe the use of the HERP1.x series of *Saccharomyces* genome editing cassettes, which are available to the public for non-commercial use. The HERP cassettes are composed of a novel positively and negatively selectable marker, *TkMX* (derived from human herpes simplex virus thymidine kinase), an inducible double-strand break generator in the form of a meganuclease and its cut site, and multiple secondary markers. These cassettes allow for manipulation of a genomic locus at rates approaching 1% of surviving cells, or approximately 1000x more efficiently than currently reported CRISPR/Cas9 rates in wild *Saccharomyces* species. Improvements on these cassettes that allow for action in other fungal organisms are outlined. Finally, we illustrate a collaboration with the Joint Genome Institute's Synthetic Biology program to produce a library of parts, mostly derived from diverse yeast species requiring genome-wide alterations to their codons, to expedite recombinatorial engineering for the production of biofuels from lignocellulosic biomass. These parts are designed to be screened on plasmids, assembled onto plasmids in user-specified or random combinations, and stably integrated into the *Saccharomyces* genome via HERP technology.

Resequencing of Switchgrass, an Outcrossing Bioenergy Crop Species

Laura E. Bartley* (lbartley@ou.edu)[1], G. Albert Wu[2], Jeremy Schmutz[3,2], Daniel Rokhsar[2], Yanqi Wu[4], Malay C. Saha[5], Thomas E. Juenger[6], David Lowry[6,7], Sandra Thibiviliers[1], Kerry Berry[2], Christian Tobias[8], Michael D. Casler[9], Katrien Devos[10], Carol Auer[11], Zengyu Wang[5]

[1] University of Oklahoma, Norman, OK, 73019; [2] DOE-Joint Genome Institute, Walnut Creek, CA 94598; [3] HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806; [4] Oklahoma State University, Stillwater, OK 74078; [5] Samuel Roberts Noble Foundation, Ardmore, OK 73401; [6] University of Texas, Austin, TX 78712; [7] Michigan State University, MI 48824; [8] USDA-Agricultural Research Service, Albany, CA 95125; [9] USDA-Agricultural Research Service, Madison, WI 53706; [10] University of Georgia, Athens, GA 30602; [11] University of Connecticut, Storrs, Connecticut 06269

Switchgrass (*Panicum virgatum* L.) is a large-stature, stress-tolerant North American Tallgrass Prairie species being developed as a bioenergy crop. Switchgrass is primarily outcrossing and consists of two

ecotypes, uplands and lowlands. The main objectives of this project are to produce 1) sequence of reduced diversity genotypes to support the switchgrass genome project and 2) a high-quality, uniform variant catalog to support switchgrass improvement efforts. In this project, 48 genotypes of switchgrass have been resequenced with whole genome shotgun sequencing, with 150 bp paired-end reads and an average of 60 Gbp generated per genotype. Lowlands are primarily allotetraploid, behaving essentially as diploids; whereas, uplands consist of both tetraploid and octaploid individuals. Resequenced genotypes consist of lowland tetraploids (33), uplands (11), including octaploids (4/11), and lowland-upland hybrids (4). Two of the lowlands possess upland chloroplast sequences, consistent with hybridization across the ecotypes *in situ*. Sequencing confirmed that three genotypes have greatly reduced heterozygosity relative to the reference, including a dihaploid genotype, and two third-generation selfed genotypes. These sequences are being used to improve the reference genome assembly. Thirty of the resequenced genotypes are being used in genetic mapping or breeding programs. Preliminary analysis has found that many of these genotypes possess genomic regions that are largely homozygous. This suggests the possibility of improving diversity in breeding programs by incorporating genotypes with variation in these regions. The whole genome resequencing data will allow us to provide a deep characterization of the genetic differentiation between switchgrass ecotypes and subpopulations within each ecotype. Data generated in this project will be used to improve switchgrass as a bioenergy species.

Do fungi have a volatome?

Joan W. Bennett* (profmycogirl@yahoo.com)

Department of Plant Biology and Pathology, Rutgers University, New Brunswick, NJ 08901

Abstract: The ability to measure metabolites in any given biological system and to determine their relationship to the genome, transcriptome and proteome of that system has become a kind of holy grail in contemporary biology. Volatile organic compounds (VOCs) are low molecular weight compounds that can vaporize and enter the gas phase at normal atmospheric temperatures and pressure. Fungi and other microbes emit a variety of VOCs that occur in chemically complex mixtures of acids, alcohols, aldehydes, aromatics, esters, ketones, terpenes, thiols and other compounds, many of which have distinctive odors. Several laboratories have shown that the relative quantity of different VOCs produced by molds is dependent on substrate, temperature and other growth parameters. My laboratory has been studying the physiological properties of fungal VOCs and shown that the most common fungal VOC, which is known as “mushroom alcohol” (1-octen-3-ol), induces apoptotic, inflammatory and neurotoxic effects in a *Drosophila melanogaster* model. When tested in a yeast knock out library, genes disrupted in vacuolar/vesicular transport or endosomal transport were highly prevalent in providing resistance to 1-octen-3-ol. Specifically, when either the retrograde (retromer and GARP complexes) or anterograde (ESCRT) sorting pathways are disrupted, yeast cells are resistant to 1-octen-3-ol. The term “volatome” has been used to describe the suite of VOCs produced by the pathogenic species *Aspergillus fumigatus*. Since fungi are excellent biodegraders, capable of producing many exoenzymes that break down numerous substrates, it is important to remember that not all the VOCs in a given mixture found from a growing mold are biosynthetic products determined by the fungal genome. This talk will review molecular, physiological and linguistic aspects of research on fungal VOCs and discuss their possible toxicological importance in indoor environments.

Protein regulatory networks

Steven Briggs* (sbriggs@ucsd.edu)[1], Justin Walley[2], Ryan Sartor[1], Kevin Wu[1], Zhouxin Shen[1], Mark Urich[3], Robert Schmitz[4], and Joseph Ecker[3]

[1]UC San Diego, [2]Iowa State University, [3]Salk Institute, [4]University of Georgia

Integrated molecular atlases make possible systems biology approaches aimed at understanding biological phenomena. Using RNA-seq and quantitative mass spectrometry we generated an atlas comprised of 41,272 transcripts, 18,646 proteins and 32,000 phosphopeptides, quantified across maize development. Analysis of the atlas has revealed complex spatiotemporal patterns of gene activity. For example, there is poor correlation between protein and mRNA levels and for many of the most abundant proteins there is little to no detectable cognate mRNA. The atlas has also enabled generation and interrogation of a number of different types of regulatory networks including mRNA and protein co-expression networks, a gene regulatory network (GRN) and a novel predicative kinase-substrate network. Together, these studies highlight the complex interplay of transcriptional, translational and post-translational events in dynamically remodeling the proteome.

Sequence-enabled gene discovery in *Setaria viridis*, a model system for the Panicoid grasses

Thomas P. Brutnell* (TBrutnell@danforthcenter.org)[1], Hui Jiang[1], Pu Huang[1], Jeremy Schmutz[2], Kerrie Barry[2], Anna Lipzen[2], Xiaoping Li[1], Zhonghui Wang[1]

[1]Donald Danforth Plant Science Center, St. Louis, MO; [2]US Department of Energy Joint Genome Institute, Walnut Creek, California.

The panicoid grasses are among the world's most productive plants and include major food, feed and bioenergy feedstocks such as maize, sorghum, miscanthus, switchgrass and sugarcane. These large, often polyploidy, and long-lived plants produce the majority of grain, sugar and biomass supporting our current bioenergy economy, but these same attributes make them challenging systems for genetic analysis. *Setaria viridis* is one of the smallest and rapid cycling diploid grasses in the panicoideae, making it an attractive system for genetic analysis. To develop the resources necessary to accelerate the broad adoption of *S. viridis* as a genetic model for bioenergy research, we generated mutant populations for forward and reverse genetics. We are in the process of creating 20,000 indexed NMU-mutagenized families and to date have generated and performed initial characterizations of approximately 3000 M3 mutant families. Several mutants were identified that likely carry defects in a range of developmental and metabolic pathways including primary carbon metabolism, hormone signaling and flowering. A total of 55 mutant individuals with interesting phenotypes are being sequenced at ~30x coverage by JGI-DOE to empirically determine mutation frequency and as a foundation for fine mapping. Initial sequence characterizations indicate the average number of non-synonymous disruptive mutations is 50 per individual. In addition, a panicle mutant has been crossed with *S. viridis* accessions to fine map the causal gene using Bulk Segregant Analysis (BSA) followed by deep sequencing. We are also developing computational pipelines and wet lab protocols to perform TILLING-by-sequencing for gene-centric reverse genetic screens of this NMU-mutagenized population. I will also discus current progress in developing populations to facilitate GWAS and fine mapping studies. This includes assembling a diverse germplasm collection of 430 *S. viridis* accessions in collaboration with several colleagues and have begun the deep sequencing of these accessions in partnership with DOE-JGI. We demonstrate that phenotypes of interest can be mapped to fine resolution with a thorough characterization of the standing

phenotypic variation in a subset of this panel. To date, 60 lines have been characterized for phenotypic traits and have been propagated for seed distribution at the USDA GRIN (http://www.ars-grin.gov/cgi-bin/npgs/html/taxon.pl?430573).

Production Strains via Systems Biology

Mark J. Burk* (mburk@genomatica.com)

EVP and CTO, Genomatica

In this presentation, we will describe a tightly integrated biotechnology platform that enables precision engineering of microorganisms for high performance production of a range of molecules ranging from simple non-natural basic and intermediate chemicals to more complex entities. Our ability to design and surgically manipulate organisms, combined with a data-rich systems approach to diagnosing metabolic and physiological bottlenecks, has provided innovative solutions that can be applied to the production of many valuable products across the chemical and fuel markets and beyond. We will demonstrate how we have leveraged our platform to optimize organisms and processes, de-risking projects from an early stage, and driving down costs and timelines for full development of novel biological processes. Case studies will include Genomatica's processes for 1,4-butanediol, the first successful commercial bioprocess for a commodity intermediate chemical, and bio-based butadiene, as well as examples showing how more complex products can be produced.

Translating a trillion points of data into therapies, diagnostics, and new insights into disease

Atul Butte* (abutte@stanford.edu)

Division of Systems Medicine, Stanford University and Lucile Packard Children's Hospital

With the end of the United States NIH budget doubling and completion of the Human Genome Project, there is a need to translate genome-era discoveries into clinical utility. The difficulties in making bench-to-bedside translations have been well described. The nascent field of translational bioinformatics may help. Dr. Butte’s lab at Stanford builds and applies tools that convert trillions of points of molecular, clinical, and epidemiological data -- measured and often released to the public by researchers and clinicians over the past decade and now colloquially termed “big data” -- into diagnostics, therapeutics, and new insights into disease. Dr. Butte, a bioinformatician and pediatric endocrinologist, will highlight his lab’s work on using publicly-available molecular measurements to find new uses for drugs and evaluating patients presenting with whole genomes sequenced.

Antarctic Microbial Omics – a Path to Learning About Life in the Cold

Ricardo Cavicchioli (r.cavicchioli@unsw.edu.au)

School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, NSW 2052, Australia

Low temperature is a critical environmental factor controlling the evolution and biodiversity of life on Earth. Most (~85%) of the Earth’s biosphere is permanently below 5°C, dominated by ocean depths,

glacier, alpine and polar regions. Antarctica is arguably the world's most important continent for influencing the Earth's climate and global ocean ecosystem. For example, 90% of the Earth's ice is present in Antarctica, and by maintaining freezing temperatures, cold water masses sink near the Antarctic continent and drive thermohaline circulation of global ocean currents. The Southern Ocean accounts for ~ 30% of global ocean uptake of CO_2 despite representing ~ 10% of the total surface area of the global ocean.

In the Vestfold Hills (68° 33' S, 78° 15' E) more than 3000 water bodies from freshwater (<0.1%) to hypersaline (32%) exist, many of which are marine-derived having been separated from the marine environment ~3–7,000 years ago. As a result they provide unique opportunities for studying the evolution of marine microbial populations.

In 1995 I started studying cold adapted archaea (methanogens), and in 2006 commenced an environmental genomics and proteomics program to evaluate whole lake and Southern Ocean ecosystems. One of the real values we have realised from applying shotgun metagenomics and metaproteomics to the Antarctic microbial communities is letting the data empirically describe the system, thereby discovering what is present and important – discoveries which do not necessarily match preconceived expectations. For example, Ace Lake and Organic Lake revealed unexpected roles for viruses (*e.g.* virophages and phage resistant bacteria) and dominant cellular species (*e.g.* green sulfur bacteria), and Deep Lake an extraordinary level of inter-genera gene exchange. As a result we learned about unanticipated and important roles for specific components of the microbial loop, and the apparent 'peculiarities' of polar ecosystems.

In this talk I will reflect on our path from studies of model laboratory strains to field-based environmental omics, highlight our work with JGI which spans back to about 2002, and provide some detail on haloarchaea and virus-host interactions in frigid (down to -20°C), hypersaline, Deep Lake. I'll also provide a perspective on our 2015 CSP which is based on a 17 month Antarctic expedition with samples taken November 2013 to February 2015: Seasonal variation in Antarctic microbial communities: ecology, stability and susceptibility to ecosystem change.

Designing New Biosynthetic Pathways for Biofuel Production

Michelle Chang* (mcchang@berkeley.edu)

University of California, Berkeley

Living systems have evolved the capacity to carry out many chemical transformations of interest to synthetic chemistry if they could be redesigned for targeted purposes. However, our ability to mix and match enzymes to construct *de novo* pathways for the cellular production of small molecule targets is limited by insufficient understanding how chemistry works inside a living cell. Our group is interested in using synthetic biology as a platform to study how enzymes function *in vivo* and to use this understanding to build new synthetic pathways for the production of pharmaceuticals, nanomaterials, and fuels using living cells.

Microbial communities regulate the cycling of energy and matter in the marine environment, yet how they respond to environmental change, and the variability of their activities in space and time, are not well understood

Ed DeLong* (edelong@hawaii.edu)

Univeristy of Hawaii

One of the larger challenges remaining is to define taxa, gene and process distributions on appropriate spatial and temporal scales in the ocean. How much does metabolic activity of an specific planktonic microbial population vary of the course of minutes, hours, days and weeks ? Over what spatial scales ? Put another way, exactly what is “a day in the life” of wild planktonic microbial species ? How does the that variation in any specific population correlate with the variability of other co-occurring taxa or populations, and corresponding environmental variation ? New robotic sampling strategies, coupled with genome-wide gene expression analyses in wild planktonic microbial populations, have potential to answer some of these questions. New results using such approaches show that individual populations, as well as very different bacterial and archaeal species, display remarkably similar, time-variable patterns of synchronous gene expression over extended periods of time. These new results suggest that specific environmental cues may elicit cross-species coordination of gene expression among diverse microbial groups, that potentially enable multispecies coupling of metabolic activity. How the temporal compartmentalization of metabolism across different across different species in complex populations is an exciting area to explore. The coupling of genome-enabled data with observational oceanography promises to significantly advance our understanding of the inner workings of complex ecosystems.

Kryptonia: A new bacterial candidate phylum discovered through global metagenomic surveys

Emiley A. Eloe-Fadrosh* (eaeloefadrosh@lbl.gov)[1], Jessica Jarett[1], David Paez-Espino[1], Peter F. Dunfield[3], Brian P. Hedlund[2], Danielle Goudeau[1], Rex Malmstrom[1], Amrita Pati[1], Edward M. Rubin[1], Tanja Woyke[1], Nikos C. Kyrpides[1], Natalia N. Ivanova[1]

[1]Joint Genome Institute, Walnut Creek, CA 94598, USA; [2]University of Nevada, Las Vegas, Las Vegas, NV 89154, USA; [3]University of Calgary, Calgary, AB T2N 1N4, Canada.

Analysis of the increasing wealth of metagenomic data collected from diverse environments offers new strategies for the discovery of novel branches on the tree of life. Interrogation of 4,290 samples collected globally comprising 5.2 Tb of metagenomic data led to the discovery of a new bacterial candidate phylum, Kryptonia. This phylum was found exclusively in high temperature, neutral pH geothermal springs in North America and Asia. Our metagenomic survey enabled the detection of this phylum previously missed due to mismatches between its 16S rRNA sequence and primers typically used to amplify this marker gene. Genome reconstruction from metagenomic data combined with single-cell genome sequencing of cells from environmental samples yielded twenty-two high-quality draft genomes, representing four different Kryptonia genera. Metabolic predictions indicate a heterotrophic, motile lifestyle with the putative capability to respire iron and the inability to synthesize most amino acids, along with the capacity to degrade a suite of polysaccharides as well as aromatic compounds at high temperature. Kryptonia genomes harbor a CRISPR-Cas system indicative of defense against viral attack, and we found evidence for a unique hybrid system type that expands the known diversity of

CRISPR-Cas loci. The discovery of this new bacterial candidate phylum with extremely restricted ecological distribution highlights the potential diversity of microbial life still awaiting discovery.

How microbial evolution influences environmental ecosystem services

Jack A. Gilbert* (gilbertjack@gmail.com)

Argonne National Laboratory, Biosciences, 9700 South Cass Avenue, Argonne, IL, 60439

Microbial ecology is the study of how these communities assemble and structure within ecosystems. This ecology does not occur in isolation, but is the result of manifold interactions and feedbacks between the biological, chemical and physical components of the system. The extraordinary diversity of the microbial world has made it very difficult to elucidate the mechanism of interaction that underlies observed trends in structure and function. Advances in metagenomic sequencing and genome re-assembly provide access the individual functional potential of community members, and the ability to examine how strain level variation influences the properties of the ecosystem. The evolutionary processes that shape how bacteria respond to specific stressors, including horizontal gene transfer and synonymous/non-synonymous mutations, can help explain observed trends in the stability and resilience of community structure. Modeling the metabolism of individual bacteria can provide a framework in which to test hypotheses about potential metabolic interactions that shape ecosystem biogeochemistry. When we bring individual metabolic models together as a network of interacting components, we can extrapolate community scale observations to explain whole ecosystem properties. These can then be extrapolated to higher spatial and temporal scales to explain the interaction between microbial metabolism and climate-scale processes. Here I will highlight a number of published and ongoing studies that have started to quantify the impact of microbial evolution and natural selection on ecological dynamics. By specifically exploring bacterial communities at the strain-level we can determine how micro-evolutionary selection shapes the biochemical potential of an ecosystem, including gas flux and carbon/nitrogen metabolism.

Tight, inducible gene expression in plants with alternatively spliced suicide exons

Tania L. González* (tlgonzalez@berkeley.edu)[1], Yan Liang[2], Bao N. Nguyen[1], Geoffrey Liou[1,3], Brian J. Staskawicz[1], Dominique Loqué[2,4], Ming C. Hammond[1]

[1]University of California, Berkeley (Berkeley, CA); [2]Joint BioEnergy Institute (Emeryville, CA); [3]Massachusetts Institute of Technology (Cambridge, MA); [4]Lawrence Berkeley National Laboratory (Berkeley, CA).

Inducible promoters alone are insufficient regulation for many toxic transgenes, resulting in unwanted background phenotypes such as the hypersensitive response (HR) associated with effector-triggered immunity. We previously engineered HyP5SM, a completely plant-derived alternatively spliced cassette which can be inserted directly and tracelessly into a variety of open reading frames to inducibly regulate protein expression in dicot plants. HyP5SM takes advantage of a gene regulation strategy common in nature – alternative splicing coupled to nonsense-mediated decay – to produce “default off, inducible on” gene expression. Here, we demonstrate that HyP5SM can regulate the HR phenotype, a defensive programmed cell death response initiated by disease resistance plants upon detection of specific

pathogen effector proteins. We combine the dexamethasone inducible promoter and the HyP5SM cassette exon to regulate pathogen effector proteins. The inducible promoter alone result in leaky transcript, but HyP5SM renders the leaky transcript non-productive, thus eliminating leaky protein detected by Western blot, and leaky HR phenotype. Furthermore, plants inducibly recover both effector protein expression and the HR phenotype. We have tested this with Bs2/AvrBs2- and RPP1/ATR1Δ51-dependent hypersensitive response pathways, in *Nicotiana benthamiana* and *Nicotiana tabacum,* respectively. We also show that *Arabidopsis thaliana* plants transgenic for these resistance/effector gene pairs are viable, healthy, and can complete their full life cycle (seed-to-seed), unless the HR immune phenotype is induced with dexamethasone. With JGI, we are also engineering new sequence diverse variants of HyP5SM for multi-gene regulation in transgenic plants. The alternatively spliced HyP5SM cassette can be generally applied to regulate other genes in dicot plants, and may be utilized with conditional, constitutive, or native promoters.

Back from the dead; the curious tale of the predatory cyanobacterium *Vampirovibrio chlorellavorus*

Philip Hugenholtz* (p.hugenholtz@uq.edu.au)[1,3], Donovan H. Parks[1], Rochelle M. Soo[1], Gene W. Tyson[1,2], Ben J. Woodcroft[1]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia; [2]Advanced Water Management Centre, The University of Queensland, St Lucia, QLD 4072, Australia; [3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia

Vampirovibrio chlorellavorus was first described in 1972 by Gromov and Mamkaeva who observed it attaching to and killing a freshwater microalga, *Chlorella vulgaris*. The bacterium was originally classified as a *Bdellovibrio* due to its predatory behaviour and cell shape, however its epibiotic mode of predation called this classification into question. Several attempts to grow *V. chlorellavorus* from culture collection stocks have been unsuccessful, however its 16S rRNA gene was sequenced directly from lyophilised cells in 2010 suggesting that the bacterium is actually a member of the phylum Cyanobacteria. We have sequenced the genome of *V. chlorellavorus* from a 36 year-old lyophilised microalga co-culture. We constructed phylogenetic trees of conserved marker genes confirming that *V. chlorellavorus* is a member of the Cyanobacteria, belonging to the recently described non-photosynthetic class Melainabacteria. We identified attributes in the genome that may allow *V. chlorellavorus* to function as an obligate predator of *C. vulgaris,* and predict that it is the first described predator to use an *Agrobacterium tumefaciens*-like conjugative type IV secretion system to invade its host. We hypothesise that a 50 kb plasmid encoding multiple efflux pumps is transferred into the microalgal prey along with numerous hydrolytic enzymes synthesised in the attached bacterial cell. Prey hydrolysates may then be exported to the surrounding media via expressed plasmid efflux pumps to support the replication of *V. chlorellavorus. V. chlorellavorus* is the first cyanobacterium recognised to have a predatory lifestyle which expands the number of bacterial phyla with known predatory representatives to five, and further supports the assertion of a major non-photosynthetic cyanobacterial lineage. Our study also indicates that sequencing genomes from lyophilised material may be a viable option for the study of cultures that are not able to be resuscitated.

Data and Science Publishing in the 21st Century

Barbara R. Jasny* (bjasny@aaas.org)

Deputy Editor, Insights Section, Science

As the manner in which scientific research is done and the tools/materials for doing the research have evolved, so too science publishing is facing a need to evolve. For example, big data, multidisciplinary research, and large-scale collaborations all pose new challenges. Questions regarding reproducibility that are facing various scientific disciplines have come to the attention of the general public and policymakers. I will describe some of these issues from the point of view of Science magazine and how we are striving to publish papers that will maintain public trust and provide the solid foundation required for science to advance.

Analysis of Natural Variation for Biofuel Traits in Grasses

S.M. Kaeppler* (smkaeppl@wisc.edu)[1,2], C.N. Hirsch[1,3,4], M. Heckwolf[1,2], B. Vaillancourt[1,4], Kerry Berry[5], Erika A. Lindquist[5], S. Deshpande[5], G. Muttoni[1,2,6], J.M. Foerster[1,2,7], R.S. Sekhon[8], Michael D. Casler[1,9], N. de Leon[1,2], C.R. Buell[1,4]

[1] Great Lakes Bioenergy Research Center, [2]University of Wisconsin-Madison, [3]University of Minnesota, [4]Michigan State University, [5]DOE-Joint Genome Institute, [6]Monsanto, Inc., [7]Pioneer Hi-Bred, [8]Clemson University, [9]USDA-ARS

Annual and perennial grass species have been identified as important sources of lignocellulosic biomass production. Perennial species such as switchgrass have the potential for biomass production with the energy balance and sustainability benefits concomitant with perenniality. Annual species such as maize are a near-term source of biomass due to the large production acreage and serve as an agriculturally relevant model species. We have analyzed natural genetic variation in maize and switchgrass using genome re-sequencing and reduced representation approaches. In maize, we have observed tremendous sequence diversity including single nucleotide polymorphism variation as well as presence-absence and copy number variation. Initial evidence suggests similar genomic diversity in switchgrass, another outcrossing grass species. Analysis of natural variation has the potential to accelerate cultivar development by application of genomic selection models, and through identification of causal genes and pathways underlying important recalcitrance and productivity traits for lignocellulosic fuel production. The discovery process is enhanced by high-throughput biochemical assays and analysis of anatomical traits. Outcomes of the genomic discovery pipeline will be highlighted.

Keeping up with the Plant Destroyers – The Two-Speed Genomes of Filamentous Plant Pathogens

Sophien Kamoun* (Sophien.Kamoun@sainsbury-laboratory.ac.uk)

The Sainsbury Laboratory, Norwich Research Park, Norwich, United Kingdom

Many species of fungi and oomycetes are plant pathogens of great economic importance. The genomes of these filamentous plant pathogens have revealed a remarkable diversity in genome size and architecture. Whereas the genomes of many parasites and bacterial symbionts have been reduced over time, the genomes of several lineages of filamentous plant pathogens have been shaped by repeat-driven expansions. In these lineages, the genes encoding proteins involved in host interactions are

frequently polymorphic and reside within repeat-rich regions of the genome. This talk will review the properties of these adaptable genome regions and the mechanisms underlying their plasticity. I will also provide an update on our work on genome evolution in the lineage of the Irish potato famine organism *Phytophthora infestans*. Many plant pathogen species, including those in the *P. infestans* lineage, have evolved by host jumps followed by adaptation and specialization on distinct plant species. However, the extent to which host jumps and host specialization impact genome evolution remains largely unknown. The genomes of representative strains of four sister species of *P. infestans* revealed extremely uneven evolutionary rates across different parts of these pathogen genomes - a two-speed genome architecture. Genes in low density and repeat-rich regions show markedly higher rates of copy number variation, presence/absence polymorphisms, and positive selection. These loci are also highly enriched in genes induced *in planta*, such as disease effectors, implicating host adaptation in genome evolution. These results demonstrate that highly dynamic genome compartments enriched in non-coding sequences underpin rapid gene evolution following host jumps.

Raffaele, S., and Kamoun, S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology, 10:417-430.

Raffaele, S., Farrer, R.A., Cano, L.M., Studholme, D.J., MacLean, D., Thines, M., Jiang, R.H.Y., Zody, M.C., Kunjeti, S.G., Donofrio, N.M., Meyers, B.C., Nusbaum, C., and Kamoun, S. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science, 330:1540-1543.

Dong, S., Stam, R., Cano, L.M., Song, J., Sklenar, J., Yoshida, K., Bozkurt, T.O., Oliva, R., Liu, Z., Tian, M., Win, J., Banfield, M.J., Jones, A.M.E., van der Hoorn, R.A.L., and Kamoun, S. 2014. Effector specialization in a lineage of the Irish potato famine pathogen. Science, 343:552-555.

http://www.KamounLab.net

The Human Microbiome – A New Frontier in Human Health

Susan Lynch* (susan.lynch@ucsf.edu)

University of California San Francisco

Prevalence of childhood allergic (atopic) asthma is rapidly increasing in westernized nations, leading to speculation that it is primarily driven by environmental exposures. Humans have co-evolved with environmental microbes, though Western lifestyle, which includes protracted periods spent in the built environment, mitigate our microbial exposures. Epidemiological studies have identified several factors associated with increased risk of childhood allergic asthma e.g. Caesarian section, lack of exposure to furred pets, which are centered around the pre-natal and neonatal stages of life and plausibly influence the composition and function of the developing neonatal gut microbiome. We therefore hypothesized that the development of allergic asthma in childhood involves a lack of environmental bacterial exposure in early infancy during the critical period of microbiome and immune maturation. Indeed households with a significantly lower risk of disease development are enriched in bacteria, many of which have previously been identified in the human gut microbiome. Murine studies confirmed that exposure to protective house dust shapes gut microbiome composition in a manner that protects the airways against allergen challenge and viral respiratory infection. Moreover, microbiome profiling identified a *Lactobacillus* species that could confer this protection when fed to mice. Given this strong data linking gut microbiome composition to airway health status, we have examined the gut microbiome of approximately 300 infants (ages 0-11 months) for whom allergic status was assessed at age 2. Using an unsupervised clustering approach, we have identified three distinct microbiome configurations in the

infant gut, one of which is associated with a significantly higher risk for allergic disease development at age 2. Collectively these data indicate that the foundation for allergy development in childhood is associated with a compositionally distinct and dysfunctional neonatal gut microbiome.

Unearthing the Roots of Fungal Symbioses

Francis M. Martin* (fmartin@nancy.inra.fr) [1], Annegret Kohler[1], Alan Kuo[2], László G. Nagy[3], Emmanuelle Morin[1], Igor Grigoriev[2], David Hibbett[3] & Mycorrhizal Genomics Initiative (MGI) Consortium

[1]INRA, Laboratory of Excellence ARBRE, UMR 1136, Champenoux, France. [2]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [3]Department of Biology, Clark University, Worcester, Massachusetts, USA.

Forests health, productivity and sustainability depend on above- and below-ground microbial associations to exchange nutrients, recycle carbon, and sustain diseases and harsh environmental conditions. Fungi are often described as either saprotrophs, which degrade complex organic substrates, or biotrophs, which obtain carbon compounds from living hosts. Among the latter, ectomycorrhizal (ECM) fungi provide crucial ecological services in interacting with most forest trees. They are portrayed as mutualists trading plant host photoassimilates for nutrients and having limited capacity to decompose soil lignocellulose, but recent studies are challenging this view. It is now evident that carbon sequestration is highly influenced by ECM symbionts. An improved understanding of the role of ECM fungi and their evolutionary adaptive history in the face of changing environmental conditions will create tools to predict how they are likely to adapt to future climate change. A major goal of mycorrhizal studies is also to define the symbiosis in molecular terms, i.e. to identify the 'symbiosis genes' that encode the molecules that mediate and regulate symbiosis development and the coordinated symbiotic metabolic pathways.

To identify the genetic innovations that led to convergent evolution of the mycorrhizal lifestyle from ancestral saprotrophic species, a comparative genomics project has been implemented by the MGI consortium. We have conducted the first broad, comparative phylogenomic analysis of mycorrhizal fungi, drawing on 49 fungal genomes, 18 of which were sequenced for this study. The 18 new fungal sequences included 13 mycorrhizal genomes, from ECM fungi that colonize tree roots, and including species that commingle with orchids and heathland plant roots. The analyses of these genomes and fossils suggested that in comparison to wood decayers, such as brown rot fungi and white rot fungi, that evolved over 300 million years ago, ECM fungi emerged more recently from several species of wood and litter decayers, and then spread out across lineages less than 200 million years ago during the expansion of forest ecosystems. It appears that mushroom-forming fungi evolved a complex mechanism for breakdown of plant cell walls in 'white rot' and then cast it aside following the evolution of ECM associations.

Transcript profilings also showed that in mycorrhizal lineages there is a huge turnover in genes that are involved in the symbiosis. Many of these have no homologs in even closely related species, suggesting that the evolution of mutualism is associated with massive genetic innovation. A subset of these genes is likely used to control plant immunity during the massive colonization of root tissues by the fungus. This study suggests that the genes required for mutualism were reinvented each time it developed in evolutionary history, although similar functional categories (e.g. nutrient transporters, secreted effector proteins) appear to be expressed in a similar manner.

We have now a better understanding on how plants and fungi developed symbiotic relationships, and how the mutualistic associations provide host plants with beneficial traits for environmental adaptation. These resources will facilitate field studies aiming to predict responses of mycorrhizal communities to environmental shifts, such as altered forest-management practices and climate change.

Raman microspectroscopy based sorting of active microbial cells for single cell genomics

Marton Palatinszky* (palatinszky@microbial-ecology.net)[1], Roman Stocker[2], Kang Soo Lee[2], Tae Kwon Lee[1], Tanja Woyke[3], Michael Wagner[1]

[1]Department of Microbiology and Ecosystem Science, Division of Microbial Ecology, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria; [2]Environmental Microfluidics Group, Ralph M. Parsons Laboratory, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 15 Vassar Street, Cambridge MA 02139; [3]Targeted Genomics lab, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

Single cell genomics is about to transform microbial ecology research, however it still lacks a generally applicable approach to selectively study microbes catalyzing a specific function of interest. By a combination of stable isotope labeling, RAMAN-microspectroscopy and microfluidic manipulation, we managed to isolate microcolonies and single cells that utilize certain compounds, and apply single-cell genomic analysis on these sorted cells. Here we present an application of this method to investigate the microdiversity of ammonia- and nitrite-oxidizing bacteria in the complex microbial communities of wastewater treatment plants, and show the latest advances of bringing this method to high-throughput by combination with microfluidics. It is anticipated that this next generation sorting device for microbial cells of specific functional properties can be directly implemented into the JGI single-cell genomics pipeline in order to add a new dimension to many of the projects of JGI users.

Rationally Ranking Candidate Genes from Quantitative Trait Loci (QTLs) to Facilitate Causal Gene Discovery using Ensemble Learning

Sue Rhee* (srhee@carnegiescience.edu), Jue Fan

Carnegie Institution for Science, Department of Plant Biology, 260 Panama St, Stanford, CA 94305, USA

Many agriculturally important traits are quantitative and controlled by multiple genetic loci. QTL mapping is a popular approach for dissecting the genetic landscape of such traits. However, QTL mapping suffers from low resolution and each QTL usually contains hundreds to thousands of genes. To identify the causal gene within a QTL, fine mapping is usually implemented using dense markers in the identified confidence intervals, which is time-consuming and labor intensive. Despite a rich collection of resources and data for plants, little effort has been put into the systematic development of candidate gene prioritization algorithms.

We are developing a candidate gene prioritization algorithm by leveraging the genomic and functional annotation resources available for model organisms *Arabidopsis thaliana* and *Setaria italica*. We are incorporating features that help distinguish causal genes from non-causal genes found in a QTL region. We trained unsupervised and supervised learning models on features such as functional annotations, functional networks and polymorphisms, trained on a curated list of 52 *Arabidopsis* causal genes. We will expand the features and train semi-supervised models to boost the performance. We will

benchmark the performance of each feature and model using cross-validation as well as newly curated causal genes of other plant species from the literature. An ensemble-learning model will be proposed based on the training results described above.

We expect this algorithm to be applicable not only to QTL mapping, but also to any studies that need to rank a list of candidate genes such as genome-wide association studies.

The Evolution of Fungal Chemodiversity

Antonis Rokas* (antonis.rokas@Vanderbilt.Edu)

Department of Biological Sciences, Vanderbilt University, VU Station B #35-1634, Nashville, TN 37235

A defining characteristic of Fungi is that they live embedded in their food, digesting it externally through the secretion of enzymes and absorbing the products of digestion from their surrounding medium. Extracellular digestion poses a considerable challenge for fungal organisms; the digested food, which the fungus has spent considerable energy on, is not just accessible to them, but also to every other nearby organism. To survive in such a hostile environment, fungi have evolved a bewildering diversity of metabolic capabilities. Importantly, this phenotypic diversity is reflected in their genomes. Thus, by examining the fungal DNA record we can gain valuable insights into the evolution of their metabolic lifestyles. One conspicuous characteristic of fungal metabolic pathways is that genes participating the same metabolic pathway are frequently physically linked, forming metabolic gene clusters. My talk will focus on the fascinating biology and evolutionary history of fungal metabolic gene clusters, including the mechanisms that underlie their formation, maintenance, and decay in fungal populations, and their ecological and evolutionary impact.

Genomics-enabled Development of Ionic Liquid Tolerant Enzymes and Microbial Biofuel Hosts

Blake Simmons* (basimmons@lbl.gov)

Joint BioEnergy Institute, Emeryville, CA 94608

Ionic liquids (ILs) are emerging as superior solvents for numerous industrial applications, including the pretreatment of biomass for the microbial production of biofuels. However, some of the most effective ILs used to solubilize cellulose inhibit enzyme activity and microbial growth, thereby decreasing efficiency in the overall process. In order to overcome the limitations imposed by certain ILs on conventional enzymes and microbes, we have carried out a research program to identify, characterize, and optimize enzymes and microbes that are tolerant to both the ILs and industrially relevant process conditions in which they are used. For the development of novel enzymes, we synthesized and expressed 175 GH1s that were selected from over 2000 candidate sequences to cover maximum sequence diversity. These enzymes were functionally characterized over a range of temperatures and pHs using nanostructure-initiator mass spectrometry (NIMS). When combined with HPLC-based sugar profiling, we observed GH1 enzymes active over a broad temperature range and toward many different β-linked disaccharides. An area of particular interest was the identification of GH1 enzymes compatible with the ionic liquid 1-ethyl-3-methylimidazolium acetate ([C2mim][OAc]). We thus searched for GH1 enzymes active at 70 °C and 20% (v/v) [C2mim][OAc] over the course of a 24-h saccharification reaction. Using this approach, we identified multiple enzymes of different phylogentic origin with desired activities. The approach of characterizing sequence diversity through targeted gene synthesis coupled to

high-throughput screening technologies is a broadly applicable paradigm for a wide range of biological problems. For the development of IL tolerant microbial biofuel hosts, we identified an IL-resistance mechanism consisting of two adjacent genes from *Enterobacter lignolyticus*, a rain forest soil bacterium that is tolerant to an imidazolium-based IL. These genes retain their full functionality when transferred to an *Escherichia coli* biofuel host, with IL resistance established by an inner membrane transporter, regulated by an IL-inducible repressor. Expression of the transporter is dynamically adjusted in direct response to IL, enabling growth and biofuel production at levels of IL that are toxic to native strains. This natural auto-regulatory system provides the basis for engineering IL-tolerant microbes, which should accelerate progress towards effective conversion of lignocellulosic biomass to fuels and renewable chemicals. These advances establish the foundation for the realization of a fully integrated IL biomass conversion that does not require excessive water washing of the pretreated substrate to reduce levels to those acceptable for commercially available enzyme mixtures and biofuel hosts.

The Immune System of Bacteria – CRISPR and Beyond

Rotem Sorek* (rotem.sorek@weizmann.ac.il)

Weizmann Institute of Science

The perpetual arms race between bacteria and phage has resulted in the evolution of efficient resistance systems that protect bacteria from phage infection.

Such systems, including the CRISPR-Cas system, have major influence on the evolution of bacteria and phage genomes, and have also proven to be invaluable for molecular and biotechnological applications.

This talk will revolve around the CRISPR adaptation process, where short pieces of DNA ("spacers") are acquired from foreign elements and integrated into the CRISPR array. It so far remained a mystery how spacers are preferentially acquired from the foreign DNA while the self-chromosome is avoided. In a genome-assisted study we found a two component mechanism that allows the system to minimize spacer acquisition from self DNA- these results will be discussed in the talk.

Marsh madness: microbial communities driving greenhouse gas cycling in coastal wetlands

Susanna Theroux* (stheroux@lbl.gov), Wyatt H. Hartman, Susannah Tringe

DOE Joint Genome Institute, Walnut Creek, California

Climate change is predicted to increase the average salinity of coastal wetlands as sea levels rise and alpine snow volume decreases. Increasing salinities will impact the soil microbial communities that drive greenhouse gas (GHG) cycling in these sensitive environments, disrupting the balance between aboveground primary production and below-ground carbon burial. Current projections of wetland response to climate change are poorly constrained due to a limited understanding of soil microbial carbon cycling coupled with the large variations in wetland tidal regimes, organic carbon content, restoration status, and other geochemical and biological variables. To better understand controls on GHG flux in wetland soils, we probed the microbial communities along a salinity gradient in the San Francisco Bay-Delta region using 16S rRNA gene sequencing and shotgun metagenomics in tandem with greenhouse gas monitoring. Methane production was consistent with sulfate availability across all sites and enhanced in freshwater and brackish restored wetlands. Microbial community composition and

metabolic potential clustered strongly according to sampling site, plant type, and salinity. Individual methanogen species were highly correlated with methane production, and methanogen substrate preferences helped explain surprising methane production rates in some freshwater and hypersaline ponds. As the most extensive survey of coastal wetland microbial communities to date, our study provides a catalog of the microbial metabolic arsenals that dictate carbon cycling in these rapidly changing environments.

Comparative and Population Genomics: Understanding the factors driving Genome-Wide Natural Selection

Stephen Wright* (stephen.wright@utoronto.ca)

Department of Ecology and Evolutionary Biology, University of Toronto, Ontario, Canada

One of the most striking features of genome evolution concerns cases where large-scale genomic regions experience widespread gene degeneration and loss. This includes sex chromosome degeneration and diploidization following whole genome duplication, where gene copies that once shared common ancestry undergo degenerative evolution. Such patterns are surprising, since they reflect a dramatic loss of selective constraint. Although the features of degenerative genome evolution have been well documented, much of the focus has been on ancient events, making it difficult to investigate the processes driving the earliest stages of genome degeneration. Plants often experience recurrent transitions in ploidy and mating system, allowing for important opportunities to investigate the early stages of genome evolution. Here I discuss our work investigating how rapidly plant populations experience major shifts in selective pressure following transitions in mating system and ploidy. The results suggest the potential for rapid and major shifts in selection pressure following evolutionary transitions to self-fertilization, polyploidy, asexual reproduction and sex chromosome formation. Large-scale comparative plant genomics efforts will enable powerful replicated tests factors governing differences in selection across lineages.

Identification and Isolation of Genes that Trigger *Laccaria* Colonization in *Populus*

Gerald A. Tuskan* (tuskanga@ornl.gov)[1,5], David J. Weston[1], M. Robeson[1], Dale A. Pelletier[1], Christopher W. Schadt[1], W. ellington[1], R. Vilgalys[3], J. Plett[4], Francis M. Martin[4], Anthony Bryan[1], G. Bonito[3], T. H. Ahn[1], Chongle Pan[1], J. Labbe[1], T. Tschaplinski[1], Xiaohan Yang[1], H. Yin[1], Kerry Berry[5], and Jeremy Schmutz[6]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of Tennessee, Knoxville, TN; [3]Duke University, Durham, NC; [4]INRA, Nancy, France; [5]Joint Genome Institute, Walnut Creek, CA; [6]HudsonAlpha, Huntsville, AL

There is a hidden molecular language that plants and microbes use to communicate. Proteins, metabolites and possibly even RNAs are exchanged between plants, bacteria and fungi. The beneficial, commensal, detrimental or neutral outcome is determined through such signal prior to intimate contact among the communal participants. Small secreted proteins discovery, subcellular localization, secretion in yeast system and *Laccaria* nuclear localization are presented. Results indicate that there are several hundred small proteins produced in *Populus* that have the ability to move into *Laccaria*, with a subset that invokes anatomical changes in hyphal growth. Using a GWAS approach, containing 1084 unrelated

genotypes clonally replicated in four contrasting environments, with each genotype resequenced to a minimum 18X depth, for identifying genes involved in *Laccaria* colonization, we created a candidate gene list, validated this subset in a QTL population and identified a D-mannose lectin kinase that was wholly correlated with successful colonization. Overexpressing the *Populus* D-mannose kinase in transgenic *Arabidopsis* resulted in intercellular hyphal growth in *Arabidopsis*.

Poster Presentations

Posters alphabetical by first author. *Presenting author

Shedding light on 'Aigarchaeota', an uncultivated lineage of obligate thermophiles

Timothy W. Alba* (albat@unlv.nevada.edu)[1], Senthil K. Murugapiran[1], Asaf Levy[2], Jeremy A. Dodsworth[3], Gisele B. Goertz[1], Natalia N. Ivanova[2], Tanja Woyke[2], Brian P. Hedlund[1,4]

[1]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA; [2]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA; [3]Department of Biology, California State University, San Bernardino, CA, 92407, USA; [4]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV 89154, USA

The phylum '*Aigarchaeota*' is a yet-uncultivated "microbial dark matter" lineage that inhabits terrestrial, marine, and deep subsurface geothermal environments. Clustering approaches, combined with phylogenetic analysis of all nearly complete '*Aigarchaeota*' 16S rRNA gene sequences reveals at least nine genus-level groups. '*Aigarchaeota*' seem to be obligate thermophiles and primarily inhabit circumneutral pH environments, although each group differs in the apparent optimal growth temperature and environmental distribution. Some geothermal springs are hotspots for '*Aigarchaeota*' abundance and diversity, including Great Boiling Spring (NV), Little Hot Creek (CA), and Gongxiaoshe, China. Substantial single-cell genomic or metagenomic datasets exist for six of the nine genus-level groups. All groups are predicted to be facultative chemoautotrophs, with highly variable predicted capacity for the oxidation of carbon monoxide, formate, and reduced sulfur compounds and dissimilatory reduction of oxygen, nitrate, nitrite, nitrous oxide, and sulfate. It's likely that '*Aigarchaeota*' lineages play important roles in global carbon, nitrogen, and sulfur cycles.

2500 genomes provide a comprehensive view of microbial metabolism in a subsurface ecosystem

Karthik, Anantharaman* (karthik.a@berkeley.edu)[1], Christopher T. Brown[2], Itai Sharon[1], Brian C. Thomas[1], Andrea Singh[1], Laura A. Hug[1], Cindy J. Castelle[1], Kenneth H. Williams[3], Eoin L. Brodie[3], Susannah Tringe[4], Jillian F. Banfield[1,5]

[1] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA 94720 USA; [2] Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720 USA; [3] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [4] Metagenome Program, DOE Joint Genome Institute, Walnut Creek, CA 94598 USA; [5] Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720 USA.

Microbes drive the transformations of carbon and impact other linked biogeochemical cycles in the terrestrial subsurface, an important reservoir of carbon on earth. Despite this, little is known about the identity, function and roles of the majority of microorganisms in the subsurface. These knowledge gaps have restricted the scope and ability of climate models to capture key aspects of the carbon cycle. Here, we used a genome-resolved metagenomic approach to decipher the metabolic capabilities of the uncultivated microbial majority in the subsurface. We sequenced groundwater and sediment samples collected across four different geochemical environments from an aquifer adjacent to the Colorado

River, near Rifle, CO, USA. Sequence assembly, manual curation and binning resulted in the recovery of 2,542 high-quality genomes, 20 of which are complete. Phylogenetic analyses involving ribosomal proteins and 16S ribosomal RNA genes revealed that less than 11% of these genomes belonged to the 4 most commonly represented phyla in public databases, that constitute 93% of all currently available genomes. A large proportion of the genomes we recovered belong to phyla that were hitherto unknown or lack cultivated representatives. Genome-specific analyses of metabolic potential revealed the co-occurrence of important traits involving carbon fixation, nitrogen fixation, electron donors (e.g. organic carbon, sulfur, hydrogen, methane, ammonia) and electron acceptors (e.g. oxygen, nitrate, sulfate). Coupled analyses of genomic abundance and metabolic potential provide a unique perspective on microbial functional redundancy and heterogeneity across time (~6 years) and geochemical environments (sediments, groundwater). This study significantly advances the understanding of the genetic underpinnings of microbial diversity and function in the subsurface. Our analyses of microbial community composition and predictions of metabolic potential will serve as inputs into Lawrence Berkeley National Laboratory's Genome-Enabled Watershed Simulation Capability (GEWaSC) modeling effort.

Global host-virus transcriptomic profile: differential gene expression in the viral infection of the marine alga *Micromonas pusilla* under nutrient limitation

Charles Bachy* (cbachy@mbari.org)[1], Christina Charlesworth,[2] A.J. Limardo,[1] Jan Finke,[2] Emily Reistetter,[1] Amy Chan,[2] Chee-Hong Wong,[3] Chew Yee Ngan,[3] Chia-Lin Wei,[3] Curtis Suttle[2] and Alexandra Z. Worden[1].

[1]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039, USA.; [2]Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada; [3]Sequencing Technology Group, Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598, USA.

Approximately half of the world's organic carbon is fixed by marine phytoplankton via photosynthesis. These organisms often grow rapidly and can form massive blooms in the oceans, however the suite of abiotic (e.g., nutrients and light availability) and biotic (e.g., viruses) interactions acting on phytoplankton growth and mortality remain to be fully characterized.

Within unicellular eukaryotic phytoplankton are small taxa termed "picoeukaryotes" (≤ 2 micron cell diameter). The green alga *Micromonas* is a particularly widespread picoeukaryote belonging to the prasinophytes which together with the chlorophytes (e.g., *Chlamydomonas reinhardtii*) share a common ancestor with algal and terrestrial streptophytes (plants). The *Micromonas pusilla* genome has been completely sequenced and is 22 Mb with 9,830 predicted protein-encoding genes. The 173 kb genome of a *virus* infecting *M. pusilla* (MpV-SP1) has also been completely sequenced and has 243 predicted protein-encoding genes. We are interested in understanding *M. pusilla* at physiological and molecular levels and in particular responses to viral infection along with nutrient depletion. Using pair-ended directional Illumina mRNA sequencing, we investigated gene expression of *M. pusilla* in experiments that allowed characterization of the cell cycle in synchronously growing cultures on a 14:10 Light:Dark cycle. Experiments were also performed on the impact of nutrient-limitation on cell growth and gene expression with, and without, concurrent viral infection.

Over the diel the majority of nuclear genes (89%, FDR-adjusted p-value <0.05) show significant changes in expression, including genes encoding proteins involved in photosynthesis, pigment synthesis and light harvesting. The greatest differential change was observed between time points at "dusk" and

"predawn" (71%). Nutrient limitation experiments show more moderate differences between nutrient-replete controls and nutrient-limited cultures than observed during the diel, with 33% being the maximum percentage of genes showing significant differences. The global gene expression of host and viral genes along the course of infection provide additional insights into key viral genes that are differentially expressed in nutrient limited conditions. This research provides valuable data for understanding green lineage evolution as well as growth and mortality of this important marine alga.

Seasonal dynamics of fungal and bacterial activity in forest soil reflects changes in ecosystem properties

Petr Baldrian* (baldrian@biomed.cas.cz)[1], Salvador Lladó Fernández[1], Tomáš Větrovský[1], Adina Chuang Howe[2], Lucia Žifčáková[1]

[1] Institute of Microbiology of the ASCR, v.v.i., Vídeňská 1083, 14220 Praha 4, Czech Republic; [2] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA

Understanding the ecology of coniferous forests is very important because these environments represent one of the largest carbon sinks globally, especially in the Northern Hemisphere. Recognition of microorganisms as fundamental mediators of C-cycling processes in these ecosystems raises the question of how microbial communities function in a temperate zone where environmental conditions show dramatic seasonal changes. The aim of this work was to combine metatranscriptomics, microbial community analysis and enzyme analyses to describe the role of individual microbial taxa in the functioning of the *Picea abies*-dominated coniferous forest soil in two contrasting seasons. These seasons were the summer at the peak of plant photosynthetic activity and late winter after an extended period with no photosynthate input. The results show that the microbial communities in soil and litter were highly diverse both in composition and function. Additionally, these communities were characterised by a high abundance of fungi. The differences in seasonal functioning of the ecosystem consisted of a combination of moderate changes in microbial community composition and profound changes in taxon-specific microbial transcription profiles. These differences were more significant in soil than in litter. Most importantly, fungal contribution to total microbial transcription in soil decreased from 33% in summer to 16% in winter. In particular, the activity of the ectomycorrhizal fungi that quantitatively dominate this environment was reduced in winter. The results indicate that plant photosynthetic production was likely the major driver of changes in the functioning of microbial communities in the studied ecosystem across seasons.

The annotation of metatranscriptomic reads to higher microbial taxa and functions indicated a profound difference in our ability to identify bacterial versus eukaryotic (fungal) transcripts. While of bacterial transcripts, 72% can be annotated to function, it is the case only for 19% of fungal transcripts. In addition, the assignment of fungal transcripts to lower taxa on the level of divisions or classes is currently also largely unreliable. The reason is not only the lesser number of fungal genomes that are annotated and publicly available at the moment, but also their incorporation into the annotation pipelines. To overcome the problems of annotation, we have used the isolation of microbial taxa from soil and the analysis of single-cell-derived genomes . While this approach might be useful to identify specific transcripts of individual taxa, it is limited in its use for the annotation the whole of the transcript pool. Despite the current limitations, metatranscriptomics and its combination with metaproteomics and metagenomics represent viable approaches for the analysis of microbial functions in the soil whose importance will largely increase in the future.

A genomic survey of genes encoding H_2O_2-producing GMC oxidoreductases in Polyporales

Jose M. Barrasa* (josem.barrasa@uah.es) [3], **Juan Carro**[1], **Patricia Ferreira**[2], Angel T. Martínez[1], and Ana Serrano[1]

[1]Centro de Investigaciones Biológicas, CSIC, Ramiro de Maeztu 9, E-28040-Madrid, Spain
[2]Departamento de Bioquímica and BIFI, Universidad de Zaragoza, E-50009 Zaragoza, Spain
[3]Departamento de Ciencias de la Vida, Universidad de Alcalá, Ctra Barcelona Km 33.6, E-28871 Alcalá de Henares, Madrid, Spain

The genomes of three representative Polyporales (*Bjerkandera adusta, Phlebia brevispora* and a member of the *Ganoderma lucidum* complex) were sequenced, in the JGI Saprotrophic Agarycomycotina Project (SAP) coordinated by David Hibbet (Clark University), to expand our knowledge on the diversity and distribution of genes involved in the attack on wood polymers by the species of this Basidiomycota order, which includes most wood-rotting organisms. Oxidases, including members of the glucose-methanol-choline oxidase (GMC) oxidoreductase superfamily, play a central role in the above degradative process since they generate extracellular H_2O_2 acting as the ultimate oxidizer in both white-rot and brown-rot decay. The survey was completed by analyzing the GMC genes in the already available genomes of seven more species to cover four Polyporales clades (up to a total of 10 genomes analyzed). First, an *in silico* search for sequences encoding members of the aryl-alcohol oxidase, glucose oxidase, methanol oxidase, pyranose oxidase, cellobiose dehydrogenase, and pyranose dehydrogenase families was performed. The curated sequences were subjected to an analysis of their evolutionary relationships, followed by estimation of gene duplication/reduction history during fungal evolution. Secondly, the molecular structures of near one hundred GMC oxidoreductases identified in the genomes were modeled to gain insight into their structural variation and expected catalytic properties confirming the previous classification. In contrast to ligninolytic peroxidases, whose genes are characteristic of the typical white-rot fungal Polyporales and absent from those of brown-rot species, (1) the H_2O_2-generating oxidases are widely distributed in both the white-rot and brown-rot fungal genomes. This indicates that the same enzyme type providing H_2O_2 to ligninolytic peroxidases (in white-rot decay) supply the H_2O_2 required for Fenton attack on cellulose (in brown-rot decay) after the transition between both decay patterns in Polyporales occurred.

1. Ruiz-Dueñas, F. J., T. Lundell, D. Floudas, L. G. Nagy, J. M. Barrasa, D. S. Hibbett, and A. T. Martínez. 2013. Lignin-degrading peroxidases in Polyporales: An evolutionary survey based on ten sequenced genomes. Mycologia 105:1428-1444.

Necessary but not sufficient: towards a casual analysis of the role that functional genes play in greenhouse gas emissions

Jason Barker* (rhizo@mail.ubc.ca)

Dept. of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada.

Genomics offers a wealth of opportunities and pitfalls for ecologists. Assessments of functional gene activity from “omics” studies offer ecologists a window into the fine -scale genetic processes that underpin ecosystem processes, such as GHG emissions. Given the complexity of interactions between microbes and the environment, establishing causal relationships between genetic factors and GHG emissions can be challenging, especially to get models past the conceptual stage and into empirical verification. I present a basic framework for incorporating functional genes into models of methane

cycling using a causally-explicit structural equation models. I use partial least square (PLS) structural equation modelling, with PLS multivariate regression, to translate the structural models into practice. In the first model, I use unpublished data from a Vancouver Island study to model methane emissions from wetland and upland forested sites. Here, I model the gene abundances of pmoA and mcrA as a direct and mediated factor (through soil moisture) in determining methane fluxes. The proposed PLS SEM model has a good fit (0.56), with moisture a stronger driver of methane emissions (0.6) than functional genes (0.14).The model estimated a moisture mediation through functional genes as 0.20 of the total gene effect. However, the influence of functional genes was greater in the upland sites (-2.0) verses wetland sites (0.4). The moisture mediation effect remained comparable, but was negative in the upland (-0.19) and positive in the wetland site (0.20). In a second model, I used PLS regression to model the relationship within gene transcription pathways related to methanogenesis and methnotrophy from a published Alaskan permafrost study. The results indicate support the development for incorporating transcription of functional genes into latent variables. Finally, I present a conceptual model for incorporating a wider range of genomic data, such as from secretome analysis, as well as modifications of the model to handle incorporating emergent ecosystem properties across spatial scales.

Nucleotid.es - objective benchmarking of bioinformatics software

Michael Barton* (mbarton@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, CA, USA

As novel technologies emerge, fueled by scientific questions, the number bioinformatics tools continues to grow. When publishing new bioinformatics software, the tool is typically compared against others' using test data. Naturally, researchers will select the results that portray their software in the best way, which can lead to a subjective comparison, sometimes even lacking reproducibility. Combining this with the current large numbers of software publications in bioinformatics it is difficult for researchers to objectively evaluate which software will work best in their analysis, since different tools are rarely tested on identical data and settings for running tools can be very custom.

To address these challenges, we developed the nucleotid.es project, which places bioinformatics tools in software containers that allow them to be used on any platform, and reproducibly benchmarked against reference sequencing data. This allows software to be evaluated simultaneously and without requiring manual installation or setting of parameters. The tools benchmarked in nucleotid.es are submitted by the bioinformatics community or created from the existing literature. This effectively crowd-sources software from anyone who wishes to participate. The latest bioinformatics research can thus constantly be evaluated against the current existing corpus and the result of this benchmarking then shared with the wider bioinformatics community as they are generated.

This talk will present nucleotid.es and the results of running this analysis on a range of genome assemblers.

Assembly, Succession, and Activity of Iron Oxide Microbial Mat Communities in Acidic Geothermal Springs

Jacob P. Beam* (jprbeam@gmail.com)[1], Zackary J. Jay[1], Ryan M. Jennings[1], Hans C. Bernstein[2], Mark A. Kozubal[1], Susannah Tringe[3], James J. Moran[2], Roslyn Brown[2], Mary S. Lipton[2], Margaret F. Romine[2], James K. Fredrickson[2], and William P. Inskeep[1,2]

[1]Thermal Biology Institute and Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT; [2]Pacific Northwest National Laboratory, Richland, WA; [3]Department of Energy-Joint Genome Institute, Walnut Creek, CA.

Iron oxide microbial mats are ubiquitous geobiological features on Earth and occur in extant acidic hot springs of Yellowstone National Park (YNP), WY, USA, and form as a result of microbial processes. Our group has studied high-temperature iron oxide mat ecosystems utilizing physicochemical and stable carbon isotope measurements combined with multiple 'omics' technologies to dissect the functional role of microbial populations *in situ*. Moreover, spatiotemporal measurements of iron oxide accretion and community structure have revealed the importance of specific microbial populations to the formation of Fe(III)-oxide mats in acidic geothermal outflow channels. We hypothesized that chemolithoautotrophic organisms contribute to the early development and production of Fe(III)-oxide mats, which could support later-colonizing heterotrophic microorganisms. Sterile glass slides were incubated in the outflow channels of two acidic geothermal springs in YNP, and spatiotemporal changes in Fe(III)-oxide accretion and abundance of relevant community members were measured using 16S rRNA gene Illumina Tag (iTag) sequencing. Lithoautotrophic *Hydrogenobaculum* spp. were first colonizers and the most abundant taxa identified during early successional stages (7 – 40 days). Populations of the iron-oxidizing crenarchaeon, *Metallosphaera yellowstonensis* colonized after ~ 7 days, corresponding to visible Fe(III)-oxide accretion. Fluorescence *in situ* hybridization probes targeting *Hydrogenobaculum* spp. and *M. yellowstonensis* confirmed their importance during early mat assembly. Organoheterotrophic archaea colonized after 30 days, and emerge as the dominant functional guild in mature iron oxide mats (1 – 2 cm thick) that form after 70 – 120 days. Metatranscriptomes from several mature Fe(III)-oxide mats showed that *Hydrogenobaculum* and *M. yellowstonensis* populations are highly active and collectively account for 50-60 % of the total mRNA reads, which was further supported by metaproteomics. First-order rate constants of iron oxide accretion ranged from 0.05 – 0.046 day^{-1}, and correspond to 70 moles of Fe oxidized per carbon mole of biomass produced for *M. yellowstonensis*. Stable carbon isotope measurements revealed that autotrophic derived biomass accounts for no less than 40 % of the total organic carbon in iron oxide mat ecosystems. Micro- and macroscale microterracettes were identified during iron oxide mat development, and suggest that the mass transfer of oxygen limits microbial growth. This was also demonstrated using microelectrode measurements of oxygen as a function of mat depth, which showed steep gradients in oxygen from the aqueous-mat interface to ~ 1 mm. Aerobic *Hydrogenobaculum* spp. and *M. yellowstonensis* are more abundant at the aqueous-mat interface, and metatranscriptome analyses confirm their importance to oxygen consumption *in situ*. The formation and succession of amorphous Fe(III)-oxide mat communities follows a predictable pattern of distinct stages and growth. The successional stages and microbial signatures observed in these extant Fe(III)-oxide mat communities may be relevant to other past or present Fe(III)-oxide mineralizing systems.

Pioneering fungal mutagenesis using Tn-seq

Rachel B. Brem* (rbrem@buckinstitute.org)[2,3], **Monica R. Mascarenas**[1], Jeffrey M. Skerker[2], Jeremy I. Roop[2], Kyu Chul Chang[2], Gina Geiselman[2], Adam P. Arkin[2], Maitreya J. Dunham[1]

[1]University of Washington, Seattle, WA. [2]University of California, Berkeley, Berkeley, CA. [3]Buck Institute for Research on Aging, Novato, CA.

Functional annotation is a critical need in modern fungal genomics, largely owing to a dearth of experimental methods for the assignment of functions to unknown fungal genes. Transposon mutagenesis and pooled sequencing (Tn-seq) is a powerful approach for the generation and screening of pooled deletion libraries that does not require labor-intensive targeted deletion design or maintenance of individual knockout lines. To date, Tn-seq strategies have not been broadly applied in fungi. We are using *Saccharomyces* yeasts as a testbed for the development of Tn-seq methods that enable the novel assignment of genes to growth attributes.

In a first methods development effort, we have developed a platform that uses the PiggyBac transposase, expressed in fungal cells, to mutagenize the host genome by insertion of a drug marker flanked by PiggyBac recognition sites. To date, we have established integrated and plasmid-borne versions of this PiggyBac machinery in *S. cerevisiae* and carried out an initial round of sequencing to verify the location of transposon insertion sites. In a second, more advanced pilot project, we have focused on identification of essential genes in *S. uvarum*, a long-diverged relative of the model species *S. cerevisiae*. We generated a plasmid library of long fragments of the *S. uvarum* genome, mutagenized these with the Tn7 transposon, and re-introduced the mutant fragments into wild-type diploid and haploid *S. uvarum* strains by massively parallel homologous recombination. We maintained each mutant collection as a pool, and cultured each in standard media conditions for ten generations to drive fit mutants to high frequency and allow unfit mutants to drop out of the population. We then isolated DNA from each pool and, from it, amplified and sequenced a region corresponding to the Tn7 insert of a given clone and the flanking genomic DNA. For mutants in 1034 genes, we detected no representation in the strain pool in the haploid background after selection. This set of candidate essential genes in *S. uvarum* was enriched for those known to be refractory to deletion in multiple *S. cerevisiae* backgrounds. Remarkably, however, mutants of 115 genes essential in *S. cerevisiae* were viable in our *S. uvarum* libraries, and mutants of 24 genes amenable to deletion in *S. cerevisiae* were absent from our *S. uvarum* mutant libraries. In compelling candidate cases, we have used single-gene knockout experiments to validate the species-specific essentiality and non-essentiality of these *S. uvarum* genes.

Together, our data attest to the power of the Tn-seq approach for the rapid generation and phenotyping of genome-scale mutant libraries in fungi. In *Saccharomycetes*, our methods will lead to the discovery of genes that underlie species-specific traits, including essentiality as well as metabolic behaviors relevant for biofuel production. And with our proof of concept established in the yeast model, the tools for Tn-seq we develop will be of immediate use for functional annotation across the fungal kingdom.

The coding potential of *Pseudomonas aeruginosa*: information from sequence 3-base periodicities, computational predictions, conservation, and experimental evidence of translation.

Luciano Brocchieri* (lucianob@ufl.edu)[1,2], Steve Oden[1,2], Anna Picca[1], and Silvia Tornaletti[3]

[1] Department of Molecular Genetics and Microbiology; [2] Genetics Institute; [3] Department of Medicine; University of Florida, Gainesville, FL 32610

The NPACT (N-Profile Analysis Computational Tool: http://genome.ufl.edu/npact), quantifies the principle of frame analysis implementing two procedures for detecting sequence regions with significant 3-base compositional periodicities indicative of codon structures. We used the information provided by NPACT in conjunction with annotations, gene predictions, and evolutionary conservation, to computationally evaluate the coding potential of nine strains of the opportunistic human pathogen *Pseudomonas aeruginosa*. We identified in annotated intergenic regions of the nine *P. aeruginosa* genomes, regions of significant contrast corresponding to 158-259 ORFs not included in the published annotations per genome. Popular gene predictors (Prodigal, Glimmer3.0 GeneMarkHMM, and GeneMark2.5) predicted in the nine strains more than 3,800 genes not included in annotations (almost half of which are uniquely predicted by Glimmer3.0). Among all predicted genes not included in the annotations, more than 900 genes were conserved across different bacterial genera or phyla. We found that among these conserved genes, 550 were characterized by significant compositional periodicities, whereas 204 other conserved ORFs with sequence periodicity were not predicted by any method.

Experimental determinations of expression by transcriptome sequencing (RNA-seq) can be complicated in prokaryotes by the polycistronic nature of bacterial mRNA and by presence of extensive anti-sense transcription. More accurate measures of expression, and precise determination of the position of translation-initiation sites of individual genes, can be obtained using the technique of ribosome profiling (Ingolia et al 2009), by which only regions of the mRNA that are actively translated can be sequenced. We performed ribosome profiling on the transcriptome of *P. aeruginosa* strain PAO1, providing experimental evidence of translation for many of the computationally predicted sequences. Ribosome-footprint profiles provided convincing evidence of expression of most annotated genes, allowed us to correct mis-predicted start-of-translation sites, and verified expression of about 30% of the predicted genes not included in the published annotation. Furthermore, ribosome footprints provided evidence of the existence of coding regions not predicted by any computational method, and of previously unknown potential mechanisms of post-transcriptional regulation of translation, in the form of 'leader peptides' and alternative translation-initiation sites.

Sequence three-base periodicities and the missing genes from prokaryotic genome annotations.

Luciano Brocchieri* (lucianob@ufl.edu)[1,2] and Steve Oden[1,2]

[1] Department of Molecular Genetics and Microbiology; [2] Genetics Institute

Gene annotation in prokaryotic genomes is challenged by inconsistencies among gene predictors and by incomplete or ambiguous information on evolutionary conservation. Evidence of genes missed by genome annotations has been provided in recent analyses. Comparing genes annotated in 1000 genomes to automated gene predictions from four popular prediction methods (Prodigal, Glimmer3.0, GeneMarkHMM, GeneMark2.5), we identified more than 570,000 predicted genes that were not included in published annotations. Although lack of convincing evidence of conservation could justify the exclusion from annotations of most of these predictions, we also found among these excluded genes almost 115,000 genes that appeared to be conserved in sequence and in length across bacterial genera or phyla. To facilitate recognition of potential genes during the annotation process, we introduced the NPACT (N-Profile Analysis Computational Tool) procedure and web-based application (http://genome.ufl.edu/npact), for the identification of sequence regions of statistically-significant 3-

base periodicity and associated ORFs, and for their graphical representation in comparison with sets of pre-annotated genes and genome-phase-specific compositional profiles. The NPACT graphical interface allows users to quickly "surf" through a complete bacterial genome, as opposed to "browsing" it on a window-by-window base, and highlights the differences between a "reference" set provided as a "pre-annotation", and the set of ORFs identified by their periodic structure (a "test" set), providing annotators with the ability to quickly incrementally improve the annotation and to focus on potential problem areas for further analysis.

Although compositional contrasts among codon phase-positions are stronger in coding sequences of high GC content, regions of significant three-base periodicities were identified by NPACT in "intergenic" regions annotated across all 1000 genomes. Among these three-periodic regions, about 77,000 could be associated with ORFs whose coding potential was also supported by some, and in the majority of cases by all, gene predictors. Furthermore, almost half of the predicted genes with three-base periodicity were also supported by long-range evolutionary conservation in sequence and in length, compared to less than 16% of the predicted genes excluded from annotations that did not show three-periodicity in their composition. Combining information from gene prediction and sequence three-base periodicity, we predict that as much as 30% of the currently annotated genomic inter-genic space could be occupied by unrecognized coding regions.

Phylogenetic Richness in the Evolution of Molecular Sequences and its Application to Bacterial Genomes.

Luciano Brocchieri* (lucianob@ufl.edu) and Hye-Won Lee[1,2]

[1] Department of Molecular Genetics and Microbiology and [2] Genetics Institute, University of Florida, Gainesville, FL 32610

Measures of diversity are relevant for the development of sampling strategies to optimize diversity of datasets of molecular sequences (e.g., microbial genomes) and for the analysis of ecological communities, among which those sampled in metagenomics studies (e.g., microbiomes). Current measures of diversity frequently used in ecological studies are based on the assumption of parsimonious evolution. We developed a measure of Phylogenetic Richness and related measures based on evolutionary trees obtained from alignments of molecular sequences, taking into account their non-parsimonious probabilistic pattern of evolution.

A natural measure of how much diversity from common ancestry of the state of a character is represented within a sample of *N* sequences is given by the number of states per site not identical by descent that we expect to observe across sequences, given their evolutionary relations. Genetic relatedness induces dependencies among sequences reducing their "effective number": two sequences with evolutionary distance $d = 0.0$ are effectively one sequence, and can become two independent sequences only when their evolutionary distance d tends to infinity. In general, a sample of *N* related sequences corresponds to an effective number *E* of sequences between 1.0 and *N*. Measures of phylogenetic diversity based on phylogenetic-tree branch lengths, assume Camin-Sokal (1965) parsimonious accumulation of features and a rooted tree. Our calculations assume time reversible probabilistic substitution of states and do not require a rooted tree. We call our measure *Expected Phylogenetic Richness E*. The definition of *E* naturally leads to a definition of "sampling density" ρ, of relatedness *r* for groups of sequences, and a related normalized definition of diversity *D* normalized to the interval [0-1].

We calculated Expected Phylogenetic Richness on an evolutionary tree generated from the concatenation of 169 protein families conserved across most of 1,800 bacterial genomes comprising 31 distinct phyla. *E* and *ρ* provide information accounting for phylogenetic relations on, for example, which groups are more diversely represented, on the density with which different bacterial groups have been sampled, and on how sequencing new bacterial genomes has increased overall sample diversity. Our measures are useful for constructing balanced samples of phylogenetic groups and for comparing diversity within and between microbial communities from metagenomics or metatranscriptomics studies.

Unusual Biology Across a Group Comprising >15% of Domain Bacteria

Christopher T. Brown* (ctb@berkeley.edu)[1], Laura A. Hug[2], Brian C. Thomas[2], Itai Sharon[2], Cindy J. Castelle[2], Andrea Singh[2], Michael J. Wilkins[3,4], Kelly C. Wrighton[4], Kenneth H. Williams[5], and Jillian F. Banfield[2,5,6]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA; [2]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA; [3]School of Earth Sciences, The Ohio State University, Columbus, OH, USA; [4]Department of Microbiology, The Ohio State University, Columbus, OH, USA; [5]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [6]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA.

A prominent feature of the bacterial domain is a radiation of major lineages that are defined as Candidate Phyla (CP) because they lack isolated representatives. Bacteria from these phyla occur in diverse environments and are suggested to mediate carbon and hydrogen cycles. Genomic analyses of a few representatives suggested that metabolic limitations have prevented their cultivation. We reconstructed 8 complete and 789 draft genomes from bacteria representing >35 phyla and documented features that consistently distinguish these organisms from other bacteria. We infer that this group, which may comprise >15% of the bacterial domain, has shared evolutionary history and describe it as the Candidate Phyla Radiation (CPR). All CPR genomes are small and most lack numerous biosynthetic pathways. Due to divergent 16S rRNA gene sequences, 50-100% of organisms sampled from specific phyla would evade detection in typical cultivation-independent surveys. CPR organisms often have self-splicing introns and proteins encoded within their rRNA genes, a feature rarely reported in bacteria. Further, they have unusual ribosome compositions. All are missing a ribosomal protein often absent in symbionts, and specific lineages are missing ribosomal proteins and biogenesis factors considered universal in bacteria. This implies different ribosome structures and biogenesis mechanisms, and underlines unusual biology across a large part of the bacterial domain.

Design and analysis of medium-throughput compound synergy screens with SynergyScreen.

Yury V. Bukhman* (ybukhman@glbrc.wisc.edu), Scott Bottoms, Quinn Dickinson, Li Hinchman, and Jeff S. Piotrowski

Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, University of Wisconsin – Madison, 1552 University Ave., Madison, WI 53726, USA

Screening sets of chemical compounds for potential synergy or antagonism has a wide range of applications, from medicine to bioenergy research. For bioenergy research purposes, it helps to understand the effects of multiple lignotoxins that are present in plant biomass hydrolysates. These lignotoxins have deleterious effects on microbial fermentation and biofuel yields.

In order to facilitate design and analysis of synergy screens, we have developed an R package, SynergyScreen. Given a set of compounds and dose ranges, SynergyScreen can produce a design of a screen testing their pairwise combinations. The package generates layouts for a set of 96-well microtiter plates. It includes titration series for each compound and each combination in a fixed ratio, comprising a set of single-ray synergy experiments. Once the experiments have been carried out, SynergyScreen can analyze the data to detect synergistic and antagonistic compound pairs. The analysis includes normalization to remove potential plate bias, modeling each individual dose-response curve, and computing interaction index values for a set of effect sizes. The package can produce tabular outputs and visualizations. We demonstrate SynergyScreen functionality using example data.

BBMerge: Accurate Paired Read Merging via Overlap

Brian Bushnell* (bbushnell@lbl.gov)[1], **Jonathan Rood**[2]

[1]Joint Genome Institute, Walnut Creek, CA; [2]NERSC, Oakland, CA

Read-pair merging is an important part of many bioinformatics pipelines, including variant detection, 16S classification and quantification, and assembly. Overlap detection without merging is useful in even more areas, such as error-correction and alignment-free insert-size determination. Yet despite being a seemingly simple problem, the most commonly-used overlap-detection and merging tools have major problems in performance and accuracy that can lead to incorrect conclusions.

BBMerge is a relatively new tool designed to reduce false-positive overlaps to near zero while maintaining a high rate of correct overlaps. It does this while also meeting the sub-goal of being the absolute fastest read merger in existence.

http://sourceforge.net/projects/bbmap/

Multi-'Omic' Analyses of the Dynamics, Mechanisms, and Pathways for Carbon Turnover in Grassland Soil

Cristina N. Butterfield* (butterfc@berkeley.edu)[1], Zhou Li[2], Peter Andeer[3], Susan Spaulding[1], David Burstein[1], Susannah Tringe[3,4], Brian C. Thomas[1], Robert Hettich[2], Trent R. Northen[3], Chongle Pan[2], and Jillian F. Banfield[1]

[1]University of California, Berkeley; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Lawrence Berkeley National Laboratory, Berkeley, CA; [4]DOE Joint Genome Institute, Walnut Creak, CA

To learn about how carbon cycling will be altered by climate change we are studying the roles of microorganisms in processing the largely vegetation-derived soil carbon in the period around the first Fall rainfall event, when soil-associated carbon fixed during Spring growth is rapidly metabolized. Genomic information was obtained for ten soil samples from two sites at 10- 20 cm and 30- 40 cm depth, two of which were collected before the first rainfall and the rest after. Metagenomic data were assembled and draft genomes, including dozens of partial to near complete genomes, were binned and reconstructed based on time series coverage analysis and tetranucleotide frequency using ggKbase and

emergent self-organizing maps. Phylogenetic analysis showed that the samples are dominated by archaea and several phyla of bacteria, including branches with closely related species radiating from terminal branches in a dandelion-like structure. Preliminary metabolite composition data clustered with depth; for example, simple sugars decrease with depth. Proteomics analysis revealed that corresponding sugar transporters are present and that the most abundant proteins in all samples are those involved in methylotrophy and ammonification. The results from these "omics" methods are quickly coming together to form a narrative about the dynamic microbial community and its processes below the grass root zone with respect to carbon flow during the Fall rainfall event. We anticipate these below ground terrestrial system feedbacks will impact the grassland ecosystem and the global carbon cycle.

http://ggkbase.berkeley.edu

Differences in potential bacterial growth rates as assessed by paired metagenome/metatranscriptome analyses

Barbara J. Campbell* (bcampb7@clemson.edu)[1], and David L. Kirchman[2]

[1]Department of Biological Sciences, Clemson University, Clemson, SC, 29634; [2]School of Marine Science and Policy, University of Delaware, Lewes, DE, 19958

Bacterial contributions to biogeochemical processes in the environment are not well understood in part because we know little about activity and growth rates. We used a paired metatranscriptomic and metagenomic approach to understand genus-level expression and abundance of different growth-related genes in bacterial communities. Samples were analyzed from Delaware coastal waters in March, May and December, 2010. Relative abundances of different genera were estimated via growth-related gene bins from unassembled metagenomes or after assembly of the 16S rRNA gene. Gene bins, segregated by gene type and phylogeny, clustered together according to sample, indicating a tight relationship between genus-specific gene abundance within each metagenome. However, relative abundances of taxa binned by corresponding transcripts or rRNA clustered independently, away from their equivalent genes. Statistical analyses indicated that various genera differentially expressed growth related genes depending on the season. Most taxa within the Bacteroidetes phylum had much higher mRNA:DNA ratios of growth related genes in March than May, even though their abundances did not significantly change between months. Other genera, such as *Rhodobacter* sp. and several Gammaproteobacteria, always had high ratios of mRNA:DNA compared to their low abundance, while *Pelagibacter* sp. within the SAR11 clade had low ratios with high abundances. This data indicates potential complex growth and activity patterns of marine bacteria that vary with phylogeny and season and suggest distinct controls of bacterial growth and abundance resulting in differential contributions to important biogeochemical processes.

Transposable element expansions in *Pleurotus ostreatus* and other fungi.

R. Castanera* (raul.castanera@unavarra.es)[1*], Antonio G. Pisabarro[1], J. Stajich[2], Lucía Ramírez[1].

[1] Public University of Navarre, Pamplona, Spain; [2] Department of Plant Pathology & Microbiology. University of California, Riverside, California, USA

Transposable elements are exceptional contributors to eukaryotic genome diversity. Their ubiquitous presence impacts nearly all species by causing mutations that disrupt genes, mediating chromosomal rearrangements and modulating gene expression. Due to the small size of fungal genomes, detection and comparative genomics of repeated sequences is more addressable than in plants or animals. We have developed a bioinformatics pipeline for TE annotation in fungi, based on a set of Python scripts for merging and mining the results of de novo and homology based searches of TEs in assembled genomes. We focused on the comparative analysis between strains of the same species and between species of the same genera (a total of 17 genomes), to investigate the basis of the differences in TE content. The results obtained uncovered large variations between fungal species (from 0.1% to 43.3 % of the genome comprised of mobile elements). A detailed analysis was performed on two strains of *Pleurotus ostreatus* demonstrating high variability in TE content (4.9% in PC9 strain vs. 9.9% in PC15 strain). This variability could be attributed to the differential expansion of Class I elements, primarily LTR retrotransposons in the *Gypsy* and *Copia* families. The analysis showed that the PC15 genome contains many more putatively active families of class I, while levels of Class II DNA elements are quite similar. The more active amplification of TEs in PC15 genome compared to PC9 generated a higher number of genes carrying mutations. The PC15 genome contains 41 interrupted genes that were intact in PC9, while PC9 has 9 mutated genes that were intact in PC15. In this sense, the average percentage of genes carrying TE fragments in *P. ostreatus* was 0.8%. Analyses of Repeat Induced Mutation patterns failed to uncover significant differences between both strains, and studies of TE methylation are ongoing. Our results suggest that the equilibrium between host genome defense and TE success critically regulates the expansion of TEs in fungi, and thus their impact on genome architecture and functionality.

The Microbiome of the Passalid Beetle (*Odontotaenius disjunctus*) and its Metabolic Potential for the Transformation of Lignocellulosic Materials

Javier A. Ceja-Navarro* (JCNavarro@lbl.gov)[1], Ulas Karaoz[1], Zhao Hao[1], Markus Bill[1], Mark Conrad[1], Richard A. White III[2], Mary S. Lipton[2], Joshua N. Adkins[2], Timothy Filley[3], Meredith Blackwell[4], **Jennifer Pett-Ridge[5], Eoin L. Brodie[1]**

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Pacific Northwest National Laboratory, Richland, WA; [3]Purdue University, West Lafayette, IN; [4]Louisiana State University, Baton Rouge, LA. [5]Lawrence Livermore National Laboratory, Livermore, CA.

Insects are the most abundant group of metazoans on Earth and have evolved to exploit a diverse array of environments, an ability that may be facilitated and enhanced by their gut-associated microbes. The wood-feeding beetle *Odontoatenius disjunctus* displays morphologically differentiated gut regions where stratified microbial communities degrade lignocellulosic materials and fix nitrogen. Our goal is to characterize the genomic potential stored in the microbiome of the passalid beetle for the optimization of lignocellulosic-dependent energy production processes.

Metagenomes were prepared from four gut regions of replicate beetles, sequenced, co-assembled and annotated. The ability to transform lignocellulosic materials by the passalid beetle was tested by Fourier Transform Infrared Spectroscopy and by measuring the fermentation products of plant polymer decomposition (H_2 and CH_4) with gas chromatography-isotope ratio mass spectrometry (GC-IRMS). The transformation of lignin after its passage through the gut was determined by ^{13}C-labeled tetramethylammonium hydroxide thermochemolysis. Proteomics and metabolomics were applied to detect expressed proteins and produced metabolites in each gut region.

After passage through beetles, oak wood showed cellulose and hemicellulose hydrolysis and lignin side chain oxidation. C and H stable isotope fractionation indicated that hydrogenotrophic methanogenesis dominates methane production the in beetle. These processes are mediated by a diverse range of bacteria, archaea and fungi in the metagenomes. Cellulose, starch, and xylan degradation genes were particularly abundant in the midgut (MG) and posterior hindgut (PHG). Genes involved in hydrogenotrophic production of methane and nitrogen cycling were more abundant in the anterior hindgut, confirming our previous phylogenetic and nitrogen fixing studies of compartmentalization in the passalid beetle gut. A filtered isolate database and predicted protein sequences from the metagenomes were used to search peptide spectra for proteome reconstruction. Proteomics supported metagenome observations, detecting expression of carbohydrate active enzymes principally in the MG and PHG, and nitrogenases and the enzymes involved in methane production in the AHG. Co-assembly and binning identified linked functions within microbial genomes including organisms with cellulosomes and a combined potential for cellulose, xylan and starch binding and hydrolysis.

Our multi-scale approach demonstrates that the passalid beetle harbors and expresses the functional potential to deconstruct lignocellulosic materials and produce H_2, CH_4 and potentially other biofuels. By studying such highly spatially and biochemically evolved polymer deconstruction and fermentation systems we hope to identify design criteria for improved lignocellulosic fuel production processes.

This research was supported by the LLNL Biofuels Scientific Focus Area, funded by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research Genomic Science program under FWP SCW1039. Work was performed at Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. Part of this work was performed at the Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 and at Pacific Northwest National Lab supported by the OBER/GSP under the Pan-Omics project. We also acknowledge EMSL support to E. Brodie through proposal #47988.

High-resolution analyses of bacterioplankton diversity during spring phytoplankton blooms in the German coastal North Sea

Meghan Chafee* (mailto:mchafee@mpi-bremen.de)[1], Christin Bennke[1,2], Bernhard M. Fuchs[1], Hanno Teeling[1], Rudolf I. Amann[1]

[1]Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany; [2]Leibniz Institute for Baltic Sea Research Warnemünde, Seestraße 15, 18119 Rostock, Germany

Marine microalgae dominate primary production in the oceans. Approximately 50% of the organic matter generated by these algae is decomposed and remineralized by heterotrophic bacteria as part of the marine microbial loop. This essential component of the global carbon cycle is particularly relevant during massive annual spring phytoplankton blooms occurring in temperate latitudes worldwide where they trigger rapid responses of specifically adapted clades of marine bacterioplankton. In particular copiotrophic members of *Gammaproteobacteria* and *Flavobacteriia* are metabolically equipped for the uptake and degradation of complex algal polysaccharides released during blooms, particularly upon bloom termination.

Microbial interactions between autotrophic and heterotrophic plankton community members during blooms are complex and dynamic, often resulting in swift successive changes in bacterioplankton composition. Distinct clades that remain rare most of the year expeditiously reach abundances as high as 40% of the total community in a matter of days. Complexity notwithstanding, these successions tend to be resilient with recurrent annual patterns. To determine the temporal stability and recurrence of

these patterns, sampling must occur at weekly to sub-weekly time scales over consecutive years. Owing to the lack of appropriate datasets, our current understanding of these processes is limited.

We have an ongoing intensively sampled time-series study from surface waters of the coastal North Sea at the long-term ecological research station (LTER) at the German island of Helgoland. After a 2009 pilot study [1], three additional years of data (2010-12) are currently being analyzed. In addition to CARD-FISH abundance data for spring bloom timepoints, bacterioplankton have been collected and size-fractionated from 153 free-living (0.2-3 µm) and 139 algae-attached (3-10 µm) samples with corresponding v4 Illumina tags for analyses of annual and seasonal patterns.

The first level of recurrence we find is a strong response of *Gammaproteobacteria* and *Bacteroidetes* (mostly *Flavobacteriia*) during spring as relative abundance of SAR11 rapidly declines. Spring bloom bacterioplankton contain more conditionally rare taxa (CRT, [2]) compared to samples collected outside spring showing the strong selective response, particularly of these two groups, to algal-derived organic compounds. The 5 most abundant CRTs contribute up to 27%, 30% and 42% of the total community abundance in 2010-12, respectively. Of these taxa, *Polaribacter* (*Flavobacteriia*) is most recurrent and comprises 44%, 60% and nearly 100% of the maximum CRT abundance. High-resolution decomposition using oligotyping [3] revealed a change in the dominant *Polaribacter* oligotype in 2010 versus 2011 and 2012 differing by only 0.8% identity, a change in taxa that would have been unresolved using conventional 16S rRNA tag clustering.

Our intensively sampled dataset contains a wealth of information that can be exploited using newly published methods aimed at achieving 'strain-level' demarcations of taxa. With this improved taxonomic resolution, we hope to expose ecological patterns often hidden by noise in such complex and dynamic environments that is further compounded by flaws in traditional OTU clustering methods. We ultimately aim to identify differential abundance of and direct interactions between microbes to more precisely define shared community niche spaces and responses to the environmental cues that drive these blooms.

References

1. Teeling, H. *et al.* Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–11 (2012).

2. Shade, A. *et al*. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. MBio **5**, e01371–14 (2014).

3. Eren, A. M. *et al.* Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, (2013).

Genome-wide analysis of the regulations of genes involved in Carbon Catabolite through expression Quantitative Trait Loci (eQTL) in *Coprinopsis cinerea*

Jinhui Chang* (tjchang@cuhk.edu.hk), Chun Hang Au, C. K. Cheng, Hoi Shan Kwan

The Chinese University of Hong Kong , Hong Kong SAR, PRC

Objective: Understanding the variation in regulation of carbohydrate-active enzymes is fundamental to the use of wood-decaying basidiomycetes in lignocellulose conversion. Our goal is to identify eQTLs of

lignocellulolytic enzymes in *Coprinopsis cinerea*, of which the genome harbors high number of Auxiliary Activities enzymes.

Methods: We sequenced *C. cinerea* reference strain Okayama 7#130 and its mapping partner #172 to develop a panel of SNP markers. Single spore isolates from crosses of the two strains were sequenced. The parental strains and the 46 single spore isolates were cultured on softwood-enriched sawdust to induce lignocellulolytic enzymes. RNAs from these cultures were sequenced. The RNA-seq results were aligned to the reference transcriptome using BWA. To assess the genetic contribution to growth variations among the 46 segregants, we mapped eQTL for the whole genome using a linear regression model.

Results: The expression level of the lignocellulolytic enzyme genes was correlated to individual SNPs. 1,332 cis-eQTLs and 36,139 trans-eQTLs ($P < 5 \times 10^{-5}$) were obtained. These eQTLs have provided us with a wealth of information that the expression of genes turning off the Carbon Catabolite Repression (CCR) is correlated with certain SNPs. Genes expressed in similar patterns with the lignocellulolytic enzyme genes across the 46 segregants were also clustered and analyzed, and SNPs are estimated for their contribution to explain the expression variances.

Conclusion: The eQTL approach has identified the transcriptional regulation that may contribute to the CAZymes expression. The results will be practically important to the enzyme production, which will benefit the bioethanol production from lignocellulose.

Biomimetic expression of cellulosomal genes of *Clostridium thermocellum* in *Bacillus subtilis* for enhanced cellulose degradation

Jui-Jen Chang* (lancecjj@gmail.com)[1,2], Mei-Yeh Lu[2], M. Anandharaj[2,3,4], Cheng-Yu Ho[5], Kenji Tsuge[6], Tsung-Yu Tsai[2], Yueh-Chin Wu[2], Chieh-Chen Huang[5], Wen-Hsiung Li[2]

[1]Department of Medical Research, China Medical University Hospital, Taichung, Taiwan; [2]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan; [3]Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, National Chung Hsing University and Academia Sinica, Taipei, Taiwan; [4]Graduate Institute of Biotechnology, National Chung Hsing University, Taichung, Taiwan; [5]Department of Life Sciences, National Chung Hsing University, Taichung, Taiwan ; [6]Institute for Advanced Biosciences, Keio University, Japan

Enzymatic conversion of lignocellulosic biomass into soluble sugars is a major bottleneck in the production of plant based biofuels. Several anaerobic organism cope these issue via multiple-enzyme complex system so called 'cellulosome'. However, it is very difficult to use these anaerobic organism for the production of biofuels. Hence we proposed a "Designer cellulosomes" concept for making an artificial cellulosome which can be used as a promising tool for the expression of cellulosomal enzymes in to industrial hosts. In this study, we used two types of "designer operon" via biomimetic approach according to a proteome-wide analysis of *Clostridium thermocellum* ATCC27405, which induced by avicel and cellobiose respectively. We selected eight celulosomal genes including one scaffolding protein gene (*cipA*), one cell-surface anchor gene (*sdbA*), two exoglucosidase genes (*celK* and *celS*), two endoglucanase genes (*celA* and *cel R*), and two xylanase genes (*xynC* and *xynZ*) of *C. thermocellum.* All these genes were cloned and co-expressed on the polycistronic operons in desired order via an ordered gene assembly in *Bacillus subtilis* method (OGAB). We successfully expressed the Type I and Type II designer cellulosomes in *B. subtilis* RM125 and demonstrated their cellulose-binding ability, thermo-stability, assembly of protein-complexes, and cellulytic activity. Our data show that, the variations in the

abundant enzyme activity among the Type I and Type II cellulosomes were successfully regulated by designer operon and this could be applied for different cellulose material saccharification, such as Avicel, filter paper and napiergrass. Overall, our findings suggest that xylanase could be the key enzyme during the initial stage of enzymatic saccharification of napiergrass.

Keywords: Designer cellulosomes, *Bacillus subtilis, Clostridium thermocellum*, biomimetic, Cellulosic bioethanol.

Expression Quantitative Trait Locus Mapping in *Populus*

Jay Chen* (chenj@ornl.gov)[1], Wellington Muchero[1], Priya Ranjan[1], David J. Weston[1], Xiaohan Yang[1], Sara Jawdy[1], Lee Gunter[1], Anna Lipzen[2], Erika A. Lindquist[2], Kerrie Barry[2], Jeremy Schmutz[2,3], and Gerald A. Tuskan[1,2]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; [2]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; [3]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA

Populus is one of DOE's "flagship" plant species that is of special interest as a biofuel feedstock. This JGI Community Science Program project aims to perform gene expression Quantitative Trait Locus (eQTL) mapping based on RNAseq data from 400 biological samples representing 200 progeny derived from a *Populus trichocarpa* × *Populus deltoides* pseudo-backcross pedigree. Each individual in this pedigree has been intensively characterized for phenotypic traits related to cell wall chemistry, biomass productivity and sustainable biomass production. These phenotypic data are complemented by fully resequenced parental genotypes of the pedigree and an ultra-dense genetic map with 3,568 single nucleotide polymorphism markers with an average marker distance of less than 0.75 cM. With the cumulative resources, we have identified large effect QTLs related to cell wall chemistry and biomass productivity. However, molecular characterization and functional validation of these QTLs have proven to be the major limiting factor. The resource generated in this JGI Community Science Program project will enable an unprecedented RNAseq-based eQTL analysis in *Populus* that facilitate the determination of genetic basis of transcriptional variation in the developing xylems. The resulting data will also facilitate the rapid characterization of high-value QTLs already mapped in the same pedigree and the identification of both genetic loci and genetic networks contributing to complex phenotypic traits of biomass quality, overall productivity and cell wall chemistry. In the short term, this project will provide novel knowledge to inform genetic improvement of *Populus* for cost-effective, sustainable biomass production, a topic that has become increasingly important in research areas highly relevant to DOE's missions.

Whole-genome transcriptomic analysis over the development of the mushroom, *Coprinopsis cinerea*

Xuanjin Cheng* (xcheng@cuhk.edu.hk), Jerome Ho Lam Hui, Yung Yung Lee, Patrick Tik Wan Law, Hoi Shan Kwan

School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China

Coprinopsis cinerea is a multicellular basidiomycete with a typical mushroom form that has limited edible value. In this study, we analyzed the gene expression profiles for *C. cinerea* development. We identified a 'developmental hourglass' over *C. cinerea* lifecycle: the 'young fruiting body' (YFB) is the stage that expresses the evolutionarily oldest transcriptome (lowest transcriptome age index, TAI) and

gives the strongest signal of purifying selection (lowest transcriptome divergence index, TDI). The expression of genes in 'information storage and processing' reaches the highest at the YFB, while genes involved in 'metabolism' become more active later. We found that two transitions, the mycelium-to-initials (MYC-to-INI) transition and the young fruiting body-to-mature fruiting body (YFB-to-MFB) transition, show most dramatic changes in gene expression. By cross-kingdom analysis, we found that all three kingdoms – animals, plants, and fungi – show underrepresented 'signal transduction mechanisms' at the waist stage of the hourglass. Overall, our study provides an overall picture on mushroom development and reveals a mutual strategy used by eukaryotes to incorporate evolutionary innovations.

Toward Comprehensive Genomics Analysis with De Novo Assembly

Jason Chin* (jchin@pacificbiosciences.com), Paul Peluso, David Rank

Pacific Biosciences, Inc.

Whole genome sequencing can provide comprehensive information important for determining the biochemical and genetic nature of all elements inside a genome. The high-quality genome references produced from past genome projects and advances in short-read sequencing technologies have enabled quick and cheap analysis for simple variants. However even with the focus on genome-wide resequencing for SNPs, the heritability of more than 50% of human diseases remains elusive. For non-human organisms, high-contiguity references are deficient, limiting the analysis of genomic features. The long and unbiased reads from single molecule, real-time (SMRT®) Sequencing and new *de novo* assembly approaches have demonstrated the ability to detect more complicated variants and chromosome-level phasing. Moreover, with the recent advance of bioinformatics algorithms and tools, the computation tasks for completing high-quality de novo assembly of large genomes becomes feasible with commodity hardware. Ongoing development in sequencing technologies and bioinformatics will likely lead to routine generation of high-quality reference assemblies in the future. We discuss the current state of art and the challenges in bioinformatics toward such a goal. More specifically, explicit examples of pragmatic computational requirements for assembling mammalian-size genomes and algorithms suitable for processing diploid genomes are discussed.

Ten years of collaboration with the JGI: a major impact on state-of-the-art in microbial methane oxidation

Ludmila Chistoserdova* (milachis@uw.edu) [1], David A.C. Beck[1,3], Maria E. Hernandez[1,2]

[1]Department of Chemical Engineering and [3]eScience Institute, University of Washington, Seattle, WA, USA; [2]Biotechnological management of resources network, Institute of Ecology A.C. Xalapa, Veracruz, Mexico

We will present a retrospective of our ten-year collaboration with the JGI, featuring five different CSP projects aimed at a better understanding of the microbial methane cycle in lake sediments. These five projects have focused on Lake Washington (Seattle), as a model for freshwater lake environments. We will highlight the major findings from each of these projects, and we will describe a progression in which each new sequence-based dataset has enabled novel hypotheses and concepts in our field of study. As a result, significant advances have been achieved in the fundamental understanding of microbial methane oxidation, especially in freshwater environments. One of the new concepts that resulted from our long-term Lake Washington project is the concept of cooperative behavior in microbial methane oxidation. In

this scenario, specific functional communities, rather than a single type of microbe are involved in the process, and the composition of such communities and specific partnerships among different functional guilds are controlled by environmental factors such as oxygen and nitrogen availability. Our data strongly suggest that speciation within each functional guild must be responsible for niche adaptation, also driven by environmental factors, so that specific ecotypes are tailored to form specific partnerships that possess competitive advantage in specific conditions, such as the placement within the methane/oxygen counter gradient. Our most recent CSP project (FY2015) is devoted to deciphering the mechanistic details of such partnerships. This ten-year collaboration with the JGI has resulted so far in 34 publications and helped secure over $6M in grant funds from the NSF and the DOE for studying methylotrophy in Lake Washington sediment.

Single Molecule, Real-Time Sequencing of Full-length cDNA Transcripts Uncovers Novel Alternatively Spliced Isoforms

Tyson A. Clark* (tclark@pacificbiosciences.com), Ting Hon, and Elizabeth Tseng

Pacific Biosciences, Menlo Park, CA

In higher eukaryotic organisms, the majority of multi-exon genes are alternatively spliced. Different mRNA isoforms from the same gene can produce proteins that have distinct properties such as structure, function, or subcellular localization. Thus, the importance of understanding the full complement of transcript isoforms with potential phenotypic impact cannot be underscored. While microarrays and other NGS-based methods have become useful for studying transcriptomes, these technologies yield short, fragmented transcripts that remain a challenge for accurate, complete reconstruction of splice variants.

The Iso-Seq™ protocol developed at PacBio offers the only solution for direct sequencing of full-length, single-molecule cDNA sequences to survey transcriptome isoform diversity useful for gene discovery and annotation. Knowledge of the complete isoform repertoire is also key for accurate quantification of isoform abundance. As most transcripts range from 1 – 10 kb, fully intact RNA molecules can be sequenced using SMRT® Sequencing (avg. readlength: 10-15 kb) without requiring fragmentation or post-sequencing assembly. Our open-source computational pipeline delivers high-quality, non-redundant sequences for unambiguous identification of alternative splicing events, alternative transcriptional start sites, polyA tail, and gene fusion events.

The standard Iso-Seq protocol workflow available for all researchers is presented using a deep dataset of full length cDNA sequences from the MCF-7 cancer cell line, and multiple tissues (brain, heart, and liver). Detected novel transcripts approaching 10 kb and alternative splicing events are highlighted. Even in extensively profiled samples, the method uncovered large numbers of novel alternatively spliced isoforms and previously unannotated genes.

Genetic control of plant root colonization by the biocontrol agent, *Pseudomonas fluorescens*

Benjamin J. Cole* (bjcole@lbl.gov)[1], Meghan Fletcher[2], Jordan Waters[3], Kelly M. Wetmore[3], Matthew J. Blow[1], Adam M. Deutschbauer[3], Jeffery L. Dangl[2], and Axel Visel[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA; [2]University of North Carolina, Chapel Hill, Chapel Hill, NC; [3]Energy Biosciences Institute, Berkeley CA

Plant growth promoting rhizobacteria (PGPR) are a critical component of plant root ecosystems. PGPR promote plant growth by solubilizing inaccessible minerals, suppressing pathogenic microorganisms in the soil, and directly stimulating growth through hormone synthesis. *Pseudomonas fluorescens* is a well-established PGPR isolated from wheat roots that can also colonize the root system of the model plant, *Arabidopsis thaliana*. We have created barcoded transposon insertion mutant libraries suitable for genome-wide transposon-mediated mutagenesis followed by sequencing (TnSeq). These libraries consist of over 10^5 independent insertions, collectively providing loss-of-function mutants for nearly all genes in the *P.fluorescens* genome. Each insertion mutant can be unambiguously identified by a randomized 20 nucleotide sequence (barcode) engineered into the transposon sequence. We used these libraries in a gnotobiotic assay to examine the colonization ability of *P.fluorescens* on *A.thaliana* roots. Taking advantage of the ability to distinguish individual colonization events using barcode sequences, we assessed the timing and microbial concentration dependence of colonization of the rhizoplane niche. These data provide direct insight into the dynamics of plant root colonization in an *in vivo* system and define baseline parameters for the systematic identification of the bacterial genes and molecular pathways using TnSeq assays. Having determined parameters that facilitate potential colonization of roots by thousands of independent insertion mutants in a single assay, we are currently establishing a genome-wide functional map of genes required for root colonization in *P.fluorescens.* Importantly, the approach developed and optimized here for *P.fluorescens>A.thaliana* colonization will be applicable to a wide range of plant-microbe interactions, including biofuel feedstock plants and microbes known or hypothesized to impact on biofuel-relevant traits including biomass productivity and pathogen resistance.

Spatially-Resolved Metabolic Cooperativity Within Dense Bacterial Colonies

John Cole* (cole15@illinois.edu)[1], Lars Kohler[2], Jamila Hedhli[3], & Zaida Luthey-Schulten[1,2]

Departments of [1]Physics, [2]Chemistry, & [3]Bioengineering, University of Illinois, Urbana-Champaign

The exchange of metabolites and the reprogramming of metabolism in response to shifting microenvironmental conditions can drive subpopulations of cells within colonies toward divergent behaviors. Understanding the interactions of these subpopulations—their potential for competition as well as cooperation—requires both a metabolic model capable of accounting for a wide range of environmental conditions, and a detailed dynamic description of the cells' shared extracellular space. Here we show that a cell's position within an in silico Escherichia coli colony grown on glucose minimal agar can drastically affect its metabolism: "pioneer" cells at the outer edge engage in rapid growth that expands the colony, while dormant cells in the interior separate two spatially distinct subpopulations linked by a cooperative form of acetate crossfeeding that has so far gone unnoticed. Our hybrid simulation technique integrates 3D reaction-diffusion modeling with genome-scale flux balance analysis (FBA) to describe the position-dependent metabolism and growth of cells within a colony. Our results are supported by imaging experiments involving strains of fluorescently-labeled E. coli. The spatial patterns of fluorescence within these experimental colonies identify cells with upregulated genes associated with acetate crossfeeding and are in excellent agreement with the predictions. Furthermore, the height-to-width ratios of both the experimental and simulated colonies are in good agreement over a growth period of 48 hours. Our modeling paradigm can accurately reproduce a number of known features of E. coli colony growth, as well as predict a novel one that had until now gone unrecognized. The acetate crossfeeding we see has a direct analogue in a form of lactate crossfeeding observed in

certain forms of cancer, and we anticipate future application of our methodology to models of tissues and tumors.

trichoCODE: A MODAL GENOME ANNOTATION PIPELINE FOR GOOD FUNGI FROM THE GENUS *TRICHODERMA*

Irina S. Druzhinina* (irina.druzhinina@tuwien.ac.at)[2], Dongqing Yang[1], Christian P. Kubicek[2], Igor Grigoriev[3], Alexey Kopchinskiy[2], Jian Zhang[1], Lea Atanasova[2], Ruifu Zhang[1], Qirong Shen[1],

[1]Jiangsu Key Lab for Organic Waste Utilization and National Engineering Research Center for Organic-based Fertilizers, Nanjing Agricultural University, Nanjing, China; [2]Microbiology Group, Research Area Biotechnology and Microbiology, Institute of Chemical Engineering, Vienna University of Technology, Vienna, Austria; [3]JGI, Walnut Creek, CA, USA

Trichoderma (Hypocreales, Ascomycota) - a large and relatively well studies genus of mycotrophic fungi, - is mainly known because of several "good" species that are used as agents of biological control of plant pathogenic fungi (*bio*fungicides) and/or as *bio*fertilizers because they promote plant's growth and development. Moreover, one species, *T. reesei*, is used in biotechnology for production of cellulases, hemi-cellulases and as a cell factory for protein expression and biosynthesis of secondary metabolites.

The genomics of *Trichoderma* began almost a decade ago with first genomes sequenced and annotated in JGI. However in recent years numerous research institutions published genomes of *Trichoderma* strains making the total number of whole genome projects approaching 15 (March 2015). However the comparative analysis shows considerable incongruence in genomic properties of even closely related species or strains that have been sequenced and annotated using different gene prediction pipelines.

In this work we present *trichoCODE*, the modular genome annotation pipeline that includes a high quality training set for *ab initio* gene calling in *Trichoderma*. The pipeline is based on the fungal genome annotation protocol of the Broad Institute. It includes four stages for gene structure annotation: (i) preparation of the training set; (ii) training and prediction; (iii) combination and fusion of multiple outputs; (iv) updating of gene structures with UTR (if EST/RNA-Seq data provided). The novelty of the pipeline comes from the fact that it is semi-automated, i.e. it allows the user to set up the best possible training set and to evaluate the annotation at several checkpoints.

The comparative analysis of genome annotations of the standard *Trichoderma* genomes, - *T. reesei, T. virens* and *T. atroviride* made using JGI, MAKER2 and *trichoCODE* pipelines will be demonstrated. Moreover the novel genome of *T. guizhouense* NJAU 4742 will be presented and compared with the genome of the closely related *T. harzianum* CBS 226.95. Our work also demonstrates the need to unify the annotations of previously published *Trichoderma* genomes, such as *T. longibrachiatum* and *T. hamatum* annotated with different pipelines and with no use of a *Trichoderma*-specific training sets.

Although prepared for *Trichoderma*, *trichoCODE* pipeline is open for any other organism as it is training-set dependent.

Using controlled selective pressure for targeted evolution of high-lipid content algal strain Chlamydomonas reinhardtii

J. Fedorko* (Fedorko.j@czechglobe.cz), D. Buzová, J. Červený

Department of Adaptation Biotechnologies, Global Change Research Centre AS CR, Drásov 470, 664 24, CZ

Green microalgae are among the most widely distributed microorganisms in the biosphere. They are significant contributors to global photosynthetic productivity and they are interesting for biotechnology due to their large variety of high value compounds accumulation and range of applications. To achieve profitable microalgae cultures for biotechnology it should combine two usually antagonistic properties: a fast growth and a high accumulation of specific target compounds.

Here, we focuse on development of advanced cultivation strategies and breeding methods applied on *a model* algae *Chlamydomonas reinhardtii, for* optimized production of lipids. For identification, isolation and subsequent selection of the optimal subpopulation with the high lipid content we use high-throughput fluorescence-activated cell sorting (FACS) in combination with imaging flow cytometry (IFC) on the cells stained with lipids specific fluorescent dye. We observed that post-sort cell viability is not negatively influenced by external parameters used during the sorting procedure (pressure, light quality and quantity, influence of the sorting electro-magnetic field, toxic effects of both fluorescent marker and fluidic system medium composition).

Additional outcome of developed methods and cultivation strategies will be set of genes that contribute to increased lipid accumulation and can be used as a clue for enhanced artificial genetic manipulation.

Keywords

Chlamydomonas reinhardtii, imaging flow cytometry, fluorescence-activated cell sorting, breeding, BODIPY, lipids, selection breeding

The use of 16S itags and metagenomics to assess bacteria-driven biogeochemical processes along a freshwater estuary to pelagic gradient

Masanori Fujimoto* (masafuji@umich.edu) [1], Ryan S. Mueller[2], Ann McCarthy[1], Henry A. Vanderploeg[3], and Vincent J. Denef [1]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109; [2] Department of Microbiology, Oregon State University, Corvallis, OR 97331; [3] NOAA Great Lakes Environmental Research Laboratory, Ann Arbor, MI 48108

Freshwater systems cover only 1% of the Earth's surface, but *emit* a net amount of CO_2 to the atmosphere that is similar to the net *uptake* by the oceans, largely due to subsidies of terrestrial carbon. Bacteria are key to terrestrial organic carbon transformation, yet the key taxa and genes are poorly characterized compared to those in marine systems. In this project, complementing NOAA's long term food web research program, 24 water samples were collected from Lake Michigan (at offshore and near-shore stations) and adjacent Muskegon Lake (a freshwater drowned river mouth estuary of Lake Michigan) during Spring, Summer, and Fall 2013 at the water surface and lake bottom. Replicate water samples were filtered to differentiate bacterial communities that were associated with particles (20 µm – 3 µm) from the free living (3 µm – 0.22 µm). DNA and RNA were extracted simultaneously from each filtrate. The 16S V4 itags from DNA and cDNA as well as metagenomic data were generated by JGI using Illumina sequencing. For 16S itags analysis, paired quality reads were merged in JGI itagger pipeline and were further analyzed in mothur for bacterial composition, alpha and beta diversity.

The particle-associated fraction consistently had a larger number of OTUs than the free-living fraction ($p<0.001$) and spring communities were more diverse than the other seasons ($p<0.001$). Bray-Curtis dissimilarity based PCoA plots revealed that the free living bacterial communities differed from particle-associated bacterial communities regardless of season and lake ($p = 0.001$, ANOSIM). Aquatic bacterial community structures based on DNA sequences differed from those based on cDNA sequences, though mostly due to shifts in relative abundance of the same taxa. Spring bacterial communities differed significantly from summer and fall ($p = 0.001$, ANOSIM). Depth (surface vs. bottom) of the water column had a significant effect on bacterial community composition during the summer stratified period ($p=0.001$, ANOSIM). Microbial communities also differed between Lake Michigan and the Muskegon Lake estuary ($p=0.05$ and $p=0.001$ for spring and summer/fall, respectively). The significant differences in community composition suggest that bacteria-driven biogeochemical processes differ between free-living and particle-associated communities, between seasons, depth, and along the estuary to pelagic gradient. A first analysis to determine how different microbial composition may lead to differences in biogeochemical cycling focused on a taxon that accounted for up to 25% of all reads in the deep offshore waters during stratification. This taxon belongs to the phylum Chloroflexi and is closely related (99.6% full 16S rRNA gene (reconstructed with EMIRGE) sequence similarity) to CL500-11, which was first found in the hypolimnion of Crater Lake, Oregon. This taxon is emerging as an endemic species to deep lakes and is a freshwater sister group to the marine group SAR202. We reconstructed a population genomic dataset for this taxon using Lake Michigan metagenomic data and determined this organism to be an aerobic, motile, likely psychrophilic, heterotroph capable of cellulose degradation. It also carries a bacteriorhodopsin gene, which may benefit this organism during the mixed period, when it becomes more abundant in the surface than deep waters.

Synthetic Biology in a Chip: A digital microfluidic platform for cell transformation, culture and expression

Philip C. Gach* (pcgach@lbl.gov) [1, 2], Steve C.C. Shih,[1, 2] Jess Sustarich,[1, 2] Jay D. Keasling,[1,3,4,5,6,7], Nathan J. Hillson,[1,3,8] Paul D. Adams,[1,3,4] and Anup K. Singh[1, 2]

[1] Technology Division, Fuels Synthesis Division, Joint Bioenergy Institute (JBEI), Emeryville, California 94608, United States; [2] Applied Biosciences and Engineering, Sandia National Laboratories, Livermore, California 94550, United States; [3] Physical Biosciences Division, Lawrence Berkeley National Lab, Berkeley, California 94720, United States; [4] Department of Bioengineering, University of California, Berkeley, California 94720, United States; [5] Department of Chemical & Biomolecular Engineering, University of California, Berkeley, California 94720, United States; [6] QB3 Institute, University of California, Berkeley, California 94720, United States; [7] Synthetic Biology Engineering Research Center, Emeryville California 94608, United States; [8] DOE Joint Genome Institute, Walnut Creek, California 94598, United States

Synthetic biology experiments require optimization of pathways consisting of many genes and other genetic elements and given the large number of alternatives available for each element, optimization of a pathway can require very large number of experiments. Currently, these experiments are done manually using fairly large amount of costly reagents per experiment making the process very expensive, extremely slow and irreproducible. We have developed a digital microfluidic platform that uses aqueous droplets suspended in an oil phase as discrete reaction chambers (or incubators) to carry out molecular biology steps. It integrates and automates all critical steps of transformation and culture including plasmid addition, transformation by heat-shock, addition of selection medium, culture and assay in one chip. Small volume droplets (~200nL) reduce reagents consumption by an order of

magnitude and also provide more efficient mixing reducing the time required for various steps. Thermal regulation was achieved by positioning three Peltier heaters below the device to provide the 4°C, 30°C, 37°C and 42°C temperatures employed during chemical heat-shock. The flexibility of digital microfluidics and peltier modules affords quick optimization of heat-shock parameters. Utilizing optimized procedures, a 0.2 μL droplet composed of 100 chemically competent DH5α cells was initially mixed with 0.2 μL GFP plasmid DNA at 4°C, moved to the 37°C region for heat-shock, merged with a media/antibiotic containing droplet at 4°C and allowed to incubate at 37°C. Following 24 hrs culture, all droplets possessed viable DH5α cells expressing GFP and the device generated transformation efficiencies up to $4.4 \times 10^6 \pm 0.2 \times 10^6$ CFU/μg DNA, similar to benchtop results. Additionally, complete cell heat-shock was afforded at a throughput of 0.1 droplets / s, though with reduced transformation efficiencies ($9.0 \times 10^5 \pm 0.9 \times 10^5$ CFU/μg DNA). Flexibility of this platform was demonstrated by transforming DNA plasmids into *E. coli*, *S. cerevisiae* and *A. niger* cells. Unlike previous microfluidic systems, the DMF transformation device allows completely automated bacteria transformation and assay. This technology will be of great utility for systematic interpretation of gene delivery methods and high-throughput screening of gene variants with minimal reagent requirements.

Genomics of endophytism

Gazis, Romina* (rgazis@clarku.edu)[1], Kuo, Alan[2], Riley, Robert[2], Hesse, Cedar[3,4], LaButti, Kurt[2], Lipzen, Anna[2], Lin, Junyan[2], Amirebrahim, Mojgan[2], Spatafora, Joseph[3], Grigoriev, Igor[2], Hibbett, David[1]

[1]Clark University, Biology Department, 950 Main Street, Worcester, Massachusetts 01610, U.S.A; [2]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek,California 94598, U.S.A; [3]Oregon State University, Department of Botany and Plant Pathology, Corvallis, Oregon 97331, U.S.A; [4]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, U.S.A.

Surveys of horizontally transmitted fungal endophytes have exponentially risen in the last decade, but it is only until recently that genomic data have become available for this ecologically and taxonomically diverse group. Here, we report the genome of *Xylona heveae*, which has only been isolated as an endophyte and, until this publication, remained the only representative of the Xylonomycetes. We had two main objectives: (1) elucidate the phylogenetic relationships and diversity of the Xylonomycetes and (2) investigate how the genome content of *X. heveae* compares with the genomes of other Ascomycota with diverse lifestyles and nutritional modes. We performed a series of comparative analyses using *X. heveae* genomic data in addition to data from other 36 Ascomycota including saprotrophs, plant pathogens (biotrophs, nectrotrophs and hemibiotrophs), mutualists (lichens, beetle endosymbionts and mycorhizae), animal pathogens and endophytes (vertically and horizontally transmitted). We focused on genes that are known to be important in the host-fungus interaction interface and that presumably have a role in determining the lifestyle of a fungus. Using genomic data, we inferred the position of the Xylonomycetes as sister to the clade composed by the Lecanoromycetes and the Eurotiomycetes and discovered that *Symbiotaphrina* is a member of the Xylonomycetes. Through the mining of ITS databases, we found that the taxa *Trinosporium guianense* is sister to *Xylona* and represents a third genus in this class. The comparative analyses suggested that *X. heveae* might have a strict endophytic lifestyle as it contained a low number of enzymes needed for plant parasitism or to be a competitive saprotroph outside its hosts. Data from this study have allowed us to resolve the phylogenetic position of the Xylonomycetes, expand its diversity and start understanding the genomic background of the different types of fungal endophytism across the endophytic continuum.

Brachypodium distachyon pan-genome: large-scale characterization of core and dispensable genes in grasses

Sean P. Gordon* (seangordon07@gmail.com)[1], Bruno Contreras Moreira[2], David L. Des Marais[3], Wendy Schackwitz[1], Joel Martin[1], Shengqiang Shu[1], Ludmila Tyler[4], Kerrie Barry[1], Alan Schulman[5], Thomas Wicker[6], Hikmet Budak[7], Luis Mur[8], Mike Freeling[9], Thomas Jeunger[10], Pilar Catalan[11] and John P. Vogel[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA; [2]Estación Experimental de Aula Dei, Zaragoza, Spain; [3]Harvard University, Cambridge, MA; [4]University of Massachusetts, Amherst, MA; [5]University of Helsinki, Helsinki, Finland; [6]University Zurich, Zurich, Switzerland; [7]Sabanci University, Istanbul, Turkey; [8]Aberystwyth University, UK; [9]Plant and Microbial Biology, UC Berkeley, Berkeley, CA; [10]The University of Texas at Austin, Austin, TX; [11]Escuela Politecnica Superior de Huesca, Universidad de Zaragoza, Spain

Most plant species are represented by a single de novo assembled genome obtained via sequencing one or a few distinct individuals. However, an entire species cannot be characterized by a single individual nor is a species characterized by population genetic variant data derived from projecting divergent sequence onto the genome of a single reference line. Characterizaton of a species requires de novo sequencing, assembly and gene annotation of diverse individuals to identify both genes Genomes at the species level are dynamic, with genes present in every individual (core) and genes in a subset of individuals (dispensable) that collectively constitute the pan-genome. Using transcriptome sequencing of seedling RNA from 503 maize (*Zea mays*) inbred lines to characterize the maize pan-genome, we identified 8681 representative transcript assemblies (RTAs) with 16.4% expressed in all lines and 82.7% expressed in subsets of the lines. Interestingly, with linkage disequilibrium mapping, 76.7% of the RTAs with at least one single nucleotide polymorphism (SNP) could be mapped to a single genetic position, distributed primarily throughout the nonpericentromeric portion of the genome. Stepwise iterative clustering of RTAs suggests, within the context of the genotypes used in this study, that the maize genome is restricted and further sampling of seedling RNA within this germplasm base will result in minimal discovery. Genome-wide association studies based on SNPs and transcript abundance in the pan-genome revealed loci associated with the timing of the juvenile-to-adult vegetative and vegetative-to-reproductive developmental transitions, two traits important for fitness and adaptation. This study revealed the dynamic nature of the maize pan-genome and demonstrated that a substantial portion of variation may lie outside the single reference genome for a species.

present in every individual (core) and genes in a subset of individuals (dispensable) that collectively constitute the pan-genome. Here we present the pan-genome of *Brachypodium distachyon*. To our knowledge this is the first plant pan-genome based on complete de novo genomes and gene annotations for a large (54) set of maximally diverse inbred lines. The accuracy of our genomes and gene annotations allow us to identify a much larger set of novel genes than previous studies and clearly define the core and dispensable genes within this species.

de novo Genome Assembly of Eukaryotic Species Mixtures – An *in silico* Evaluation

Bastian Greshake* (bgreshake@googlemail.com)[1], Simonida Zehr[1], Francesco Dal Grande[2], Anjuli Meiser[2], Imke Schmitt[2], Ingo Ebersberger[1]

[1]Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany; [2]Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

Lichens, an association of a filamentous fungus and one to several algal or cyanobacterial photobionts, are a hallmark for the success of mutualistic symbioses involving eukaryotes. They can colonize extreme ecological niches, frequently act as pioneering organisms, and are a promising resource for novel bioactive substances of medical and economical relevance. Still, the full potential of lichens for evolutionary and biotechnological studies has not been tapped, mainly since comprehensive genomic data are lacking. Extending the collection of lichen genomes is not trivial as a separate sequencing of the closely interacting symbionts is often not possible. Genome skimming of the lichen metagenome is an obvious and cost-effective solution to rapidly broaden the data basis for genomic research on lichens. Here we address the questions how to best assemble genome skimming data from a eukaryotic species mixture, what pitfalls can occur, and at what quality one can expect to reconstruct the individual genome sequences from a given experiment.

For a comprehensive benchmarking of different assembly approaches we introduce the concept of simulated twin sets. These are *in silico* generated whole genome shotgun sequencing data resembling the outcome of a real world sequencing experiment. Starting from a genome skimming of the lichen *Lasallia pustulata* we simulated 11 twin sets based on the draft genomes of the lichenized fungus *Cladonia grayi* and its photobiont *Asterochloris sp.* To evaluate how coverage differences between the fungal and algal genomes affect the assembly outcome we varied the ratio of sequencing reads stemming from the two organisms in each of the 11 twin data sets. Six assemblers (*MIRA*, *Omega*, *sga*, *Velvet*, *MetaVelvet* & *SPAdes*), representing 2 different Overlap-graph based methods, 2 *de Bruijn Graph* based methods and 2 dedicated metagenome assemblers, were then benchmarked on these data sets.

Our results reveal a substantial variation in the quality of genome reconstructions by the individual assemblers. Differences become more pronounced with increasing bias in the coverages of the two genomes, with mainly the DBG assemblers, including the metagenome assembler MetaVelvet, suffering most from low and uneven coverages. Moreover, we show that the common practice of empirical assembly parameter choice to maximize the assembly N50 value can substantially interfere with completeness of genome reconstructions from metagenome skimming data. We highlight cases where this effect culminates in precluding an entire genome from the assembly. For our particular example, we could show that MIRA, an all-purpose overlap assembler consistently generates the best assemblies regardless of the coverage ratio for the two organisms. The resulting genome reconstructions facilitated the identification of almost all genes annotated in the original data.

IMG-ABC: An Atlas of Biosynthetic Gene Clusters to Fuel the Discovery of Novel Secondary Metabolites

Michalis Hadjithomas* (michalis@lbl.gov) [1], I-Min Amy Chen[2], Ken Chu[2], Anna Ratner[2], Krishna Palaniappan[2], Ernest Szeto[2], Jinghua Huang[2], T.B.K. Reddy[1], Peter Cimermančič[3], Michael A. Fischbach[3], Natalia N. Ivanova[1], Victor Markowitz[2], Nikos C. Kyrpides[1], Amrita Pati[1]

[1] Prokaryotic Super Program, DOE Joint Genome Institute, Walnut Creek, CA 94598 USA; [2] Biological Data Management and Technology Center, Computational Research Division Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720 USA; [3] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, CA USA

In the discovery of secondary metabolites, analysis of sequence data is a promising exploration path that remains largely underutilized due to the lack of computational platforms that enable such systematic approach on a large scale. In this work, we present **IMG-ABC** (https://img.jgi.doe.gov/abc) -- An **A**tlas of **B**iosynthetic gene **C**lusters within the Integrated Microbial Genomes (IMG) system to pave the way from

big data to small molecules. IMG-ABC harnesses the power of IMG's integrated environment for structural and functional genomic analysis for the investigation of biosynthetic gene clusters (BC) and associated secondary metabolites (SM). SMs and BCs serve as the two interconnected foci of this platform, each with a rich collection of attributes. As a unique feature, IMG-ABC incorporates both experimentally validated and computationally predicted BCs in genomes as well as metagenomes, uncovering BCs in uncultured populations and rare taxa. We also report the discovery, using this resource, of phenazine producing clusters for the first time in alphaproteobacteria, thus showcasing IMG-ABC's focused integrated analysis tools that enable the exploration of microbial secondary metabolism on a global scale. IMG-ABC strives to fill the long-existent void of resources for computational exploration of the secondary metabolism universe; its underlying scalable framework enables traversal of uncovered phylogenetic and chemical structure space – serving as a doorway to a new era in the discovery of novel molecules.

A workflow for the analysis of contigs from the metagenomic shotgun assembly of SMRT® Sequencing data

Richard J. Hall* (rhall@pacificbiosciences.com)[1], Meredith Ashby[1], Brett Bowman[1], Ken Dewar[2], Cheryl Heiner[1], Nikoleta Juretic[2], Sudeep Mehrotra[2], Jessica Wasserscheid[2]

[1] Pacific Biosciences, Menlo Park, CA 94025; [2] McGill University, Montreal, Quebec, Canada

The throughput of SMRT® Sequencing and long reads allows microbial communities to be analyzed using a shotgun sequencing approach. Key to leveraging this data is the ability to cluster sequences belonging to the same member of a community. Long reads of up to 40 kb provide a unique capability in identifying those relationships, and pave the way towards finished assemblies of community members. Long reads are highly valuable when samples are more complex and containing lower intra-species variation, such as a larger number of closely related species, or high intra-species variation.

Here, we present a collection of tools tailored for the analysis of PacBio® metagenomic assemblies. These tools allow for improvements in the assembly results, and greater insight into the complexity of the study communities.

Supervised classification is applied to a large set of sequence characteristics (e.g. GC content, raw read coverage, kmer frequency, and gene prediction information) and to cluster contigs from single or highly related species. Assembly in isolation of the raw data associated with these contigs is shown to improve assembly statistics. A unique feature of SMRT Sequencing is the availability to leverage simultaneously collected base modification / methylation data to aid the clustering of contigs expected to comprise a single or very closely related species. We demonstrate the added value of base modification information to distinguish and study variation within metagenomic samples based on differences in the methylated DNA motifs involved in the restriction modification system.

Application of these techniques is demonstrated on a mock community and monkey intestinal microbiome sample.

Ecological Niches of Uncultivated Freshwater Phyla Revealed through Comparative Genomics

Joshua J. Hamilton* (jjhamilton2@wisc.edu)[1], Shaomei He[1], Katherine D. McMahon[1,2]

Department of Bacteriology[1] and Department of Civil and Environmental Engineering[2], University of Wisconsin-Madison, Madison, WI, 53706

Microbes are critical players in all freshwater ecosystems, where they support essential ecosystem functions such as nutrient and carbon cycling. Because many of these microbes cannot be cultured, metagenomics approaches are increasingly being utilized to study the genetics and physiology of these communities. This work utilizes metabolic reconstructions to identify metabolic features and potential niches for members of two freshwater phyla: the numerically dominant Actinobacteria and the poorly-understood Verrucomicrobia.

We have obtained 35 Actinobacterial and 19 Verrucomicrobial genomes from metagenomes (GFMs) from temperate lakes in the United States and additional assemblies of 16 Actinobacterial single-cell amplified genomes (SAGs). The Verrucomicrobial genomes span four of the six subdivisions, and the Actinobacterial genomes span three of seven major freshwater lineages.

We performed metabolic reconstructions of each genome to identify pathways present in each freshwater tribe, which in turn suggest substrate utilization abilities and requirements. We also performed comparisons of COG (Clusters of Orthologous Group) and Pfam (protein family) abundance between tribes to identify differences in gene content, and identify potential ecological niches. These reconstructions reveal the Verrucomicrobia have a heterotrophic life style using plant and algal polymers as the main carbon source. Analysis of Actinobacterial genomes reveals specialization between lineages for substrate acquisition.

Large scale sequencing of Dothideomycetes provides insights into genome evolution and adaptation.

Sajeet Haridas* (sharidas@lbl.gov)[1], Pedro Crous[2], Manfred Binder[2], Joseph Spatafora[3], Igor Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA. USA; [2]CBS Fungal Biodiversity Centre, Utrecht, The Netherlands; [3]Oregon State University, Corvallis, OR. USA.

Dothideomycetes is the largest and most diverse class of ascomycete fungi with 23 orders 110 families, 1300 genera and over 19,000 known species. We present comparative analysis of 70 Dothideomycete genomes including over 50 that are as yet unpublished. This extensive sampling has almost quadrupled the previous study of 18 species and uncovered a 10 fold range of genome sizes. We were able to clarify the phylogenetic positions of several species whose origins were unclear in previous morphological and sequence comparison studies. We analyzed selected gene families including proteases, transporters and small secreted proteins and correlate them to the varied lifestyles of these Dothideomycetes.

Nutrient Stoichiometry Drives Carbon Turnover and Microbial Community Composition in Mineral and Organic Soils Under Rice Cultivation

Wyatt H. Hartman* (whhartman@lbl.gov)[1], Rongzhong Ye[2], William R. Horwath[2], Susannah Tringe[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA, USA; [2]Department of Land, Air and Water Resources, University of California, Davis, Davis, CA, USA.

Project goals: We aim to characterize the microbial communities in wetland environments and evaluate their impact on carbon cycling, in order to develop strategies for optimizing carbon sequestration. Here we sought to determine the impact of soil carbon, nitrogen and phosphorus on carbon turnover and microbial community composition in an experimental rice field system.

Inundation of soils combined with emergent vegetation growth may enable high rates of soil carbon sequestration, and both rice cultivation and wetland restoration have been proposed as strategies to enhance carbon storage and reverse subsidence in degraded agricultural soils of the Sacramento-San Joaquin Delta. To isolate effects of wide variation in soils of the Delta on belowground C turnover, we studied a series of rice field trials as a model system with controlled vegetation and hydrology. Although soil C cycling may depend on the availability of carbon (C), nitrogen (N) and phosphorus (P), the role of P is particularly poorly understood due to a historical emphasis on soil C:N ratios informed by terrestrial nutrient limitation. To determine the effects of N and P availability on soil C turnover, we compared soil respiration over the course of a growing season in four adjacent rice fields with 5%, 10%, 20% and 25% soil C, each with control and N addition treatment plots (80 kg N/ha urea). Although soil P was not manipulated in parallel, prior work has shown soil P concentrations decline markedly with increasing soil C content. Soil CO_2 and CH_4 fluxes were monitored using static chambers at biweekly intervals during the growing season, and soils were collected at the end of the growing season for biogeochemical analysis and DNA extraction. Seasonal CO_2 fluxes (per m^2) were highest in 10% C soils (N:P =16:1), while soil N addition increased CO_2 flux and soil C turnover (seasonal CO_2 flux per unit soil C) in lower C fields (5% and 10% C), but not in higher C fields (20% and 25%). These patterns may be more clearly interpreted in light of shifts in soil N:P stoichiometry, which increased with soil C pools. Soil carbon turnover was greatest in mineral soils and inversely related to soil N:P ratios, suggesting progressive P limitation might limit both soil metabolism and its response to N at higher levels of soil C. Microbial community composition, based on 16S rRNA sequencing, was strongly influenced by soil C and pH along the gradient, but not by N additions as commonly observed in upland soils. Like soil carbon turnover, bacterial communities were also closely linked with soil N:P and inorganic P, and these relationships were significant even after accounting for covariance with soil C and pH. Functional gene content inferred from microbial phylogeny suggests substantial shifts in the potential to utilize carbon and phosphorus substrates between low and high C soils. Our results show that soil P availability and stoichiometry may affect microbial communities and their mediation of soil C turnover, even where primary producers appear limited by N.

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. This project was funded by the DOE Early Career Research Program, grant number KP/CH57/1 to Susannah Green Tringe, and was also supported by DOE JGI Community Sequencing Program.

Natural Arabidopsis thaliana accessions differentially recruit bacterial isolates to their root in complex synthetic communities

Sur Herrera Paredes* (sur00mx@gmail.com)[1,2,5], Derek S. Lundberg[1,3,4], Jose Macalino Esteban[1], Corbin D. Jones[1] & Jeffery L. Dangl[1,2,3,5]

[1]Department of Biology, [2]Bioinformatics and Computational Biology Curriculum, [3]Genetics and Molecular Biology Curriculum, The University of North Carolina at Chapel Hill; [4]Max Planck Institute for Developmental Biology; [5] Howard Hughes Medical Institute

Plants assemble a distinct bacterial community in their roots which is subset of the taxa present in soils. However, it remains to be determined how active is the plant role in modulating the assembly of the

root microbiome. Multiple rRNA gene surveys have reported statistical support for differential recruitment of bacterial taxa across plant genotypes, but the effects remain small and survey specific. One possibility, is that true genotype-specific differences are obscured by known artifacts associated from the use of Operational Taxonomic Units (OTUs). We constructed a complex synthetic community of 65 root isolates and tested whether these isolates differentially recolonize different Arabidopsis accessions as well as other plant species. The isolates were chosen to maximize diversity while allowing unambiguous identification of each one via 16S profiling. Our results show that there is little genotypic variation in the ability of bacterial strains to colonize roots of different genotypes, but there is strong quantitative variation in the relative abundance that each isolate reaches on different genotypes. Together, our results suggest that, for members of our synthetic community, the ability to colonize plant roots is a trait mostly determined by fast evolving bacteria while plant evolution mainly influences root colonization by modulating relative abundances of bacterial community members.

Systems Biology of phototrophic consortia: Towards a multi-scalar understanding of biological carbon and nitrogen cycling

Matthias Hess* (mhess@ucdavis.edu)[4,7], Erik R. Hawley[1], Hailan Piao[1], David E. Culley[2], Erika M. Zink[3], Young-Mo Kim[3], Lye M. Markillie[3], Samuel O. Purvine[3], Dehong Hu[3], Alice Dohnalkova[3], Miranda Harmon-Smith[4], Erika A. Lindquist[4], Stephen R. Lindemann[2], Tanja Woyke[4], Michael Barton[4], Davin Coleman-Derr[4], Tijana Glavina del Rio[4], Scott Lea[3], Galya Orr[3], Errol Robinson[3], Trent R. Northen[4], Peter E. Larsen[5], Kevin P. Keegan[4], Susannah Tringe[4], and Richard W. Castenholz[6]

[1] Washington State University, Richland, WA; [2] Pacific Northwest National Laboratory, Richland, WA; [3] DOE Environmental Molecular Sciences Laboratory, Richland, WA; [4] DOE Joint Genome Institute, Walnut Creek, CA; [5] Argonne National Laboratory, Lemont, IL; [6] University of Oregon, Eugene, OR; [7] University of California, Davis, CA

Cyanobacteria are photosynthetic prokaryotes that are found in most types of environments that are illuminated. Their global biomass has been estimated to be ~3×10^{14} carbon [4] and their key role in global carbon and nitrogen cycling has been widely accepted by the scientific community [5-7]. Despite the ecological importance and the significant efforts by DOE and the scientific community to advance our understanding of this important phylum, most of the research in this field is performed on pure cultures – which represent only a few percent of the organisms in an environmental sample [8, 9]. Very little is known about how members of the phylum cyanobacteria affect and respond to changes in complex biological systems. In the project selected here we selected photosynthetic consortia – based on their culture maintenance condition and their collection site – from a total over 1,200 samples that are currently grown in our laboratories – and subjected them to a diverse set of omics and imaging techniques that are available to the scientific community at DOE's Joint Genome Institute (JGI) and DOE's Environmental Molecular Sciences Laboratory (EMSL). The overall goals of this project is to determine the nutritional and environmental requirements of microorganisms that can currently only be grown in coculture and to define the molecular processes that facilitate carbon and nitrogen sequestration by phototrophic consortia and the individual organisms that contribute to these consortia. Here we present first results from this ongoing study that is part of the JG-EMSL Joint Program.

The DOE-JGI Microbial Annotation Pipeline

Marcel Huntemann* (mhuntemann@lbl.gov)[1], Konstantinos Mavrommatis[1], Natalia N. Ivanova[1], Natalia Mikhailova[1], Galina Ovchinnikova[1], Jim Tripp[1], Krishna Palaniappan[2], Ernest Szeto[2], I-Min Amy Chen[2], Victor Markowitz[2], Nikos C. Kyrpides[1], Amrita Pati[1]

[1] DOE Joint Genome Institute, Walnut Creek, CA, USA; [2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Here we present the DOE-JGI Microbial Annotation Pipeline (DOE-JGI MAP) version 2. All microbial genomes and metagenomes sequenced at the JGI or submitted to IMG are processed by the DOE-JGI MAP on a compulsory basis. The pipeline, which performs quality analysis, screening, sequence dereplication, structural annotation, and functional annotation, has been deployed in an integrated manner since the beginning of 2012 and has processed more than 27 Billion genes till date. The models for *de novo* structural annotation used in the DOE-JGI MAP were derived after exhaustive benchmarking and comparison with best practices. Similarly, functional annotation results in genes predicted by structural annotations being associated with Pfams, COGs, KOGs, KO numbers, KEGG pathways, IMG terms and pathways, EC numbers, TC numbers, MetaCyc pathways, phenotypes, and putative biosynthetic activity/associated natural products, and homologs in other genomes. The output of this pipeline is available through the IMG-ER systems, which allow genomic and metagenomic analysis and manual curation in a comparative context of thousands of genomes and metagenomes. Constant monitoring by on-site scientists and a user base of over 10,000 people ensures that the pipeline is up to date with the latest technologies in the field and delivers accurate products. The pipeline is available to users submitting their genomes via IMG's submission system (https://img.jgi.doe.gov/submit).

The DOE-JGI MAP is deployed on computational systems at NERSC (www.nersc.gov) and the functional annotation tools are implemented within the Hadoop framework. It can handle large unassembled metagenome datasets as efficiently as smaller assembled datasets.

Mining 1000 *Populus* Genomes: Core and Pan Genome, Correlotypes and the Hidden Microbiome

D. A. Jacobson* (jacobsonda@ornl.gov)[1], D. A. Weighill,[1,2] Gerald A. Tuskan[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, [2]Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN

The goals of this project include: 1) the determination of the core and pan genome of *Populus* and the detection of the endophytic communities present in the 1000 *Populus* genomes sequence data, 2) The determination of the genomic variants (SNPs, small and large INDELs) in the core and pan genome, 3) The calculation of correlations amongst all genome variants across 1000 *Populus* genomes, 4) The calculation of an exhaustive collection of sets of genome co-variants in *Populus* and 5) The development of a new multi- test agglomerative statistical method to associate sets of variants with complex phenotypes and microbiomic profiles.

Biological organisms are complex systems that are composed of pleiotropic functional networks of interacting molecules and macro-molecules. By the very nature of these pleiotropic networks many phenotypes are multigenic and, as such, will not follow classical Mendelian laws of inheritance and are thus less amenable to discovery by traditional linkage disequilibrium or many current genome-wide association methods.

Genome sequences have previously been generated for over 1000 genotypes of *Populus* (Wellington Muchero , Jessy Labbé , Priya Ranjan , Stephen DiFazio, 2014). We are mapping the resulting 100 billion reads to the reference genome (Tuskan et al., 2006) in order to confirm bi-allelic and multi-allelic SNPs as well as discover small and large INDELs. It appears that roughly 10 billion reads will not map to the *Populus* reference genome. We are treating this as a pooled metagenome and assembling it as such. Preliminary evidence suggests that the majority of these reads actually form the *Populus* pan genome. We can also detect thousands of different microbial and viral species in these reads, thus constituting the endophytic microbiome. We intend to use this metassembly to define a pan genome for *Populus.* We will then map the reads from individual samples to this overall pan genome in order to determine the full compliment of SNPs, large and small INDELs present across these 1000 genomes.

Complex phenotypes are the result of somewhat heterogeneous collections of genome variants. However, the effects of these variants are collectively subject to selective pressure and, as such, their co-occurrence can be seen as genome-wide correlations. We will calculate over 32 trillion correlations between all pairs of genome variants and use highly correlated variants to create a correlation network. Breadth-first searches and Markov Clustering will be used in order to determine an exhaustive collection of sets of variants that we will refer to as correlotypes. We will use correlotypes in combination with a new set-based agglomerative statistical method in order to associate collections of heterogeneous genomic variants with complex phenotypes. We will be testing these correlotype profiles against a range of phenotypic variables, including morphological, microbiomic and molecular profiles, resulting in thousands of phenotypes to test for complex genotypic associations.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., … Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)*, *313*, 1596–1604. doi:10.1126/science.1128691

Wellington Muchero , Jessy Labbé , Priya Ranjan , Stephen DiFazio, and G. A. T. (2014). The Influence of Climate Change on Insect Invasions in Temperate Forest Ecosystems.

Challenges and Opportunities for the World's Forests in the 21st Century, 81, 267–293. doi:10.1007/978-94-007-7076-8

A multi-pronged approach to the recovery of phylogenetic diversity and candidate phyla from Wilbur Hot Springs

Jessica Jarett* (jkjarett@lbl.gov)[2], **Jillian Steele**[1], Alex Copeland[2], Janey Lee[2], Scott Clingenpeel[2], Natalia N. Ivanova[2], Tanja Woyke[2]

[1]Mills College, Oakland, CA; [2]United States Department of Energy Joint Genome Institute, Walnut Creek, CA

Identifying and estimating the relative abundance of microbes in environmental samples is of key importance to many studies, but continues to be a challenge despite the development of a number of different methods. Candidate phyla are often especially problematic due to the lack of appropriate reference sequences. We compared the performance of several methods for community profiling, using an artificial mock community for validation and environmental samples from a system of hot springs as a test of more complex communities containing candidate phyla. Both 16S rRNA PCR amplicons (iTags) and shotgun metagenomes were sequenced for each sample. FOCUS was used to generate a taxonomic

profile of metagenomic reads based on *k*-mer composition relative to a reference database, and 16S rRNA sequences were obtained from iTags and extracted from metagenomic data using EMIRGE and an assembly-based pipeline. Overall, analysis methods differed in the amount of diversity they recovered, with EMIRGE yielding the fewest taxa and FOCUS the most. Both FOCUS and the assembly method detected false positives when applied to the mock community, whereas EMIRGE failed to detect several low-abundance taxa. iTags most accurately reflected the overall composition of the mock community. In samples from Wilbur, iTags recovered the most total diversity and the most candidate phyla, and EMIRGE the least. The amplification step in iTag sequencing likely increases the rate of detection of rare 16S sequences relative to methods such as EMIRGE and assembly that rely on isolating the small population of 16S rRNA reads from the larger metagenome pool. The use of all metagenome reads by FOCUS would avoid both PCR bias and the problem of rare 16S rRNA reads, but is hampered by the paucity of reference genomes for candidate phyla. These results highlight the utility of multi-pronged approaches to microbial community profiling, particularly when candidate phyla are targeted.

The distribution, diversity and function of predominant Thermoproteales populations in Yellowstone National Park

Zachary J. Jay* (zackary.jay@montana.edu)[1], Jacob P. Beam[1], D.B. Rusch[2], Alice Dohnalkova[3], B. Bodle[4], Mark A. Kozubal[1], R. Jennings[1], and William P. Inskeep[1]

[1]Department of Land Resources and Environmental Sciences and Thermal Biology Institute, Montana State University, Bozeman, Montana, USA; [2]Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana, USA; [3]DOE-Pacific Northwest National Laboratory, Richland, WA, USA; [4]Department of Biochemistry, Montana State University, Bozeman, Montana, USA

Metagenome and 16S rRNA gene sequence from geothermal habitats in Yellowstone National Park (YNP) has shown that members of the order Thermoproteales (phylum Crenarchaeota) are abundant in high-temperature (> 70 °C) Fe(III)-oxide microbial mats, sulfur sediments and filamentous 'streamer' communities. However, a comprehensive study of the diversity and distribution of Thermoproteales populations across different habitat types has not been conducted. Consequently, the goals of this study were to correlate the distribution of primary Thermoproteales genera with key geochemical parameters, and to compare specific functional attributes from curated genome sequence assemblies obtained across different habitat types with varying geochemistry. Curated sequence assemblies of *Thermocladium*-, *Vulcanisaeta*-, *Caldivirga*-, *Thermoproteus*-, and *Pyrobaculum*-like populations were characterized across numerous (~ 12) high-temperature sites. The distribution of these predominant phylotypes across different geothermal systems is controlled by key environmental variables including temperature, pH, sulfide and oxygen. *Thermocladium*, *Vulcanisaeta* and *Caldivirga* spp. were the primary Thermoproteales populations present in low pH (pH < 5) habitats, whereas *Thermoproteus* populations were found in mildly-acidic (i.e., pH 5 - 6) sulfur sediments, and *Pyrobaculum* populations were confined to higher-pH (i.e., pH > 6) sulfur sediments and/or filamentous 'streamer' communities common in alkaline siliceous springs. Metabolic reconstruction and comparative genomics across different genome assemblies of these populations suggest that they are primarily chemoorganotrophs and contain numerous amino acid and peptide transporters. The majority of these Thermoproteales populations have the metabolic capacity to synthesize requirements for vitamins, cofactors, amino acids, and/or nucleotides. The primary mechanisms of energy conservation (i.e., electron transfer) vary across different populations and correlate with key environmental parameters (e.g., sulfide versus oxygen, sulfate). Lower pH *Caldivirga* and *Vulcanisaeta* populations contain *dsrAB* genes likely used to encode proteins involved in sulfate reduction. Several different clades of dimethylsulfoxide

molybdopterin (DMSO)- encoding genes were analyzed, and are likely involved in the reduction of elemental sulfur, polysulfides, and/or arsenate. One such representative of the *Pyrobaculum* spp. was isolated from YNP and sequenced; the growth attributes of this organism are consistent with the putative function of these DMSO proteins. Several Thermoproteales populations contain heme-Cu oxidases and respiratory complexes indicative of aerobic metabolism, and correlate with the presence of dissolved oxygen measured *in situ*. Numerous intron sequences were also identified in Thermoproteales 16S rRNA genes, and phylogenetic analysis of several intron sequences revealed a strong geographical control on intron diversity. Many of the intron sequences within the 16S rRNA gene are in locations that are targets of 'universal' primers used in environmental surveys. Consequently, use of additional 16S rRNA gene primers targeting the Thermoproteales is advised for work in high-temperature systems. The distribution of different Thermoproteales populations in YNP is controlled strongly by pH and the availability of key electron acceptors, including reduced phases of sulfur, arsenate, sulfate and/or oxygen.

Identification of genomic elements required for uranium resistance by *Caulobacter crescentus*

Yongqin Jiao* (jiao1@llnl.gov)[1], **Mimi C. Yung**[1], Dan M. Park[1], Cindy A. Hoover[2], and Matthew J. Blow[2]

[1] Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA; [2] DOE Joint Genome Institute, Walnut Creek, CA

Caulobacter crescentus is a ubiquitous, aerobic bacterium known to survive in nutrient limited environments and to tolerate high levels of U(VI). We have recently demonstrated that *C. crescentus* is able to facilitate U(VI) biomineralization through the formation of U-P_i precipitates, highlighting is potential to be used for bioremediation in the environment. In order to gain a further understanding of how *C. crescentus* resist U, we employed a transposon mutagenesis screening approach (Tn-seq) to identify essential genomic elements that are required for U resistance. In our method, highly saturated transposon (Tn) mutagenesis was first performed to generate a library of 10^6 mutants. The library was then grown on solid agar plates containing U, Cd, or no metal control. Mutants surviving each exposure were subsequently sequenced via high-throughput Illumina sequencing at JGI to identify Tn insertion sites. Genes that accumulated fewer Tn insertions under U stress compared to the Cd or no stress controls were identified as genes specific for U tolerance. Using this method, we identified 15 genes potentially involved in U tolerance, which were subsequently tested through mutational analysis. Genes identified to be involved in U tolerance include TolC-like transporters $RsaF_a$ and $RsaF_b$, previously identified to be involved in S-layer protein transport, and stress factors CztR and CztA. The functions of these key genes for U resistance were examined and will be discussed. In particular, we performed in depth study examining the functions of the outer membrane transporters RsaF, and our results suggest that besides S-layer it can also export other antimicrobial compounds including U. Overall the Tn-seq results provided important insight into the various resistance pathways employed by *C. crescentus* for survival under U stress.

(Funding provided by the BER Early Career Research Program and work conducted under Contract DE-AC52-07NA27344. LLNL-ABS-667012.)

Bacterial communities and their activities are correlated with methane yields in New Zealand sheep.

Janine Kamke* (janine.kamke@agresearch.co.nz)[1], Sandra Kittelmann[1], Weibing Shi[2], Siva Ganesh[1], Jeff Froula[2], Edward M. Rubin[2], Graeme T. Attwood[1]

[1] AgResearch Ltd, Grasslands Research Centre, Tennent Drive, Palmerston North 4442, New Zealand; [2] Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Methane emissions from ruminant livestock account for approximately 14% of anthropogenic greenhouse gases and result from complex interactions between rumen microorganisms, including methanogenic archaea, bacteria and eukaryotes. Methane yield is a heritable trait in sheep and can differ significantly between individual animals thus allowing selection for low and high methane yielding flocks. In a recent study our group described differences in expression of the hydrogenotrophic methanogenesis pathway genes of archaeal communities in naturally high and low methane-yielding sheep and showed a strong up-regulation of these genes in the high methane yield phenotype. Here, we show that there are also differences at the bacterial community composition, gene abundance and gene expression levels between the low and high methane yield phenotypes. In samples of rumen contents from the low methane yield animals, metagenome and 16S rRNA gene amplicon sequencing data showed a significantly higher abundance of sugar-fermenting organisms of the genus *Sharpea*. Genes and transcripts predicted to encode sugar transport via the phosphotransferase system and parts of the galactose metabolism pathway were significantly more abundant in the low methane yield animals. The microbial communities of the high methane yield animals were characterised by a higher abundance of several taxa within the order Clostridiales and the Verrucomicrobia RFP12 clade, as well as a larger abundance of bacterial type III secretion genes and transcripts. Reassembly of type III secretion genes from our metagenome and metatranscriptome datasets and comparison with genome sequences from cultured organisms, showed some homologies to genes of organisms affiliated with the genus *Succinivibrio,* but the majority of the type III secretion genes could not be assigned to any known organisms. Together, these findings indicate that bacterial sugar fermentation, and type III secretory systems differently influence the conditions in the rumen and contribute to the methane yield phenotype observed in sheep. The exact mechanism by which this occurs is not yet clear, but these changes in bacterial metabolism are consistent with our previous hypothesis of rumen volume and hydrogen concentrations being the main drivers of methane yield in sheep.

Comparative genome analysis of 25 novel strains of planktonic *Flavobacteriia* isolated during North Sea spring phytoplankton blooms reveal distinct glycan degradation niches

Lennart Kappelmann* (lkappelm@mpi-bremen.de), Jens Harder, Bernhard M. Fuchs, Rudolf I. Amann, Hanno Teeling

Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Bremen, Germany

The *Flavobacteriia* constitute a broad class within the *Bacteroidetes* phylum. Marine representatives are abundant in epipelagic oceanic and coastal waters and benthic habitats and can reach up to 30% of the bacterial population [1, 2]. Much like their well-studied human gut microbiota counterparts, marine *Bacteroidetes* thrive in copiotrophic habitats and possess well-adapted gene repertoires for the degradation of complex organic matter. Large amounts of complex organic matter are produced and released into the environment during blooms of marine microalgae. Such algae blooms can trigger

successive blooms of distinct bacterioplankton clades that are often dominated by *Flavobacteriia*. In 2009 we performed a detailed study of such an event in the North Sea and showed that successively blooming *Flavobacteriia* clades exhibited distinguishable substrate spectra. This was most evident from distinct gene frequencies and expressions of carbohydrate-active enzymes (CAZymes) and corresponding transporters, in particular TonB-dependent transporters (TBDT) [3], suggesting a pronounced niche partitioning with respect to the degradation of algal polysaccharides.

Respective genes in *Bacteroidetes* are often co-located in dedicated clusters called polysaccharide utilization loci (PULs). However, so far only few PULs could be linked to dedicated functions and thus the diversity and functions of *Bacteroidetes* CAZymes and PULs as well as their possible co-dependences within and in-between taxa remain largely unexplored.

We hypothesize that distinct ecotypes and their corresponding ecological niches manifest themselves as detectable patterns of genomic CAZyme and PUL gene inventories. Hence we isolated a large collection of *Flavobacteriia* during North Sea spring phytoplankton blooms [4, 5], 50 of which are draft sequenced at the DOE-JGI in the framework of the community sequencing program COGITO.

Analysis of the so far sequenced 25 strains have revealed two distinct ecotypes: *Flavobacteriia* with smaller, more streamlined genomes and few PULs target easily degradable algae polysaccharides such as laminarin, while *Flavobacteriia* with larger genomes with many PULs and sulfatases are capable of degrading more recalcitrant, highly branched and often sulfated algal polysaccharides such as fucans. The former furthermore often exhibit common characteristics such as genes coding for bacteriorhodopsin and higher peptidase:CAZyme ratios. These strains are likely free-living and feed on peptides during polysaccharide-depleted times (outside of algae blooms). In contrast, the *Flavobacteriia* with larger genomes have lower peptidase:CAZyme ratios, lack bacteriorhodopsin, but often possess mannitol dehydrogenases. Mannitol-capped polysaccharides occur in macroalgae, and these *Flavobacteriia* likely represent algae-colonizing species

These results corroborate initial findings of similar patterns within the *Polaribacter* genus [6]. We intent to extend these comparative analyses to new flavobacterial isolates as well as existing marine flavobacterial strains from contrasting habitats like open ocean isolates to refine further ecotypes and their corresponding PUL repertoires. Preliminary results revealed as yet undescribed conserved PULs in multiple flavobacterial strains. Some of these PULs display very high degrees of conservation, while others merely contain a set of conserved, characteristic genes complemented by further divergent genes, which might reflect the high *in situ* structural versatility of certain polysaccharides.

References

[1] Eilers, H., Pernthaler, J., Glöckner, F.O., Amann, R. (2000) Culturability and in situ abundance of pelagic bacteria from the North Sea. Appl. Environ. Microbiol. 66, 3044–3051.

[2] Gómez-Pereira, P.R., Fuchs, B.M., Alonso, C., Oliver, M.J., van Beusekom, J.E.E., Amann, R. (2010) Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. ISME J. 4, 472–487.

[3] Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F.D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K.H., Glöckner, F.O., Schweder, T., Amann, R. (2012) Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336, 608–611.

[4] Hahnke, R.L., Harder, J. (2013) Phylogenetic diversity of Flavobacteria isolated from the North Sea on solid media. Syst Appl Microbiol 36, 497-504

[5] Hahnke, R.L., Bennke, C.M., Fuchs, B.M., Mann, A.., Rhiel, E., Teeling, H., Amann, R. and Harder, J. (2014) Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. Syst Appl Microbiol. 10, 1462-2920

[6] Xing, P., Hahnke, R.L., Unfried, F., Markert, S., Huang, S., Barbeyron, T., Harder, J., Becher, D., Schweder, T., Glöckner, F.O., Amann, R.I., Teeling, H. (2014) Niches of two polysaccharide-degrading Polaribacter isolates from the North Sea during a spring diatom bloom. ISME J. 5

Diel Metagenomics, Metatranscriptomics, and Metaproteomics of Elkhorn Slough Hypersaline Microbial Mat

Ulas Karaoz* (ukaraoz@lbl.gov) [2], Jackson Z. Lee[1, 4], Angela Detweiler[1], Craig R. Everroad[1], Leslie Prufert-Bebout[1], Rhona K. Stuart[3], Mary S. Lipton, Eoin L. Brodie[2], Peter K. Weber[3], Brad M. Bebout[1], Jennifer Pett-Ridge[3]

[1]NASA Ames Research Center, Mountain View, CA. [2]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA. [3]Lawrence Livermore National Laboratory, Livermore, CA. [4]Current Affiliation: MIT Lincoln Laboratory, MA.

Microbial mats are amongst the most diverse microbial ecosystems on Earth, inhabiting some of the most inclement environments known, including hypersaline, dry, hot, cold, nutrient poor, and high UV environments. Photosynthetic microbial mats found in intertidal environments are stratified microbial communities with anoxic conditions at night, generating significant amounts of H2 and organic acids. The high microbial diversity of microbial mats makes for a highly complex series of ecological interactions. To address this challenge, we are using a combination of metagenomics, metatranscriptomics, metaproteomics, iTags and culture-based simplified microbial mats to study biogeochemical cycling (H2 production, N2 fixation, and fermentation) in mats collected from Elkhorn Slough, Monterey Bay, California.

To understand the variation in gene expression associated with the daytime oxygenic phototrophic and nighttime fermentation regimes in hypersaline microbial mats, a contiguous mat piece was sampled at regular intervals over a 24-hour diel period. Additionally, to understand the impact of sulfate reduction on biohydrogen consumption, molybdate was added as an inhibitor to a parallel experiment. Four metagenome and 12 metatranscriptome Illumina HiSeq lanes were completed for samples collected from day / night, and control / molybdate experiments.

Our preliminary examination of gene expression in midday versus midnight samples (mapped using bowtie2 to reference genomes) has revealed several notable features, particularly relevant to the dominant mat-building cyanobacterium *Microcoleus chthonoplastes*.

M. chthonoplastes expresses several pathways for nitrogen scavenging, including nitrogen fixation.

Reads mapped to *M. chthonoplas*tes indicate expression of two starch storage and utilization pathways, a starch-trehalose-maltose-glucose pathway, and a UDP-glucose-cellulose-β-1,4 glucan-glucose pathway.

The overall trend of gene expression was primarily light driven up-regulation followed by down-regulation in the dark; much of the remaining expression profile appears to be constitutive.

Metaproteome analyses, conducted in collaboration with PNNL's Pan-Omics project (mapped using co-assembled metagenome), indicate upregulation of *Chloroflexi*-assigned proteins in the dark and upregulation *Cyanobacteria*-assigned proteins in the light

Co-assembly of quality-controlled reads from 4 metagenomes was performed using Ray Meta with progressively smaller K-mer sizes, with bins identified and filtered using principal component analysis of coverages from all libraries and a %GC filter, followed by reassembly of the remaining co-assembly reads and binned reads. A total of 20 near-complete (>80%) and an additional 50 minor genomic bins have been identified. Despite having relatively similar abundance profiles in each metagenome, this binning approach was able to distinctly resolve bins from dominant taxa, as well as sulfate reducing bacteria that are critical to our understanding of molybdate inhibition effects. Bins generated from this iterative assembly process are being used for downstream mapping of transcriptomic reads as well as isolation efforts for *Cyanobacteria*-associated bacteria.

Metagenomic analysis of poplar endophytes:

Mahsa Khorasani* (mkh2612@uw.edu), Sharon Doty

University of Washington

Study of the collective genomes of the members of a microbial community is important because it facilitates assessment of the microbiome regardless of culturability. Sequence-based metagenomic approaches could improve our understanding of microbial diversity and function. This in particular can be beneficial in the study of plant- microbe interactions. Bacterial endophytes that reside inside plants are able to enhance plant growth in several ways such as fixing nitrogen, producing phytohormones, increasing resistance against pathogens, detoxifying contaminants, solubilizing phosphate and generally reducing plant stress. More information about their community diversity, community dynamics, signaling and function could have a number of implications for agriculture, carbon cycling, biomass production for biofuels, forestry practices, etc. Development of increasingly fast, accurate, and inexpensive sequencing technologies, coupled with significant improvements in bioinformatics enable us to investigate these potential bio-fertilizers in their natural habitat and in association with each other and the plant.

In this work, we used a developed technique to isolate and amplify microbial 16S rRNA as well as NifH gene from polygenomic DNA isolated from leaves and stems of Populus trichocarpa The amplified bacterial 16S rRNA and NifH products were sequenced using the 2x300bp protocol on an Illumina miSeq. Analysis of the sequences obtained from 16S rRNA showed that Massilia, Sphingomonas, Methylobacterium and Burkholderia were the most common genera found in about 75-90% of the bacterial reads in both tissues. Sequences obtained from NifH gene analysis showed the existence of a diverse N-fixing bacterial community belonging to proteobacteria phylum.

Microfluidic Technology for High-Throughput Single Cell Analysis

Soohong Kim* (soohong@broadinstitute.org)[1, 2], Paul Blainey[1, 2]

[1]Broad Institute of Harvard and MIT, Cambridge, MA, USA; [2]Massachusetts Institute of Technology, Biological engineering, Cambridge, MA, USA

Next generation sequencing has revolutionized the study of biology. From evolutionary biology and environmental metagenomics to epigenomics and transcriptomics, the low cost to read nucleic acid sequences made this technology easily accessible and has transformed our understanding of practically every field in biology. However, while the cost to sequence large whole genomes has been dropping as the sequencing technology evolved, sequencing many samples remains challenging due to the primitive

state of whole genome sequencing sample preparation methods. Continual clinical resequencing of viral and pathogen genomes for the surveillance of antibiotic resistance gain, and single-cell genomics are studies that widely acknowledge the challenge posed by the limitations in the current sample preparation throughput. We developed an integrated cell sorting, culturing, and whole genome sequencing sample preparation system based on optical trapping and microfluidics. With this technology, we can sort single bacteria cells without distinctive markers or phenotypes; for example, sort two identical daughter cells from each other. The microfluidic chip also integrates the entire sample preparation process, from cell lysis to adapter tagging and barcoding, to lower the sample preparation cost by two orders of magnitude by reducing reagent consumption and achieving full automation. With our new technology, we have whole-genome-sequenced clinical isolates of 124 Pseudomonas Aeruginosa samples from six cystic fibrosis patients. In my poster, I will discuss (1) our analysis and interpretation of the clinical sequencing data and (2) our proposed plan for studying spontaneous evolution of bacteria at the single-cell level to remove biases that result from clonal interferences.

Ecological impact assessment of a biostimulation site on remediation of chlorinated ethylenes by metagenomic approaches

Nobutada Kimura* (n-kimura@aist.go.jp)[1], Saori Watahiki[1], Atsushi Yamazoe[2], Naohiro Noda[3], Yoh Takahata[4], Hideaki Nojiri[5],
Masao Fukuda[6]

([1]BPRI, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, [2]NITE, Tokyo, Japan, [3]BMRI, AIST, Tsukuba, Japan, [4]Taisei corp., Tokyo, Japan [5]Univ. of Tokyo, Tokyo, Japan, [6]Nagaoka Univ. of Tech., Nagaoka, Japan)

Bioremediation is a cost effective treatment system that involve the biological function of organisms to remove environmental pollutants from the contaminated site. To stimulate the indigenous bacteria that can degrade pollutant in the polluted site, various forms of rate limiting nutrients and electron acceptors such as carbon, nitrogen, and phosphorus are added to the contaminated site through injection. Therefore this process has possibility to impact on the ecological system around the bioremediation site. In this study, we attempted to assess the ecological impact assessment for biostimulation site on remediation of chlorinated ethylens by metagenomic appoaches using next generation-sequencing systems.

Metagenomic DNA from the soil samples that collected from the contaminated site for during of biostimulation was extracted to analyze the microbial community successions based on the 16S rRNA sequencing. To investigate the phylogenetical diversity of viable microbial communities in the contaminated sites, total RNA was also extracted from soil samples and cDNA sequencing of reverse-transcribed 16S rRNA genes were performed. Furthermore, this study applied both metagenomic and metatranscriptomic analysis to characterize gene expression of microbial community in the bioremediation site. Total DNA and cDNA were sequenced by Illumina Hi-seq2000 at a depth of 3 Gbp.

Taxonomic analysis showed bacteria were dominant in the both DNA and cDNA datasets and the bioremediation sites were dominated by phylum *Firmicutes* in both DNA and cDNA datasets after the injection of sodium lactate (electron donor). Gene expression annotation based on the COG（Clusters of Orthologous Groups）and KEGG showed the disagreement between the DNA and cDNA datasets. Gene expression focusing on the 16S rRNA genes revealed that dehalo-respiring bacteria, such as class *Dehalococcoidetes* were activated in the bioremediation site during of bioremediation. These studies

demonstrated that metagenomic and metatranscriptomic approaches could provide reference data for monitoring the ecological impact assessment of a biostimulation site on remediation of chlorinated ethylenes.

Array-derived Oligonucleotide Assembly with Static Tag Dial-Out PCR

Jason C. Klein* (jcklein@uw.edu)[1], Jerrod Schwartz[2], Jay Shendure[1]

[1]Department of Genome Sciences, University of Washington, Seattle, WA; [2]Google, Mountain View, CA

Over the past decade, DNA microarrays have decreased the cost of DNA oligonucleotide synthesis by approximately three orders of magnitude compared to column-based methods. However, the use of array-synthesized oligos in practice for gene synthesis is limited for several reasons: short synthesis lengths, high synthesis error rates, low yield, and the challenges of assembling long constructs from complex pools.

In 2012, our group developed Dial-Out PCR to overcome in part the last three limitations (Schwartz *et al.* Nature Methods 2012). This method incorporates degenerate tags on each oligo during PCR amplification, allowing for sequence verification and molecular retrieval. However, by utilizing degenerate tags, this approach requires ordering primers for every retrieval reaction, increasing cost and time. It also depends on random tags providing functional priming sites. Dropouts, wherein ~4% of targets were not retrievable, were possibly due to hairpin and dimer formations of the corresponding primers.

To increase the reliability and ease of Dial-Out PCR, we designed a set of 4637 static Dial-Out tags. Instead of relying on degenerate sequences, this set allows for in-house tags and corresponding retrieval primers that can be used for any Dial-Out reaction. We screened this set for properties such as GC, homopolymers and similarity to other tags. We then tested in silico every pair-wise Gibbs free energy, and identified the maximum clique of non-interacting tags. This set of 4637 tags was split into forward and reverse libraries with a complexity >5 million.

While Dial-Out PCR addresses several limitations of using array-synthesized oligos for gene synthesis, it does not address the challenge of assembling genes from short oligos. In order to produce sequence-verified 200-mers for the JGI gene synthesis pipeline, we optimized a “stitching” reaction that combines PCR assembly and tagging. We divided each of 104 targeted oligos into 5’ and 3’ fragments with 30 bp overlap. We utilized Uracil-Specific Excision Reagent to remove adaptor sequences, making this approach generalizable to any synthesis. After stitching the sets of 104 oligos, we identified sequence-verified constructs for 98/104 targets (94.2%). Of those, we retrieved 19/19 targeted oligos represented >4/160,000 (0.0025%) reads each. We sequenced 12/19 constructs with Sanger and verified 8/12 (66.7%). The failure modes included amplification of multiple oligonucleotides due to low sequencing depth and a likely Sanger sequencing error in a homopolymer stretch. Overall, these data indicate that our “stitching” protocol can achieve 2-piece multiplex assembly while remaining compatible with static tag Dial-Out PCR. In a subsequent experiment, we assembled 2 x 416 oligos in one reaction, and identified sequence-verified assemblies for 325/416 (78.1%). We are currently testing multiplex stitching up to 2 x 6,000 oligos.

To decrease cost and labor of retrieving assembled libraries, we also tested multiplex retrieval of sequence verified constructs. By controlling for abundance in the sequenced library and identifying sets

that will limit background amplification, we were able to retrieve 10/10 sequence-verified constructs in a single retrieval reaction.

Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes

Manuel Kleiner* (manuel.kleiner@ucalgary.ca)[1,2], **Breck A. Duerkop**[1], **Lora V. Hooper**[1,3]

[1] Department of Immunology, University of Texas Southwestern Medical Center, Dallas 75390, TX, USA ; [2] Energy Bioengineering and Geomicrobiology Research Group, Department of Geoscience, University of Calgary, Calgary T2N 1 N4, AB, Canada; [3] The Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas 75390, TX, USA.

BMC Genomics 2015, **16**:7, http://www.biomedcentral.com/1471-2164/16/7

The sequencing data for the artificial microbial community samples described here is available from the European Nucleotide Archive http://www.ebi.ac.uk/ena/data/view/PRJEB6941. This data should be well suited for benchmarking *in silico* methods (assembly, binning etc.) geared towards viruses.

Background. Viruses are a significant component of the intestinal microbiota in mammals. In recent years, advances in sequencing technologies and data analysis techniques have enabled detailed metagenomic studies investigating intestinal viromes (collections of bacteriophage and eukaryotic viral nucleic acids) and their potential contributions to the ecology of the microbiota. An important component of virome studies is the isolation and purification of virus-like particles (VLPs) from intestinal contents or feces. Several methods have been applied to isolate VLPs from intestinal samples, yet to our knowledge, the efficiency and reproducibility between methods have not been explored. A rigorous evaluation of methods for VLP purification is critical as many studies begin to move from descriptive analyses of virus diversity to studies striving to quantitatively compare viral abundances across many samples. Therefore, reproducible VLP purification methods which allow for high sample throughput are needed. We compared and evaluated four methods for VLP purification using artificial intestinal microbiota samples of known bacterial and viral composition.

Results. We compared the following four methods of VLP purification from fecal samples: (i) filtration + DNase, (ii) dithiothreitol treatment + filtration + DNase, (iii) filtration + DNase + PEG precipitation and (iv) filtration + DNase + CsCl density gradient centrifugation. Three of the four tested methods worked well for VLP purification. We observed several differences between methods related to the removal efficiency of bacterial and host DNAs and biases against specific phages. In particular the CsCl density gradient centrifugation method, which is frequently used for VLP purification, was most efficient in removing host derived DNA, but also showed strong discrimination against specific phages and showed a lower reproducibility of quantitative results.

Conclusions. Based on our data we recommend the use of methods (i) or (ii) for large scale studies when quantitative comparison of viral abundances across samples is required. The CsCl density gradient centrifugation method, while being excellently suited to achieve highly purified samples, in our opinion, should be used with caution when performing quantitative studies.

Categorization of the *DREB* gene family in common bean (*Phaseolus vulgaris* L.): orthologous relationships, protein motifs, chromosomal location and expression profiling

Enéas Ricardo Konzen* (erkonzen@cena.usp.br)[1], Gustavo Henrique Recchia[1], Fernanda Cassieri[1], Siu Mui Tsai[1]

[1] Center of Nuclear Energy for Agriculture, University of Sao Paulo, Piracicaba, SP/Brazil.

Common bean (*Phaseolus vulgaris* L.) is widely cultivated throughout the world, especially in developing countries. However, its cultivation has been limited by several environmental factors, such as abiotic stresses. This includes drought episodes, high soil salinity, nutritional deficiencies and temperature changes. Genomic approaches and detailed characterization of candidate genes are of prime importance to unravel and improve stress tolerance mechanisms. DREB (Dehydration Responsive Element-Binding) transcription factors seem to be promising candidates as they are involved in the downstream regulation of several other genes under abiotic stresses in plants. These proteins have an AP2/ERF conserved domain within their peptide sequences and some very specific motifs determine their binding ability. Nevertheless, few studies have been done with this gene family in common bean. Hereby, we performed a genomic categorization of all putative DREB proteins from sequences available on GenBank and Phytozome. A Hidden Markov Model for the AP2/ERF protein domain was obtained from Pfam v27.0, and it was searched against databases for common bean, *Arabidopsis thaliana* and *Glycine max*. A Neighbor-Joining tree was constructed with all sequences and, in total, 54 putative common bean DREB proteins were inferred based on orthologous relationships (with e-value < $1x10^{-5}$ in all comparisons). According to the phylogenetic tree, sequences were distributed within six subgroups (A-1 to A-6). Further analyses were performed to refine this categorization. First, the amino acid valine was 100% conserved at the 14^{th} position from the beginning of the AP2 domain, which may be important for binding-specificity, as prior studies have shown. Furthermore, proteins motifs were highly conserved within subgroups A-1 to A-6 and subgroup-specific motifs were identified. A Blast2Go annotation predicted all sequences have DNA-binding activity and involvement in abiotic stress responses. Chromosomal location of the putative *PvDREB* genes showed several genes in specific positions and clusters, as well as tandem duplications. Further examination of RT-qPCR profiles revealed under which stress conditions five selected genes (named *PvDREB1*, *PvDREB2A*, *PvTINY*, *PvDREB5* and *PvDREB6B*) were responsive. Plants of different common bean genotypes were subjected to different stress treatments (polyethylene glycol, high salinity, low temperature and abscisic acid) and expression analysis was performed with samples of roots, stem and leaves, collected at different time periods of stress. Each gene revealed different responses to all stresses and the variation was time, tissue and genotype dependent. The presence of SNP markers within the coding region of *PvDREB6B* influenced the expression profiling across a set of genotypes varying for stress tolerance. The categorization, expression profiling and SNP discovery in the common bean *DREB* genes are important for further studies to determine their role in abiotic stress tolerance mechanisms in common bean. We are working toward SNP, INDEL and SAP markers discrimination for all *DREB* genes and an association mapping between drought-related traits and these loci.

This work was supported by FAPESP, CAPES, CNPq.

The role of the Sphagnum microbiome in carbon and nutrient cycling in peatlands

Joel E. Kostka* (kostka@biology.gatech.edu)[1], Max Kolton[1], David J. Weston[2], Jon Shaw[3], Jennifer Glass[1], and Erik Lilleskov[4], Susannah Tringe[5]

[1]Schools of Biology and Earth & Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA; [2]Biosciences Division, Oak Ridge National Lab, Oak Ridge, TN; [3]Biology Department, Duke University, Durham, NC; [4]Northern Research Station, USDA Forest Service, Houghton, MI; [5]Joint Genome Institute, Walnut Creek, CA

Peatmosses (*Sphagnum spp.)* have a profound influence on the structure and function of peatland ecosystems that are estimated to harbor 20-30% of global terrestrial carbon pools. *Sphagnum* represents an ancient, early branching lineage of land plants that is largely responsible for the complex adaptive characteristics of northern peatlands. More carbon is sequestered in *Sphagnum* than any other plant genus on Earth, and yet the metabolic potential of its microbiome and bacteria-host interactions remain largely unknown.

This project will conduct iTag, draft metagenome, and draft metatranscriptome sequencing to characterize microbiomes associated with two representative peat moss species, *S. fallax* and *S. magellanicum*. Sampling efforts will be mainly conducted at two field sites sponsored by DOE BER, where *Sphagnum* is abundant: 1) the NGEE–Arctic site seeks to characterize the consequences of climate change on the vast amount of stored carbon in permafrost of the arctic tundra in Alaska, and 2) the SPRUCE site is assessing the response of northern peatland ecosystems to increases in temperature and exposures to elevated atmospheric CO_2 concentrations. To extend our *Sphagnum* sampling to the global scale, we will leverage samples collected from peatlands around the world via the Global Peatland Microbiome CSP project led by Erik Lilleskov. Since Lilleskov is also a coPI of the proposed work, he will work between the projects to provide the replicate samples of *Sphagnum* gametophytes under existing agreements. The project will also benefit from an ongoing CSP project led by Dave Weston (ORNL) and Jon Shaw (Duke U.), JGI (Jeremy Schmutz and others) on the genome sequencing of *S. magellanicum* and *S. fallax*, involving high-quality draft genome sequences and assemblies, supported by RNAseq to facilitate gene annotation, and SNP identification to enable construction of a high-resolution genetic map. We will determine the community composition and metabolic potential of commensal microbial populations associated with the population genetics of the same *Sphagnum* species.

This project seeks to answer the following questions:

1) Is the *Sphagnum* microbiome distinct from the microbial communities in the surrounding bulk peat?, 2) What is the metabolic potential of microbiome populations?, and 3) Does the microbiome vary with plant genotype and habitat?

Preliminary investigations were conducted using iTag sequencing of SSU rRNA genes on samples collected from the S1 bog, Marcell Experimental Forest, in northern Minnesota (SPRUCE site). In comparison to bulk peat samples from the S1 bog, the microbiomes contained a lower relative abundance of Acidobacteria and a higher relative abundance of Proteobacteria, Cyanobacteria, and Actinobacteria. We hypothesize that this reflects the influence of commensal taxa and the elevated pH (pH 5.5) of the *Sphagnum* endosphere relative to the bulk peat (pH 4 or less). A preliminary comparison indicated that *S. fallax* and *S. magellanicum* microbiomes were distinct. Results show that *S. magellanicum* microbiomes contain a higher relative abundance of Actinobacteria and Verrucomicrobia and a lower abundance of Cyanobacteria in comparison to *S. fallax* microbiomes. Further studies are needed to test these differences between plant genotypes and to explore the metabolic function of the microbial groups detected.

The genomic potential of bacterioplankton communities for algal polysaccharide degradation explored by deep metagenome sequencing in the German coastal North Sea

Karen Krüger* (kkrueger@mpi-bremen.de), Bernhard M. Fuchs, Rudolf I. Amann, Hanno Teeling

Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Celsiusstrasse 1, 28359 Bremen, Germany

As an essential component of the marine food web and the global carbon cycle, heterotrophic bacterioplankton decompose and recycle carbon produced by autotrophic plankton. This interplay is particularly marked during the massive annual spring phytoplankton blooms when large amounts of algal biomass are released into surface waters, triggering secondary blooms of heterotrophic bacteria. A substantial fraction of algal biomass is composed of polysaccharides of varying complexity, e.g. as storage or structural compounds. Distinct bacterioplankton clades are adapted for the uptake and degradation of these polysaccharides, in particular within the *Bacteroidetes* and *Gammaproteobacteria* [1]. These clades are characterized by high numbers of carbohydrate-active enzymes (CAZymes; [2]) that are responsible for the breakdown, modification and synthesis of carbohydrates and usually constitute 1-5% of a bacterial genome's genes. Planktonic bacteria capable of decomposing algal biomass are typically specialized on dedicated polysaccharide subsets and thus analysis of their CAZyme profiles provide clues about their respective glycan niches.

During the 2009 annual spring phytoplankton bloom nearshore the German North Sea island Helgoland, we observed a succession of distinct bacterial clades. Metagenomic and metaproteomic data suggested that this succession was caused by temporal fluctuations in the availability of distinct algae-derived substrates (bottom-up control) [3]. Differential frequency and expression of CAZymes and related transporters in bacterial clades suggested that algal substrates provided distinct glycan niches for bacteria mostly within *Flavobacteriia* and *Gammaproteobacteria*. In order to investigate whether this 2009 phenomenon was driven by stochastic forces or exhibits an annual recurring pattern, bacterioplankton samples for metagenomic and metaproteomic analyses have been collected for three consecutive years (2010-2012) on a monthly to sub-monthly timescale alongside corresponding 16S rRNA tag sequences collected at a weekly to sub-weekly timescale. DNA sequencing was performed at the DOE-JGI in the framework of the community sequencing program COGITO.

CAZyme annotation of 10 deeply sequenced metagenomes and corresponding metaproteomes was performed by custom integration of three different CAZyme search tools, which allowed for reproducible repertoire comparisons across datasets. The combination of these predictions with taxonomic classification of assembled metagenome contigs enabled identification of dedicated taxa with high polysaccharide degradation potential. A broad-level recurrent pattern in CAZyme profiles was detected in all three years with increasing degradation potential in bloom/post-bloom communities compared to pre-bloom time points. This increase in the number of CAZymes linked to polysaccharide degradation correlated with a shift from a SAR11-dominated community to one dominated by *Flavobacteriia* and *Gammaproteobacteria*. Furthermore, these two dominant bloom/post-bloom clades contained abundant glycoside hydrolase and polysaccharide lyase families with recurring patterns over the three investigated years. This pattern was more uniform on the class level compared to genus/clade level, suggesting that after phytoplankton blooms the recurrent response of the bacterioplankton community is likely more pronounced on the functional level than on the taxonomic level. This functional recurrence suggests that different clades occupy similar glycan niches. To further test this

hypothesis, we plan a denser temporal sampling of metagenomes using archived gDNA samples as well as an in-depth analysis of the functional roles of specific sub-populations from metagenomes.

References

[1] Buchan, A., LeCleir, G.R., Gulvik, C.A. and González, J.M. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nature Reviews Microbiology 12, 686-698.

[2] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42, D490-D495.

[3] Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F.D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K.H., Glöckner, F.O., Schweder, T. and Amann, R.I. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336, 608–611.

ETOP: Resources and protocols for generating high-quality genome assemblies.

David A. Kudrna* (dkudrna@email.arizona.edu), Dario Copetti, Jianwei Zhang, Seunghee Lee, Jayson Talag, Yeisoo Yu, and Rod A. Wing.

Arizona Genomics Institute, School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson AZ 85721, USA

BACKGROUND.
The Arizona Genomics Institute (AGI) has played significant roles in numerous genome projects over the past 15 yrs, including Asian and African rice and its 20 wild relatives, maize, *Brachypodiuum*, sugarcane, soybean and its wild relative, tomato, *Eucalyptus*, *Amaranthus*, barley, cacao, cotton, *Amborella* and *Eutrema salsugineum*. AGI's expertise is not limited to plants, and also includes model species like Drosophila (19 genomes), zebra finch, *Biomphalaria* and nurse shark, as examples. The majority of AGI's roles in these genome projects have centered on the generation of high-quality physical maps and BAC libraries with the ultimate goal of producing the highest quality reference genome sequences possible. AGI's philosophy is that the first genome of every important species under investigation should be of the highest quality (i.e. Reference sequence RefSeq, or near-RefSeq quality) which significantly affects/facilitates all subsequent downstream functional, population and evolutionary analyses.

CURRENT ETOP OBJECTIVES.
Produce high-quality, high-molecular weight genomic DNA in support of JGI projects and partners. Provide substrates for Brassicales taxa sequencing. Develop 96 well DNA isolation procedures for population re-sequencing and GBS-type studies. Investigate alternative methods of High Throughput DNA isolations.

Construct large-insert high-quality BAC libraries for JGI and partners for: *Oryza sativa* Kitaake/Ubi-Xa21; *Panicum hallii* HAL2, *Poplar deltoids* WV94.

Develop new software programs for the integration of sequence assemblies with BAC based Physical Map for *Bochera stricta*.

Develop robust protocols for large-insert Mate Pair and Long-range Linking Illumina library construction.

Develop NGS-based BAC end sequence methods.

Here we present a progress report on meeting the above objectives as well as future directions. AGI is happy to share protocols/expertise to help to meet the objectives of JGI/HA CSP and in house projects.

Whole-Genus Sequencing: 300 Aspergilli.

Alan Kuo* (akuo@lbl.gov)[1], Robert W. Riley[1], Alicia Clum[1], Asaf Salamov[1], Scott Baker[2], Blake Simmons[2], Mikael Andersen[3], Igor Grigoriev[1]

[1]Fungal Genomics Program, DOE Joint Genome Institute, Walnut Creek, CA; [2]Joint BioEnergy Institute, Emeryville, CA; [3]Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark.

Aspergillus is a ubiquitous and phenotypically diverse genus of filamentous Ascomycota, many of which play key roles as fermenters in food production, platforms for biotechnology and industrial production of enzymes and chemicals, plant and opportunistic animal pathogens, and agents of agricultural toxigenesis and biomass conversion for bioenergy. As part of a DOE Joint BioEnergy Institute initiative to characterize the entire genus, the JGI plans to sequence, assemble, and annotate the genomes of each of the ~300 species of the genus *Aspergillus*. To accomplish this massive task in a timely manner without sacrificing quality, we have sought to streamline our existing processes as well as explore alternative technologies, especially assembly and annotation of long PacBio sequencing reads. Over the past year we have released on MycoCosm the genomes of an additional 24 *Aspergillus* sp. with preliminary analyses of their phylogenies, secretomes, and secondary metabolism. The next tranche of 92 species is expected soon.

Shiitake mushroom genomics

Hoi Shan Kwan* (hoishankwan@cuhk.edu.hk), Chun Hang Au, Man Chun Wong, Lei Li, Yung Yung Lee, Qianli Huang, Wenyan Nong, Man Kit Cheung, Jinhui Chang, Xuanjin Cheng

Food Research Centre and School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China

We aim to understand the complex fruiting body development and evolution of the mushrooms. We sequenced the genome of the *L. edodes* monokaryon L54A. Over 13,000 protein-coding genes were predicted from the 40.2 Mb draft genome. Comparative analyses on genome sequences of basidiomycetes and ascomycetes revealed genes expanded in genomes of mushroom-forming fungi. Five functional categories, General function prediction only [R], Signal transduction mechanisms [T], Posttranslational modification, protein turnover, chaperones [O], Transcription [K] and Carbohydrate transport and metabolism [G], dominate in the expanded families. We examined kinome, ubiquitome, transcription factories and CAZy enzymes. AGC kinase subfamily are significantly expanded in mushroom-forming genomes. F-box and paracaspase domain-containing E3 like proteins were expanded in the mushroom-forming genomes and basidiomycota have significantly more BTB type E3 than the ascomycota genomes. We sequenced the genomes of 9 wild *L. edodes* strains to explore their relationships. Their morphologies and cultivation characteristics were investigated. With these data, we could start to link some phenotypes with the genotypes using a Genome-Wide Association Studies, GWAS, approach. We compiled the genome sequences of *L. edodes* and other fungi into an Ensembl-based platform with a battery of genomic tools. Our works have generated rich resources for genomics and transcriptomics of mushrooms.

Plant-Microbe Interfaces: Genome re-sequencing reveals a species-specific whole-gene deletion associated with *Populus-Laccaria* mycorrhizal symbiosis

Jessy Labbé* (labbejj@ornl.gov)[1], Wellington Muchero[1], Juan Wang[1], Olaf Czarnecki[1], Jin-Gui Chen[1], Priya Ranjan[2], Xiaohan Yang[1], Anthony Bryan[1], Sara Jawdy[1], Lee Gunter[1], Wendy Schackwitz[4], Joel Martin[4], **Gerald A. Tuskan**[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]University of Tennessee; Knoxville, TN; [3]West Virginia University; Morgantown, WV; [4]DOE Joint Genome Institute; Walnut Creek, CA

Mycorrhizal symbiosis between perennial plants and fungal associates has critical implications for diverse phenomena including global carbon, water and nutrient cycling, as well as agricultural and forestry productivity with limited inputs on marginal croplands. As such, characterizing the molecular genetics underlying such interactions holds tremendous potential in engineering biological systems for enhanced carbon sequestration and sustainable biomass production.

In the perennial bioenergy feedstock *Populus*, numerous studies have demonstrated the species-dependent colonization efficiency by the fungal symbiont *Laccaria bicolor*, with *P. trichocarpa* exhibiting high levels of mycorrhization compared to *P. deltoides*. This highly species-specific attribute of the interaction presents an opportune platform for the discovery of host genetic factors governing mycorrhizal interactions using inter-specific hybrids. To this end, we identified a major quantitative trait locus (QTL) contributing up to 60% of the phenotypic variance explained (PVE) in colonization of *P. trichocarpa* x *P. deltoides* F1 hybrids by *Laccaria*. Genome anchoring of this QTL using single nucleotide polymorphism (SNP) markers with known physical positions revealed its co-location with a region harboring tandemly repeated lectin-type receptor kinases. Alignment of the *P. trichocarpa* and *P. deltoides* re-sequenced parental genomes suggested major structural differences in this region including a whole-gene deletion event in *P. deltoides* involving a D-mannose lectin receptor kinase. Analysis of allelic effects of the indel revealed that individuals carrying a full copy of the gene-exhibited 2X more colonization by the fungal symbiont compared with individuals missing segments of the same gene. Further, we screened pure *P. trichocarpa* and *P. deltoides* natural variants to assess penetrance of the indel in the species' natural habitats. We could not detect a full copy of the gene in any of the 60 *P. deltoides* genotypes collected from diverse geographical origins in eastern United States whereas the gene was highly conserved in 673 re-sequenced *P. trichocarpa* genomes evaluated. Since D-mannose receptor kinases have been implicated in innate immunity and self-incompatibility responses, which require highly specific recognition of cells and microorganisms, we hypothesize that this indel polymorphism contributes substantially to the species-specificity observed in *Populus* interaction with *Laccaria*. Transgenic validation of putative effects of the receptor kinase on mycorrhization is currently underway and results of these analyses will be presented.

The Plant Microbe Interfaces Scientific Focus Area is sponsored by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research.

http://PMI.ornl.gov

Systems Biology Approach to Fracking for Environmental Monitoring

Regina Lamendella* (LAMENDELLA@juniata.edu)[1], Justin Wright[1], Nikea Ulrich[1], Christopher Grant[1], Maria Campa[2], and Terry Hazen[2]

[1] Juniata College, Huntingdon PA 16652; [2] University of Tennessee Knoxville, Knoxville, TN

Similar to most industrial activities, unconventional natural gas extraction (fracking) can lead to contamination and impact environmental health. In this Community Science Project we plan on using high-throughput sequencing to elucidate the microbial biodegradation pathways and adaptive strategies harnessed by microbial communities in response to potential fracking impacts. Characterizing the microbial consortia associated with fracturing fluids can lead to improvements in treatment and reuse of these fluids. Further, knowledge of the biodegradative capabilities of the microorganisms associated with hydrofrack operations and potential recipients (e.g., aquifers, surface waters, and streambed sediments) will enable predictions about the fate and longevity of compounds of environmental concern. The sequence data requested in this Project will be combined with more conventional molecular, geochemical and isotopic indicators of deep brines to provide a systems approach of assessing the impact of hydrofrack operations. Altogether this proposal will provide a clearer understanding of the natural microbial community's response to hydrofracking and will better inform *in situ* remediation strategies, as well as the persistence of microbial markers related to biodegradation.

Establishing a genome-wide sequence-indexed collection of grass mutants

Debbie Laudencia-Chingcuanco* (debbie.laudencia@ars.usda.gov)[3], **John P. Vogel**[1], and Richard Sibout[2]

[1]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598; [2]Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, Saclay Plant Sciences, France; [3]USDA-ARS WRRC, Crop Improvement and Genetics Research Unit, 800 Buchanan St. Albany CA 94710

Much work needs to be done to domesticate the grasses being developed as feedstocks for the nascent biofuel industry. Knowledge of the genes controlling traits relevant to biomass crops (e.g. yield, abiotic stress tolerance, disease resistance, cell wall composition) could be used to accelerate domestication and help overcome the inherent difficulties (e.g. polyploidy, self-incompatibility, difficulty evaluating yield traits) in breeding these new crops. The ability to identify and order plants with mutations in virtually any gene in the genome is a powerful research tool that can be used to accelerate research. The use of an appropriate model system is an efficient way to gain the knowledge about basic grass biology needed to serve as a foundation for the development of superior energy crops. Although Arabidopsis has been and continues to be an extremely valuable model system it is not suitable for the study of biological features unique to the grasses (e.g. cell wall composition and architecture, plant anatomy, development including intercalary meristems, secondary metabolic pathways, etc.). *Brachypodium distachyon* possesses all the biological attributes needed to be a modern model organism including simple growth requirements, fast generation time, small stature, small genome size and self-fertility. Numerous *B. distachyon* experimental resources and methods have been developed including: an outstanding reference genome sequence, high efficiency Agrobacterium-mediated transformation protocols, efficient crossing methods, a large germplasm collection, 22,000 sequence-indexed T-DNA lines, recombinant inbred lines, and microarrays, which have led to its widespread adoption as a model system for grass research. With the rapid decrease in the cost of DNA sequencing and the large number of mutations a single chemically mutagenized line can harbor, we will create a powerful reverse and forward genetic tool by sequencing a collection of 2,000 chemical and radiation mutants in the model grass *B. distachyon*. While the sequencing of the first 96 lines has just begun, we have already initiated collaborations with 40 laboratories that will characterize mutants or screen for phenotypes in this collection. Additional collaborators are welcome

A metagenomic approach to study flavin mediated public good dynamics in hydrocarbon resource environments.

Kateryna Levdokymenko* (levdokymenko@gmail.com)[1,2], Cameron R. Strachan[2,3], Anna J. Mueller[2] and Steven J. Hallam[1,2,3,4]

[1]Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, BC, Canada V5Z 4S6, [2]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada V6T 1Z3, [3]MetaMixis, Inc., Vancouver, BC, Canada V6T 1Z3, Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, BC, Canada V6T 1Z3, [4]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada V5T 4S6.

Emerging lines of evidence indicate that microbes form distributed networks of metabolite exchange based in part on public goods. These networks have the potential to drive the evolution of microbial lineages, and contribute to essential functions and services in natural and engineered ecosystems. However, experimental systems in which to evolve and perturb public good dynamics remain poorly constrained. Here, we use a functional metagenomic screening approach to recover abundant biosynthetic gene clusters with the potential to mediate microbial interactions in the environment. Specifically, 29 gene clusters involved in the production of riboflavin sourced from diverse microbial donor genotypes were recovered by functional screening from two fosmid libraries constructed from methanogenic communities enriched on hydrocarbons. Active clones were sequenced and riboflavin encoding gene cassettes were verified using transposon mutagenesis and cluster subcloning. Focusing on observed relationships between mobile genetic elements, metabolite secretion patterns and gene frequency distributions, we suggest a role for riboflavin as a public good in hydrocarbon resource environments. We tested this role by developing model co-culture systems in *E. coli* in which fosmid pools containing one or more riboflavin over-expressing clones demonstrated fitness advantages in the presence of naphtha. In addition to elucidating an important ecological interaction in hydrocarbon resource environments, we describe a model system for testing synthetic microbial communities capable of emergent hydrocarbon transformation processes based on public good dynamics.

Genomic features of plant-associated bacteria

Asaf Levy* (alevy@lbl.gov)[1], Scott Clingenpeel[1], Susannah Tringe[1], Sur Herrera Paredes[2], Derek S. Lundberg[3], Sarah Lebeis[2], Jeffery L. Dangl [2], and Tanja Woyke[1]

[1] DOE Joint Genome Institute, Walnut Creek, California 94598, USA; [2]Department of Biology and Howard Hughes Medical School, Bioinformatics and Computational Biology Program, Curriculum in Genetics, and Carolina Center for Genome Sciences, CB-3280, University of North Carolina, Chapel Hill, North Carolina 27599, USA; [3] Max Planck Institute for Developmental Biology in Tübingen, Germany.

Plants tightly associate with an array of phylogenetically and functionally diverse bacteria, with which they share complex and fascinating interactions. Greatly influencing host fitness, some plant-associated (PA) bacteria supply essential nutrients to plants thereby promoting plant growth, other PA bacteria serve as pathogens causing plant diseases, and some PA bacteria protect the plants from those pathogens. In order to fulfill their mutualistic, parasitic or commensal functions the PA bacteria have evolved a large set of genes that enable them to adapt to the plant environment. Here, we employed a large-scale comparative genomics study in order to characterize the core genome of PA bacteria. This was done by comparing the genomes of 1300 bacterial species from seven prominent taxa (from phyla *Proteobacteria*, *Actinobacteria*, and *Firmicutes*) that were classified as being PA bacteria or non-plant associated (NPA) bacteria based on their site of isolation. This set includes a list of 80 newly sequenced

bacterial genomes that were isolated from the root endophytic compartment of *Arabidopsis*. The genes enriched in PA bacteria include genes involved in nitrogen fixation, antibiotics production, sugar transport and metabolism, plant cell wall degradation, type III and type VI secretion systems, and biosynthesis of plant growth promoting hormones. Interestingly, this analysis also revealed a large set of uncharacterized genes that are enriched in PA bacteria. Many of these latter genes are conserved within and across taxa, clustered in operons, and are upregulated in the plant environment, underscoring their likely roles in plant interactions. Understanding the functions of these genes has enormous potential to improve plant growth efficiency and carbon sequestration by plants via microbiome manipulation.

Chromosome-scale genome scaffolding and organismal deconvolution of metagenomic populations using Hi-C

Ivan Liachko* (il34@uw.edu), Joshua Burton, Maitreya J. Dunham, Jay Shendure.

Department of Genome Sciences, University of Washington, Seattle, WA.

Assembly of whole genomes from next-generation sequencing is inhibited by the lack of contiguity information in short-read sequencing. This limitation also impedes metagenome assembly, since one cannot tell which sequences originate from the same species within a population. We have overcome these bottlenecks by adapting a chromosome conformation capture technique (Hi-C) for the deconvolution of metagenomes and the scaffolding of *de novo* assemblies of individual genomes.

In modeling the 3D structure of a genome, chromosome conformation capture techniques such as Hi-C are used to measure long-range interactions of DNA molecules in physical space. These tools employ crosslinking of chromatin in intact cells followed by intra-molecular ligation, joining DNA fragments that were physically nearby at the time of crosslink. Subsequent deep sequencing of these DNA junctions generates a genome-wide contact probability map that allows the 3D modeling of genomic conformation within a cell. The strong enrichment in Hi-C signal between genetically neighboring loci allows the scaffolding of entire chromosomes from fragmented draft assemblies. Hi-C signal also preserves the cellular origin of each DNA fragment and its interacting partner, allowing for deconvolution and assembly of multi-chromosome genomes from a mixed population of organisms.

We have used Hi-C to scaffold whole genomes of animals, plants, fungi, as well as prokaryotes and archaea. We have also been able to use this data to annotate functional features of genomes, such as centromeres in many fungal species. Additionally, we have applied our technology to diverse metagenomic populations such as craft beer and wine, clinical infections, soil, and tree endophyte samples to discover and assemble the genomes of novel strains of known species as well as novel prokaryotes and eukaryotes.

Integrating multiple genomics approaches to freshwater microbial ecology

Alexandra Linz* (amlinz@wisc.edu), Joshua J. Hamilton, Sarah Stevens, Shaomei He, Katherine D. McMahon

University of Wisconsin, Madison

Genomics provide unparalleled insight into the ecology of uncultured bacteria. Without next-generation sequencing, we would not even know of the existence of entire phyla of bacteria in freshwater.

However, gaining information about organism function from genomic data is difficult at best. Our research group uses multiple genomic approaches to learn more about the ecology, evolution, and function of freshwater bacteria. Our study systems include Lake Mendota, a eutrophic lake located in Madison, WI, and five dystrophic bogs near Minocqua, WI. Both field sites are part of long-term sampling efforts, providing years of metadata to complement our sequencing efforts. We use metagenomes, genomes from metagenomes, single amplified genomes, and iTags in combination. This data is used to answer questions about the natural history of specific freshwater bacterial groups, to reconstruct the metabolisms of uncultured organisms, and to detect community composition changes in response to disturbance.

Genomics-guided discovery of secondary metabolites and their regulation in *Pseudomonas protegens* Pf-5

Joyce E. Loper* (Joyce.Loper@ars.usda.gov)[2,4], Qing Yan[1], Ian Paulsen[2], Harald Gross[3], and Jennifer Clifford[4]

[1]Oregon State University, Corvallis, Oregon, USA; [2]Macquarie University, Sydney, Australia; [3]Universität Tübingen, Germany; and [4]USDA-ARS, Corvallis, Oregon, USA.

Pseudomonas protegens strain Pf-5 is a well-characterized soil bacterium known for its production of a diverse spectrum of secondary metabolites and its capacity to suppress plant diseases caused by soilborne fungal, bacterial and oomycete pathogens. Metabolites produced by Pf-5 include 2,4-diacetylphloroglucinol (2,4-DAPG), pyoluteorin, pyrrolnitrin, hydrogen cyanide, rhizoxin analogs, the cyclic lipopeptide orfamide A, toxoflavin, and the siderophores enantio-pyochelin and pyoverdine. The genomic sequence of Pf-5 was published in 2005, providing the opportunity for genomics-guided discovery of several metabolites (orfamide A, enantio-pyochelin, rhizoxin analogs, and toxoflavin) and other traits (bacteriocins and the insect toxin FitD) that contribute to multitrophic interactions. Additional orphan gene clusters have been identified through transcriptomic analysis of Pf-5 and derivatives with mutations in regulatory genes, through an approach called global-regulator-based genome mining. Our current studies focus on the coordination of secondary metabolite production in *P. protegens* Pf-5. Among the large spectrum of antibiotics produced by this bacterium, two—pyoluteorin and 2,4-DAPG— are known to function in intracellular and intercellular communication, both as autoinducers of their own production. Furthermore, the production of pyoluteorin and 2,4-DAPG is balanced in Pf-5, suggesting a coordinated regulation between the two pathways. Here, we report that phloroglucinol, an intermediate in the biosynthesis of 2,4-DAPG, regulates the transcription of genes in the pyoluteorin biosynthesis cluster in a concentration-dependent manner. Furthermore, phloroglucinol had broad effects on the transcriptome of Pf-5, significantly altering the transcription of 289 genes by at least two fold in this study. The 130 genes most highly regulated by phloroglucinol (≥4 fold) fall into 14 role categories, reflecting the influence of phloroglucinol on diverse aspects of bacterial physiology and highlighting interactions between primary and secondary metabolic pathways. The effects of nanomolar versus micromolar concentrations of phloroglucinol differed both quantitatively and qualitatively, influencing the expression of distinct sets of genes or having opposite effects on transcript abundance of certain genes. Therefore, our results support the concept of hormesis, a phenomenon associated with signaling molecules that elicit distinct responses at different concentrations. Furthermore, our results expand knowledge of the natural roles of secondary metabolites, as the 2,4-DAPG pathway functions both in competition, due to the antibiotic properties of 2,4-DAPG, and in bacterial communication, due to the role of the intermediate phloroglucinol as an intercellular and intracellular signal.

ETOP Project: Development and Implementation of High Throughput Methods for Fungal Culturing and Nucleic Acid Isolation

Jon Magnuson* (Jon.Magnuson@pnnl.gov)[1], **Beth Hofstad[1]**, David E. Culley[1], Mark Butcher[1], James Collett[1], Igor Grigoriev[2]

[1]Pacific Northwest National Laboratory, Richland, WA; [2]Joint Genome Institute, Walnut Creek, CA

As part of the Emerging Technologies Opportunity Program (ETOP; www.jgi.doe.gov/programs/ETOP/) at the Joint Genome Institute, we have developed higher throughput methods for culturing fungi, recording phenotypic growth data, and isolating quality DNA and RNA. This ETOP project also supports the goals of the 1000 Fungal Genomes Project (genome.jgi.doe.gov/programs/fungi/1000fungalgenomes.jsf) by providing DNA and RNA for sequencing the diverse fungi represented in this whole Kingdom sequencing project. The ETOP project has three aspects. 1. Method development for high throughput fungal culturing: culturing filamentous fungi in a flexible and efficient microtiter plate format to rapidly obtain phenotypic growth data, and provide samples from a wide variety of culture conditions for mRNA sequencing to support genome annotation. 2. Method development for high throughput nucleic acid isolation: lysing fungi and extracting DNA and RNA of high quality and quantity suitable for downstream sequencing and analysis at JGI. 3. Fungal nucleic acid delivery: providing DNA and RNA and associated growth data from 145 fungi to JGI. The first phase of this project focused on method development for culturing diverse fungi and simultaneously assessing their growth fluorometrically on different carbon and nitrogen sources. This has been done in 24 well standard footprint plates using the resorufrin/resazurin growth reporting reagent, which detects active aerobic respiration. DNA and RNA samples have been delivered to JGI and a significant number of genomes have been sequenced, annotated and released under the auspices of the 1000 Fungal Genomes Project. These new genomes can easily be found in JGI's Mycocosm. In the next month we will be assessing a commercial high throughput nucleic acid purification system for its effectiveness with diverse fungi. The focus of the remaining months of this project is on continued delivery of DNA and RNA to JGI using the methods developed.

Universal expression tools to improve nutrient acquisition of energy crops

Nasim Mansoori Zangir* (nmansoorizangir@lbl.gov)[1], Carlos Hernandez-Garcia[1], Patrick Shih[1], Charlee Vuu[1], Garima Goyal[1], Sangeeta Nath[2], Nathan J. Hillson[1,2], Samuel Deutsch[2], **Dominique Loqué**[1]

[1]Joint BioEnergy Institute, Physical Biosciences division, Lawrence Berkeley National Laboratory, 1 Cyclotron road, Berkeley, CA 94720, USA; [2]DOE Joint Genome Institute, 2800 Mitchell Drive, B310-346A, Walnut Creek, CA 94598, USA

Project Goals: The main goals of this research are the generation of "universal" expression tools for plant root engineering and to utilize them to improve nutrient acquisition in energy crops.

http://www.jbei.org/research/divisions/feedstock/cell-wall-engineering/

Plant growth and development relies on roots as a means to anchor the plant body as well as to absorb water and nutrients such as N, P and K. As deficiency in these nutrients negatively impacts photosynthesis, plant growth and yield, chemical fertilizers have been used to compensate and meet the growing demand for plant material. Unfortunately the excessive use of fertilizers has come at high environmental and economic costs; and its production utilizes a substantial proportion of worldwide

energy consumption. The current development of growing crops on marginal land (low water content, low nutrient supply, vulnerability to erosion and heavy-metal pollution) for bioenergy will reduce competition with food crops and the pressure on high-quality arable lands utilization. As sufficient nutrient uptake by the plant root is vital especially when soil conditions restrict its availability, root systems in energy crops can be engineered to acquire/ accumulate the required nutrients.

Our focus is to design "universal" expression tools for plant root engineering functional across diverse plant species. Through targeted-cell engineering and synthetic biology we aim to engineer plant root systems to generate metabolic pathways and improve nutrient acquisition/accumulation in energy crops. This aim will be carried out through development and validation of a large set of root-specific and nutrient-responsive promoters. In house transcriptome analysis will help identify root specific genes that are constitutively expressed, induced or repressed through conserved metabolic responses to N, P and Fe starvation in hydroponic conditions across a widely diverse selection of plant species. Promoters from the identified genes will be isolated and characterized for spatiotemporal expression patterns and expression levels across different taxonomic classes using composite plants and conventional transgenic approaches. Subsequently monocot and dicot "universal" promoter libraries of the detected genes alongside those previously reported will be generated using a semi-high-throughput promoter cloning approach and yeast homologous recombination assembly method and ultimately used for tissue specific metabolic pathway engineering to improve nitrogen use efficiency in bioenergy crops.

We expect that the results of this research will generate a diversity of building blocks for plant engineering and will directly contribute to advance the DOE's mission for the sustainable production of bioenergy.

This work was part of the DOE Early Career Award and the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research; and U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

Manganese-peroxidase subfamilies: A critical evaluation based on the *Ceriporiopsis subvermispora* genome

Angel T. Martínez* (ATMartinez@cib.csic.es)[1] Elena Fernández-Fueyo[1], Sandra Acebes[2], Victor Guallar[2], María Jesús Martínez[1], Francisco Javier Medrano[1], Antonio Romero[1], Francisco J. Ruiz-Dueñas[1]

[1]CIB, CSIC, Ramiro de Maeztu 9, E-28040 Madrid, Spain; [2]BSC, Jordi Girona 29, E-08034 Barcelona, Spain

The genome of the white-rot fungus *Ceriporiopsis subvermispora* includes two lignin peroxidase (1) and thirteen manganese peroxidase (MnP) genes, the latter classified into three subfamilies according to the C-terminal tail length (2). These subfamilies appear widespread in the genomes of white-rot basidiomycetes and absent from all the sequenced brown-rot genomes (3). *C. subvermispora* genome includes representatives for the three MnP subfamilies, a fact that enabled comparison without "interferences" from the species evolutionary history. Two long, two extralong and the only short MnP were heterologously expressed and characterized, together with several short-tail variants. The first extralong-MnP structure was solved (4), and compared with the first short-MnP structure, recently solved (5), and with the already know long-MnP structure. Its C-tail surrounding the heme-propionate access channel contributes to Mn^{2+} oxidation (K_m decreased when length increased) but prevents

oxidation of 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonate) (ABTS), which is only oxidized by the short MnP and the short-tail variants. The tail is also associated with a high stability at acidic pH. Cd^{2+} binds at the Mn-oxidation site and competitively inhibited both Mn^{2+} and ABTS oxidation. Moreover, blocking the heme-propionate channel with a double mutation prevented oxidation of both substrates. This agrees with diffusion simulations that positioned ABTS at electron-transfer distance from the heme propionates of a short-tail form. We conclude that only small differences exist between long and extralong MnPs. However, short MnPs appears as a distinct subfamily differing in structural, stability and catalytic properties, the latter including oxidation of other substrates than Mn^{2+} at the heme-propionate channel.

1. Fernández-Fueyo, E., F. J. Ruiz-Dueñas, Y. Miki, M. J. Martínez, K. E. Hammel, and A. T. Martínez. 2012. Lignin-degrading peroxidases from genome of selective ligninolytic fungus *Ceriporiopsis subvermispora*. J. Biol. Chem. 287:16903-16916.

2. Fernández-Fueyo, E., F. J. Ruiz-Dueñas, P. Ferreira, D. Floudas, D. S. Hibbett, P. Canessa, L. Larrondo, T. Y. James et al. 2012. Comparative genomics of *Ceriporiopisis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis. Proc. Natl. Acad. Sci. USA 109:5458-5463.

3. Floudas, D., M. Binder, R. Riley, K. Barry, R. A. Blanchette, B. Henrissat, A. T. Martínez, R. Otillar et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336:1715-1719.

4. Fernández-Fueyo, E., S. Acebes, F. J. Ruiz-Dueñas, M. J. Martínez, A. Romero, F. J. Medrano, V. Guallar, and A. T. Martínez. 2014. Structural implications of the C-terminal tail in the catalytic and stability properties of manganese peroxidases from ligninolytic fungi. Acta Crystallogr. D. Biol. Crystallogr. 70:3253-3265.

5. Fernández-Fueyo, E., F. J. Ruiz-Dueñas, M. J. Martínez, A. Romero, K. E. Hammel, F. J. Medrano, and A. T. Martínez. 2014. Ligninolytic peroxidase genes in the oyster mushroom genome: Heterologous expression, molecular structure, catalytic and stability properties and lignin-degrading ability. Biotechnol. Biofuels 7:2.

Biogeochemistry of the microbial soil sink of carbonyl sulfide (COS) - a carbon cycle tracer

Laura K. Meredith* (predawn@stanford.edu)[1], Paula V. Welander[1], Joe Berry[2]

[1]Environmental Earth Systems Science, Stanford University, [2]Department of Global Ecology, Carnegie Institution for Science

Soils microorganisms access their repertoire of genetic tools for survival in response to the availability of substrates from the organic, mineral, aqueous, and gaseous phases present in their microenvironment. Their collective metabolism has a profound effect on the cycling of elements within soils and the exchange of gases between the atmosphere and soils. One such example is the cycling of carbonyl sulfide (COS), a trace constituent of Earth's atmosphere (global mean of ~500 pmol mol^{-1}), which has a significant global soil sink thought to be driven by microorganisms containing carbonic anhydrase (CA) enzymes that catalyze the hydrolysis reactions of both COS and carbon dioxide (CO_2): $COS + H_2O \rightarrow CO_2 + H_2S$ and $CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$. Because of this shared reaction of COS and CO_2 via CA in plants, measurements of COS hold great promise as a tool for disentangling the large, simultaneous photosynthetic and ecosystem respiration CO_2 fluxes of the terrestrial biosphere, which is critical for

assessing changes in carbon sequestration in response to climate forcing. Cautious application of COS as a carbon cycle tracer requires that other ecosystem fluxes, such as COS soil uptake be understood well enough to be accounted for.

In this project, we are interested in linking gene-, microbe-, and meso-scale processes regarding soil cycling of COS with the hope of identifying the key processes influencing macroscopic soil-atmosphere fluxes of COS. We use gene-based discovery to link soil microorganisms, their CA enzymes, and COS consumption. In particular, we aim to link atmospheric COS uptake to the specific type and clade of CA, and use bioinformatics to determine the breadth of distribution across common soil phyla. Furthermore, we seek to shed light on whether microorganisms consume COS as part of a sulfur assimilation pathway, as is the case for plants, or if COS is simply hydrolyzed by CA due to a structural similarity of COS and CO_2. Advances made in understanding the use of COS at multiple scales in complex soil environments will help inform understanding of soil sulfur cycling and are imperative for the process-level understanding and prediction of soil-atmosphere exchange of COS.

De Novo Assembly of Genomes Using Long Span Ngs Mate Pair Libraries

Scott Monsma* (smonsma@lucigen.com), **David Mead**, Svetlana Jasinovica, Erin Ferguson, Brendan Keough, Megan Niebauer, Michael Lodes

Lucigen Corp., Middleton, WI 53562 USA

Long repetitive DNA sequences are abundant in most species, which creates technical challenges for the *de novo* assembly of even small genomes using short read next generation sequencing (NGS) methods. The incorporation of long span mate pair reads could dramatically improve the success of *de novo* assembly and closing of genomes by linking contigs. Existing methods are limited to 5-6 kb mate pairs, which is inadequate for most microbial or complex genomes. A new NGS library method that generates user defined mate pairs (MP) from 2-8 kb and 10-20 kb using bead-based and gel-based methods, respectively, has been developed. A unique barcoding strategy is used to distinguish true mate pairs from false chimeric junctions, reducing the fraction of misassembled contigs. We report the closing and finishing of four bacterial genomes using a single 10-20 kb mate pair library in conjunction with a conventional 600 bp paired end fragment library using Illumina sequencing chemistry. Genomes representing diverse sizes and %GC content were closed and finished with this simple strategy including *Thermus aquaticus* (2.2 Mb, 68% GC), *Staphylococcus aureus* (2.8 Mb, 32% GC), a *Streptomyces* spp. (8.6 Mb, 71% GC), and a *Nonomurea* spp. (10.3 Mb, 70.4% GC). Preliminary results indicate that the technology is scalable to 100 kb MP libraries, with important consequences for assembling repeat rich, complex genomes from fungi, mitochondria, chloroplasts, plants and animals. We also report on the scaffolding of human, maize, switchgrass, and a sorghum mitochondrial genome with 20-100 kb mate pair libraries. The ability to construct and sequence mate pair libraries up to 100 kb (BAC-sized paired end reads) without physical cloning simplifies the accurate closing and finishing of complex genomes economically.

Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens

Abraham Morales-Cruz* (abmora@ucdavis.edu)[1], Katherine C.H. Amrine[1], Barbara Blanco-Ulate[1], Daniel P. Lawrence[2], Renaud Travadon[2], Philippe E. Rolshausen[3], Kendra Baumgartner[4], Dario Cantu[1]

[1] Department of Viticulture and Enology, University of California Davis, CA 95616, U.S.A; [2] Department of Plant Pathology, University of California Davis, CA 95616, U.S.A; [3] Department of Botany and Plant Sciences, University of California Riverside, CA 92521, U.S.A; [4] United States Department of Agriculture - Agricultural Research Service, Crops Pathology and Genetics Research Unit, Davis, CA 95616, U.S.A

Background

Trunk diseases threaten the longevity and productivity of grapevines in all viticulture production systems. They are caused by distantly-related fungi that form chronic wood infections, but variation in wood-decay abilities and production of phytotoxic compounds are thought to contribute to their unique disease symptoms. We recently released the draft sequences of *Eutypa lata*, *Neofusicoccum parvum* and *Togninia minima*, causal agents of Eutypa dieback, Botryosphaeria dieback and Esca, respectively. In this work, we first expanded genomic resources to three important trunk pathogens, *Diaporthe ampelina*, *Diplodia seriata,* and *Phaeomoniella chlamydospora*, causal agents of Phomopsis dieback, Botryosphaeria dieback, and Esca, respectively. Then we integrated all currently-available information into a genome-wide comparative study to identify gene families potentially associated with host colonization and disease development.

Results

The integration of RNA-seq, comparative and *ab initio* approaches improved the protein-coding gene prediction in *T. minima*, whereas shotgun sequencing yielded nearly complete genome drafts of *Dia. ampelina*, *Dip. seriata,* and *P. chlamydospora.* The predicted proteomes of all sequenced trunk pathogens were annotated with a focus on functions likely associated with pathogenesis and virulence, namely (i) wood degradation, (ii) nutrient uptake, and (iii) toxin production. Specific patterns of gene family expansion were described using Computational Analysis of gene Family Evolution, which revealad lineage-specific evolution of distinct mechanisms of virulence, such as specific cell wall oxidative functions and secondary metabolic pathways in *N. parvum*, *Dia. ampelina*, and *E. lata*. Phylogenetically-informed principal component analysis revealed more similar repertoires of expanded functions among species that cause similar symptoms, which in some cases did not reflect phylogenetic relationships, thereby suggesting patterns of convergent evolution.

Conclusions

This study describes the repertoires of putative virulence functions in the genomes of ubiquitous grapevine trunk pathogens. Gene families with significantly faster rates of gene gain can now provide a basis for further studies of *in planta* gene expression, diversity by genome re-sequencing, and targeted reverse genetic approaches. The functional validation of potential virulence factors will ultimately lead to a more comprehensive understanding of the mechanisms of pathogenesis and virulence, which ultimately will enable the development of accurate diagnostic tools and effective disease management.

Barley genomics - developing genomics resources to explore diversity, adaptation and selection

Gary J. Muehlbauer (muehl003@umn.edu)

Department of Plant Biology, Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108

Barley is a diploid member of the Triticeae with a 5.1 Gb genome. It was domesticated approximately 10,000 years ago in the fertile crescent and moved throughout Asia, Europe and Africa and then subsequently to all temperate regions of the world. To enable genomic studies, an integrated draft genome sequence composed of a BAC-based physical map, BAC end sequences, complete BAC sequences, and survey sequencing of sorted chromosome arms was released in 2012. Population sequencing (POP-SEQ) of two mapping populations resulted in anchoring and ordering of 1.22 Gb and genetically anchoring the BAC-based physical map. Recent efforts to sequence the minimum tilling path of BAC clones are underway and four of the seven chromosomes are complete. The genome sequence was used to develop a 62 Mb exome capture technology and exome capture sequencing was conducted on a collection of 228 cultivars, landraces and wild species from geographically dispersed regions of Europe and Asia. Analysis of the sequence data shows that genetic diversity correlates with eco-geographical regions, the genome is represented by high and low amounts (regions of selection for domestication traits) of genetic diversity, and tremendous diversity is associated with genes that regulate flowering time. These results and related results of this study will be presented.

Expanding the Genomic Encyclopedia of Bacteria and Archeae with a Thousand Microbial Genomes (KMG)

Supratim Mukherjee* (supratimmukherjee@lbl.gov)[1], David Paez-Espino[1], Natalia N. Ivanova[1], Jan Meier-Kolthoff P.[2], Markus Göker[2], Hans-Peter Klenk[3], Jonathan A. Eisen[4], Tanja Woyke[1], Nikos C. Kyrpides[1] and Amrita Pati[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA, USA; [2]DSMZ-German Collection for Microorganisms and Cell Cultures, Braunschweig, Germany; [3]School of Biology. Newcastle University, UK; [4]University of California, Davis, CA, USA

Microorganisms play key roles in important living processes and biogeochemical cycles; yet, current sequencing efforts, with upwards of 40,000 isolate prokaryotic sequencing projects in various stages of completion, have covered only a minute fraction of their diversity. This selection bias necessitated a systematic sequencing approach in order to maximize chances of discovering novel biological phenomenon as well as to avoid populating the prokaryotic genome space with redundant information. Along these lines, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project pioneered a phylogeny-driven sequencing effort by targeting the sequencing of 250 archaeal and bacterial organisms.

Here, we significantly expand the original GEBA project by analyzing genome sequences of over one thousand phylogenetically diverse strains of Bacteria and Archaea. Majority of the strains, collectively known as KMG (One Thousand (K) Microbial Genomes), include the cultured type strain of their species sampled from different parts of the world. Their isolation sources span a variety of habitats including extreme environments, enriched soils, industrial waste site, human body sites and more. The target organisms were selected using a phylogeny-based scoring system leading to a phylogenetic distribution across 21 different phyla. An amalgamation of single-copy genes, average nucleotide identity (ANI) and

whole proteome-based distance calculation method was applied to analyze the phylogenetic diversity of the KMG dataset. A combination of in-house and publicly available tools was employed to study the abundance and distribution of CRISPR-Cas system and identify biosynthetic gene clusters. A sequence similarity based implicit phylogenetic metric was developed to identify new functions within a phylum via potential horizontal gene transfers. Phylogenetic expansion of known protein families and identification of novel protein families within KMG were performed using homology based detection methods. Effective characterization of any metagenome depends heavily on reference isolate genomes. To test if KMG improves phylogenetic annotation of available metagenomes, phylogenetic anchoring of metagenome contigs from the Integrated Microbial Genomes (IMG) database was performed. This work represents the first phase of a larger effort to improve the phylogenetic coverage of all cultured microbial species, fill evolutionary gaps in the Bacterial and Archaeal branches of the Tree of Life and increase our overall understanding of microbial evolution and function.

Genomic sequencing and high-throughput phenotyping highlight adaptive strategies in the common green alga *Chloroidium sp. DN1*

David R. Nelson* (drn2@nyu.edu)[1,2], Ashish Jaizwal[1,2] , Sudhir Sharma[3], Kenan Jijakli[1,2,3], Amphun Chaiboonchoe[1,2], Wei Qi Fu[1,2], Marc Arnoux[2], Nizar Drou[2], Bushra Saeed Dohai[1,2†], Sachin Kapli[3], Ramesh Jagganathan[3], and Kourosh Salehi-Ashtiani[1,2]

[1]Division of Science and Math, New York University Abu Dhabi, Abu Dhabi, UAE; [2] Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi, Abu Dhabi, UAE; [3] Engineering Department, New York University Abu Dhabi, Abu Dhabi, UAE

In this report, we present an analysis protocol for the rapid characterization of new algal strains or uncharacterized strains. We use a newly isolated strain of Chloroidium*,* herein referred to as *Chloroidium sp. DN1,* as an example of the breadth and scope of information that can be obtained with our approach. Originally isolated from fresh water, we found that *Chloroidium sp. DN1* can grow in high salinities and grow heterotrophically using a broad range of carbon substrates. *Chloroidium sp. DN1* was sequenced using the Illumina HiSeq 2500 to yield a genome of 52.5 Mb and 10,605 predicted ORFs. Analysis of the draft genome and ORFeome reveals potential metabolic strategies to explain its adaptation to a desert climate. We created a draft metabolic network of *Chloroidium sp. DN1* using the software Pathway Tools. The metabolic network is comprised of 1,445 genes in 194 pathways. Phenotype analysis of *Chloroidium sp. DN1* was performed with analytical bioreactors, phenotype microarrays (PMs), gas chromatography/mass spectrometry (GC/MS), and flow cytometry. *Chloroidium sp. DN1* was found to grow heterotrophically on 10 different carbon compounds including 2,3−Butanedione, Dihydroxy−Acetone, D−Arabitol, D−Mannitol, D-Sorbitol, D-Mannose, D-glucose, trehalose, α−Keto−Butyric Acid, D-Halose, and D-Galactose. Overall 38 peptides/dipeptides were found to serve as nitrogen sources promoting heterotrophic growth. Resistance to over 470 chemicals/antibiotics was tested. *Chloroidium sp. DN1* was found to be highly resistant to a number of common antibiotics such as paromomycin and tetracycline. Several genes were predicted in genomic analyses that are likely to play roles in *Chloroidium sp. DN1's* broad carbon substrate uptake capabilities and adaptation to variable salt conditions. Lipid bodies were found to accumulate in intracellular space in undisturbed cultures. Growth was analyzed in the presence of 1,728 different chemical compounds and osmolyte concentrations. Lipid analysis using GC/MS, flow cytometry, and fluorescent microscopy revealed a capability of Chloroidium sp. DN1 to accumulate high levels of palmitic acid in normal growth conditions. Overall, maximum photoautotrophic biomass production was estimated to be over 120 kg/acre/day from growth in open pond simulator analytical bioreactors. Our results validate our

platform for the rapid, comprehensive analysis of new algal isolates with the intitial characterization of a newly isolated strain of algae that may have properties beneficial for commercial applications.

Key words: Alga (green) – Genomics – Phenotype microarray – Salinity tolerance

Simplified microbial communities as models for studying functional compartmentalization

William C. Nelson* (William.Nelson@pnnl.gov), Stephen R. Lindemann, Margie F. Romine, James K. Fredrickson

Pacific Northwest National Laboratory, Richland, WA

Microbial mats are laminated biofilm communities of bacteria, archaea, and microeukaryotes embedded in extracellular polymeric substances. The complex and compact spatial structure promotes metabolic networks among community members, leading to efficient energy utilization, nutrient cycling, and obligate mutualistic relationships. Assembly of these communities is driven by a combination of stochastic forces (e.g., colonization order) and deterministic forces (energy acquisition, nutrient availability, resource competition, microbial interactions), while spatial arrangement is constrained by gradients of physical and chemical parameters formed by environmental forces and microbial function. We hypothesize that functional compartmentalization promotes metabolic interaction and interdependencies between community members. These interactions have a stabilizing effect upon diversity, and thereby lead to community resilience to environmental variation.

Studies of microbial ecology generally suffer from three common difficulties: 1) high community diversity complicates identification of rare but important members; 2) in field studies, environmental variation is uncontrolled and stochastic, clouding causative relationships; 3) in laboratory studies, constructed communities rely on cultured organisms, which can call into question the ecological relevance of the interactions. We have developed a system that addresses all three problems.

To examine metabolic functional compartmentalization and the effects of environmental variables upon community composition, we have generated two unicyanobacterial consortia (UCC-A and UCC-O) derived from the hypersaline Hot Lake phototrophic mat. Community composition is simple, being composed of a single autotroph and <20 associated heterotrophs. The consortia are stable to serial dilution passage in culture and thus tractable to controlled manipulation and, being derived from an environmental sample, interactions and dependencies between members are likely to have ecological relevance.

Metagenomic sequence data collected from the UCC-A and UCC-O cultures have been assembled and segregated into taxonomic bins resulting in 18 distinct, near-complete genome sequences, 15 of which are shared between the two communities. Each consortium contains a single cyanobacterium that is the sole autotroph and a heterotrophic assemblage comprised of members of Bacteroidetes, Gammaproteobacteria, and Alphaproteobacteria. Genome analysis indicates most organisms in the cultures can use urea as a nitrogen source, and the Gammaproteobacteria can use cyanate, a toxic product of urea decomposition. The alphaproteobacterium *Oceanicola* possesses the most diverse set of carbohydrate catabolic genes and is capable of degrading a range of mono- and disaccharides, organic acids and sugar alcohols. Other community members contain varying subsets of carbohydrate catabolism genes, with the gammaproteobacterium *Idiomarina* and alphaproteobacterium *Oceanicaulis*, putative amino acid fermenters, having fewest. Isolate strains were tested for growth on various carbon sources to test predicted carbon sources.

These examples highlight potential niche specialization and functional compartmentalization among community members that will be empirically validated using our consortia and isolate cultures. Interactions identified in the consortia cultures will be tested through field observations to evaluate the consortia as model ecological systems.

Genome of an uncharted *Euryarchaeota* clade WSA2 reveals uniquely restricted methyl-reducing methanogenesis

Masaru K. Nobu* (masarunobu@gmail.com)[1], Takashi Narihiro[1,2], and Wen-Tso Liu[1]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave, Urbana, Illinois 61801; [2] Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Central 6, Higashi 1-1-1, Tsukuba, Ibaraki 305-8566, Japan

Despite the long history of research on *Euryarchaeota* methanogenesis, we have yet to elucidate the ecophysiology of the large class-level clone cluster WSA2 (or Arc I) associated with anaerobic digestion and a variety of other anaerobic environments. To gain insight into how this clade may contribute to the methanogenic sector of the global carbon cycle, a draft genome was recovered through metagenomics sequencing of anaerobic digester sludge. Taxonomic analysis of the genes encoding 16S rRNA and methyl-coenzyme M reductase A places WSA2 as a distinct clade between *Methanomassiliicoccales* and *Methanobacteriales*; however, phylogenomics reveals that the WSA2 genome is distantly related from methanogenic clades, suggesting distinct physiology. For methanogenesis, WSA2 encodes genes for H_2 (electron-confurcating hydrogenase) and carbon monoxide (CO dehydrogenase) oxidation, H_2 cycling (coenzyme F420 dehydrogenase and energy-coverting hydrogenase), dimethylsulfide/methanethiol demethylation (methylated thiol:coenzyme M methyltransferase, MtsAB), and conventional methyl-reduction. On the other hand, it lacks typical methanogenic pathways including CO_2 reduction, coenzyme F420 biosynthesis, and acetyl-CoA pathway. This suggests that WSA2 may depend on exogenous carbon and F420 sources (*e.g.*, produced by other methanogens *in situ*). Indeed, the genome harbors heterotrophic pathways and a coenzyme F420 transporter. Thus, this metagenomic study indicates that WSA2 is a novel hydrogenotrophic methyl-reducing methanogen uniquely restricted to respiring methylated sulfide species and also physiologically reliant on exogenous resources.

Emerging Pipelines in Multi-Omics Informatics: Dealing with High Volume and High-Throughput Data

Angela D. Norbeck* (angela.norbeck@pnnl.gov)[1], Douglas Baxter[1], David Brown[1], Kevin Fox[1], Hyunjoo Walker[2]

[1] Environmental and Molecular Sciences Laboratory, Pacific Northwest National Lab, Richland, WA 99352; [2] Computational and Statistics Analysis Division, Pacific Northwest National Lab, Richland, WA 99352

As improvements in instrumentation and computation increase biological data production, developing tools to manage, store, and automate workflows on that data become imperative for any organization, and especially so for national user facilities. At EMSL, we have developed a data aggregation, storage, retrieval, and distribution system called MyEMSL, where users can manage their data files. In addition, we have a web interface for our EMSLHub environment, where users can exchange data, notes, tasks

lists, and project updates with anyone assigned to the team. This platform also provides a workspace where users can initiate analytical workflows and share their desktop with other team members. In recent months, we have been focusing on developing workflows specific to microbial community multi-omics experiments. These include using genomics data from the JGI, proteomics and metabolomics data from our high-throughput mass spectrometry instruments and transcriptomic data, analyzed using computational resources at EMSL. In some cases, we need to utilize the Cascade supercomputer in the workflows, and we are currently developing a lightweight visualization tool for browsing results.

Nitrification in Agricultural Soils under Contrasting N Management: From Soils to Genomes and Back Again.

Jeanette M. Norton* (jeanette.norton@usu.edu)[1], Marlen C. Rice[1], Yang Ouyang[1], Mussie Habteselassie[2], Jennifer Reeve[1], John Stark[1], Chee Kent Lim[3], Lisa. Y. Stein[4], Luis Sayavedra-Soto[5], Martin G. Klotz[3], Jessica Kozlowski[4], Hidetoshi Urakawa[6], Nicole Shapiro[7], Lynne A. Goodwin[8]

[1]Utah State Univ., Logan, UT, USA, Univ. Georgia Griffin GA, USA[2], , [3]Univ. of North Carolina, Charlotte, NC, USA, [3]Miami Univ., Oxford, OH, USA, [4]Univ. of Alberta, Edmonton, AB, CANADA, [5]Oregon State Univ., Corvallis, OR, USA, [6]Florida Gulf Coast University, Fort Myers, FL, USA, [8]DOE Joint Genome Inst., Walnut Creek CA, USA, [9]Los Alamos Natl. Lab., Los Alamos, NM, USA.

Nitrification changes the form of nitrogen (converting ammonium/ammonia to nitrite and nitrate) and thereby alters the fate of N in agricultural soils. Nitrification provides a link between ammonium (product of mineralization and major fertilizer input) and denitrification, the major biological loss for fixed reactive N and a source of the potent greenhouse gas N_2O. Understanding nitrification is therefore central to the ability to predict and manage soil N losses and to understand impacts of agricultural management on N_2O production. Since many of the soil nitrifiers are difficult to culture, molecular tools are used extensively to characterize the community responsible for nitrification in soils. Methods for determining the rates and kinetics of the nitrification process are advancing simultaneously. Genome sequences of soil ammonia oxidizers including *Nitrosospira multiformis* C71, *Nitrosospira briensis* C128 and *Nitrososphaera viennensis* EN 76 (Stieglmeier et al. 2014) have been key for understanding the physiology of ammonia oxidizers. A field plot experiment (multi-year replicated randomized block design) was initiated in 2011 in both Utah and Georgia USA to examine nitrogen source effects on nitrification and mineralization in agricultural systems. Nitrogen sources include high and low levels of ammonium sulfate fertilizer (100 and 200 kg N/ha) and manure composts. DNA extraction and soil characterization were accomplished for multiple time points. Real-time quantitative PCR targeting *amoA* for ammonia oxidizers was used to follow changes in bacterial (AOB) and archaeal ammonia oxidizer (AOA) populations. Selective inhibition experiments that differentiate between ammonia oxidation mediated by the two groups were applied to the same experimental system. Our results suggest that although archaeal *amoA* gene copies outnumber bacterial *amoA* gene copies; AOB in the genus *Nitrosospira* dominate ammonia oxidation, in particular under fertilized (high nitrogen) conditions. Ongoing research includes further amplicon library sequencing of the *amoA*, SSU RNA gene, preliminary metagenomic sequencing and process kinetics. This combination of soil ecology, genomics and environmental metagenomics is pursued with the goal to link the capable organisms to the process rate and extent in the environment.

Support from DOE-JGI Community Sequencing Program 2010, 2014, Utah Agric. Expt. Sta. #UTA00371, NSF:0541797 and USDA Agriculture and Food Research Initiative Competitive Grants Program Grant no. 2011-67019-30178

Stieglmeier, M., Klingl, A., Alves, R.J.E., Rittmann, S., Melcher, M., Leisch, N., Schleper, C., 2014. *Nitrososphaera viennensis* gen. nov., sp nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. International Journal of Systematic and Evolutionary Microbiology 64, 2738-2752.

Habteselassie, M.Y., Xu, L., Norton, J.M., 2013. Ammonia-Oxidizer Communities in an Agricultural Soil treated with Contrasting Nitrogen Sources. Frontiers in Microbiology 4, 326.

Cradle to grave: Using comparative metatranscriptomics to identify how roots alter the decomposition of organic matter in soil

Erin E. Nuccio* (nuccio1@llnl.gov)[1], Thea Whitman[2], Shengjing Shi[2,3], Jizhong Zhou[3], Mary S. Lipton[4], Eoin L. Brodie[2,5], Mary K. Firestone[2], **Jennifer Pett-Ridge[1]**

[1]Lawrence Livermore National Laboratory, Livermore, USA; [2]University of California, Berkeley, USA; [3]University of Oklahoma, Norman, USA; [4]Pacific Northwest National Laboratory, Richland, USA; [5]Lawrence Berkeley National Laboratory, Berkeley, USA

The soil surrounding roots, known as the rhizosphere, is the primary nexus of belowground carbon cycling in terrestrial systems. Rhizosphere microorganisms can alter the breakdown of plant tissues and root litter, and in many cases have been shown to accelerate the decomposition of detrital plant biomass. However, the soil microbial communities and carbohydrate and lignolytic gene transcripts mediating the decomposition of root litter in soil are largely unidentified. We hypothesize that root exudates stimulate the expression of enzymes that decompose macromolecular C compounds. To assess how the abundance and diversity of decomposition enzymes differs in the rhizosphere relative to the surrounding bulk soil, we sequenced the metatranscriptomes of rhizosphere and bulk soil over time under the auspices of the JGI CSP and JGI/EMSL programs. Glycoside hydrolases were the most abundant class of enzyme transcripts detected in the rhizosphere. Our initial results suggest that enzymes involved in the breakdown of plant polysaccharides were more highly expressed in the rhizosphere compared to the bulk soil. In addition, we are using genomic and proteomic approaches (included stable-isotope techniques) to identify key metabolic pathways responsible for C transformation and mineralization during root in-growth and root death/decay. By integrating stable isotopes as tracers of natural resource utilization (i.e. root litter), and analysis of the functional properties of the communities that respond to those resources, we can identify the microbial communities that are stimulated in the soil microbiome in response to root litter, living roots, and their intersection.

Immune-modulatory genomic properties differentiate gut microbiotas of infants with and without eczema

Seungdae Oh* (seungdae@illinois.edu)[1], Gaik Chin Yap[2], Pei-Ying Hong[3,4], Chiung Hui Huang[2], Marion Aw[2], Lynette Pei Chi Shek[2], Wen-Tso Liu[1], and Bee Wah Lee[2]

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA: [1]Department of Pediatrics, National University of Singapore, Singapore, Singapore; [2] Division of Biological and Environmental Science and Engineering; [3]Water Desalination and Reuse Center; [4]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Gut microbiota plays an important role in human immunological processes, potentially affecting allergic diseases such as eczema. Although the diversity and structure of the gut microbiota in infants with eczema have been previously documented, how the gene content differs from those of healthy infants remain unknown. To gain insight into this, fecal samples were collected from twelve infants at the age of one month old. These infants were followed up prospectively from birth till the age of two years. There were six cases of eczema and six controls who were matched for mode of delivery and infant feeding. Comparative metagenomic analysis showed that the six healthy control (H) communities were significantly enriched in genes associated with tetrapyrrole (potential anti-inflammatory agent) biosynthesis compared to the six eczema (E) communities. Further, specific immune-regulatory DNA motifs were also significantly enriched in H communities, many of which were encoded by *Bifidobacterium* (38% of the total motifs in H communities). Draft genomes of five *Bifidobacterium* species (*B. longum*, *B. bifidum*, *B. breve*, *B. dentium,* and *B. pseudocatenulatum*) were recovered from metagenomic datasets. The *B. longum* BFN-121-2 genome encoded more immune-regulatory DNA motifs (4.2 copies per 1 million genome sequence) than other *Bifidobacterium* genomes and was significantly overrepresented in H communities. Taken together, our results report distinct immune-modulatory genomic properties of gut microbiotas in infants associated with eczema and provide new insights into the potential roles of gut microbiota in influencing human immune homeostasis.

Functional and Comparative Genomics of Lignocellulose Degradation by *Schizophyllum commune*

Robin A. Ohm* (raohm@lbl.gov)[1,2], Heather M. Brewer[3], Akiko Carver[2], Alex Copeland[2], Jane Grimwood[4], Hanbyul Lee[2,5], Erika A. Lindquist[2], Anna Lipzen[2], Joel Martin[2], Hongjae Park[2,5], Samuel O. Purvine[3], Wendy Schackwitz[2], Margo Schuller[1], Martin Tegelaar[1], Andrew Tritt[2], Scott Baker[3], In-Geol Choi[2,5], Luis G. Lugones[1], Han A. B. Wösten[1], Igor Grigoriev[2]

[1]Utrecht University, Utrecht, The Netherlands; [2]DOE Joint Genome Institute, Walnut Creek, California, USA; [3]Pacific Northwest National Laboratory, Richland, Washington, USA; [4]Hudson Alpha Institute for Biotechnology, Huntsville, Alabama, USA; [5]Korea University, Seoul, South Korea

The Basidiomycete fungus *Schizophyllum commune* is a wood-decaying fungus and is used as a model system to study lignocellulose degradation. Version 3.0 of the genome assembly filled 269 of 316 sequence gaps and added 680 kb of sequence. This new assembly was reannotated using RNAseq transcriptomics data, and this resulted in 3110 (24%) more genes.

A large number of wild isolates of *S. commune* have been isolated from all over the world. These strains show a large diversity in phenotypes. Two strains with different wood-decaying properties were sequenced, from Tattone (France) and Loenen (The Netherlands). Sequence comparison shows remarkably high sequence diversity between the strains. The overall SNP rate of > 100 SNPs/kb is among the highest rates of within-species polymorphisms in Basidiomycetes. Some well-described proteins like hydrophobins and transcription factors have less than 70% sequence identity among the strains. Some chromosomes are better conserved than others and in some cases large parts of chromosomes are missing from one or more strains.

Gene expression on glucose, cellulose and wood was analyzed in two *S. commune* strains. Overall, gene expression correlated between the two strains, but there were some notable exceptions. Of particular interest are CAZymes (carbohydrate-active enzymes) that are regulated in different ways in the different strains. Moreover, a large number of hypothetical genes were strongly up-regulated on cellulose and/or wood, and these may encode novel enzymes involved in lignocellulose degradation. Proteomics analyses

of the secretome of these strains during growth on various carbon sources provided additional insight, as well as additional potentially novel enzymes.

In both strains the transcription factor Fsp1 was strongly up-regulated during growth on cellulose and wood, when compared to glucose. A knock-down of Fsp1 resulted in significantly lower cellulase enzyme activity, which suggests that Fsp1 is involved in regulating CAZyme gene expression.

An Intuitive '16S rRNA Biodiversity Tool'

Christian Olsen* (christian@biomatters.com)[1], **Helen Shearman**[2], C. Duran[1,2], K. Qaadri[1], S. Stones-Havas[2], B. Ammundsen[2]

[1] Biomatters, Inc. San Francisco, CA; [2] Biomatters Ltd. Auckland, New Zealand

The Geneious R8 '16S rRNA Biodiversity Tool' is a cloud-based tool for routine species classification and relative abundance measurement using high throughput 16S rRNA amplicon sequencing data from environmental samples. Preprocessed full-length bacterial 16S rRNA sequences may be utilized, but any sub-region of 16S can be used.

The user submits their next-generation sequence data through the Geneious R8 bioinformatics platform to a distributed cloud compute resource. The data are then analyzed using the Ribosomal Database Project (RDP) database Classifier. The RDP Classifier assigns sequences derived from bacterial and archaeal 16S genes and fungal 28S gene to the corresponding taxonomy model using a 'Naïve Bayesian Classifier' for rapid assignment of rRNA sequences.

The Geneious '16S Biodiversity Tool' accurately assigns a taxonomy (in the range of domain to genus) along with a confidence-estimate for each sequence by comparing them to the RDP database. A secure weblink is returned within Geneious' 'Document Viewer'. Upon clicking the weblink, the output is then displayed in a web browser using Krona, which produces an interactive HTML5 hierarchical graph of the bacterial diversity in the sample. In this poster we present an easy to use web application tool for the analysis of 16S rRNA fragment and whole sequence data.

Uncovering Earth's Virome

David Paez-Espino* (adpaezespino@lbl.gov), Alexander D. Thomas, Emiley A. Eloe-Fadrosh , Marcel Huntemann, Amrita Pati, Edward M. Rubin, Natalia N. Ivanova , **Nikos C. Kyrpides** (nckyrpides@lbl.gov)

Department of Energy, Microbial Genome and Metagenome Program, Joint Genome Institute, Walnut Creek, USA.

Viruses are the most abundant biological entities on Earth. They play an essential role in the carbon (and other) cycle(s), microbial population levels, and diversity. However, the number of the currently available viral genomes is extremely small due mainly to the complexity in the isolation of the host where they rely on, or the availability of large culture collections. To address these limitations, we have used and developed a variety of computational methods and approaches that enable the mining of the largely unexplored metagenomics sequence space. These approaches include methods aiming at the identification of both novel viral genomes as well as their host organism where possible. We identified over 135,000 viral scaffolds (~2 Gb of viral sequences), representing more than an order of magnitude fold increase over the number of previously DNA known viruses, from mining almost 2,000

metagenomics samples from a large variety of environmental and host associated habitats -comprising around 3 Tb of metagenomes sequences. We developed and implemented a novel sequence-based classification system which provides a coherent framework for clustering viral groups and putative linkages with their host. We will present the main findings of this project including interesting novel viral features, habitats with statistical significant presence of viruses, and in many cases, the identification of the virus host organism -enabling both biogeographical analysis of a massive number of viruses on a global scale, as well as gaining further insights on the host-virus antagonism.

Defining the functional diversity of the *Populus* root microbiome

Dale A. Pelletier* (pelletierda@ornl.gov)[1], David W. Ussery[1], Se-Ran Jun[1], Collin M. Timm[1], W. Nathan Cude, Tse-Yuan Lu[1], Miriam Land[1], Intawat Nookaew[1], Christopher W. Schadt[1], Gerald A. Tuskan[1], Mitchel J. Doktycz[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN. USA

Populus is a widely studied model woody plant species and a potential cellulosic feedstock for biofuels. These trees are also host to a wide variety of microbial associations within their roots and rhizosphere and thus serve as a model to study interactions between plants and microorganisms. Our long-term goal is to gain a detailed mechanistic understanding of how bacterial strains recognize their plant host, form stable interactions and ultimately affect host function. Over the past several years, our group as part of a DOE funded Plant Microbe Interfaces project (http://PMI.ornl.gov), have isolated >2500 bacterial strains from *Populus* roots representing >425 distinct OTUs. The genomes of >400 of these strains are currently being sequenced in collaboration through this JGI-CSP. To date we have >100 draft genome sequences from *Populus* bacterial isolates and we are applying comparative genomics, metabolic modeling and *Populus*-bacterial model system studies to identify functions and important mutualistic relationships. Here we compare three different genera for which genome sequence has been obtained, including >15 strains each of *Rhizobium*, *Pseudomonas* and *Chryseobacterium*. Members of these genera are commonly found in the rhizosphere and endosphere compartments of *Populus* roots. Comparative genomics of these endophytes are done within the broader context of all known sequenced genomes from the same genera. Initial comparative genomic analyses, metabolic reconstructions and phenotypic characterizations have revealed diverse functional capabilities among these microbes, including the production of siderophores, indole-3-acetic acid, acyl-homoserine lactone signals, antimicrobial compounds and diverse carbon metabolism. Additionally, plant-microbe interaction assays have led to observations of phenotypic effects on host plants including immune responses, mycorrhizal helper effects, plant growth, root biomass and architecture effects. Metabolic models were also generated for specific strains and experiments performed, and their results were compared with predictions based on the genome sequences. In addition to genome sequence analysis of isolates we have been developing methods for enriching microbes from plant tissues for metagenomic sequencing of *Populus* endosphere samples that avoid host DNA background contamination to gain insight into the functional repertoire of un-cultivated microbial endophytes. Our initial data from a *Populus deltoides* root endophyte metagenome library constructed from DNA extracted from microbially enriched fraction via differential centrifugation prior to extraction. Sequencing shows that host contaminant DNA makes up less the 10% of the final metagenome reads in this root endophyte library. These methods should enable direct comparison of soil, rhizosphere and root endosphere communities using complete metagenomic toolkits, and we are currently collaborating with JGI to analyze 30 paired samples collected from the microbiome of *Populus deltoides* and *P. trichocarpa* x *deltoides* hybrids. These genomic resources provide insight into, the genotype/phenotype relationship, putative signaling and molecular pathways

leading to host recognition and maintenance of interactions and ultimately microbial affects on host function.

Kinetic, Metabolic and Regulatory Models of Class II Methanogens

Joseph R. Peterson* (jrptrsn3@illinois.edu)[1], Piyush Labhsetwar[2], Lars Kohler[1], William W. Metcalf[3] and **Zaida Luthey-Schulten**[1,2,3]

[1]School of Chemical Sciences, [2]Center for Biophysics and Computational Biology, [3]School of Molecular and Cellular Biology, [4]Carl R. Woese Institute for Genomic Biology, University of Illinois, Champaign, IL, USA

Here we describe our work modeling *Methanosarcina acetivorans* and *M. barkeri*. These class II methanogens are capable of producing methane from the largest number of growth substrates by utilizing pathways for acetotrophic growth (acetate), carboxidotrophic growth (CO), hydrogenotrophic growth (H_2/CO_2, H_2/HCO_2^-) and methylotrophic growth (methanol, methylamines methylsulfides, and methylmercaptopropionate) and thus play a key role in biomass breakdown and global carbon cycling. We constructed a kinetic model of the methanogenesis pathways in *M. acetivorans* based on experimental RNA expression data that is capable of capturing growth on acetate, methanol and methanol/trimethylamine mixtures. It allowed identification of reactions where enzyme counts had large impact on the growth and methane production rates. Rates were most sensitive to enzyme copy numbers of three sets of proteins: 1) those flanking the branch point to the acetotrophic pathway (Mtr, Mer), 2) those used for uptake and efflux (Ack, Mta, Mcr), and 3) and those involved in the electron transport chain (Hdr, Rnf, Fpo). Further exploring the effect electron transport chain enzyme copy number has on methane production, we combined kinetic measurements with simulations of hydrogen cycling in *M. barkeri* to derive steady-state fluxes through the membrane bound hydrogenases (Vht, Ech, Frh) that were consistent with experimentally determined growth rates and proton gradients. These fluxes were used as constraints in the recent genome-scale metabolic model (iMG746) and greatly improved the predicted methane production rate over the unconstrained model under methylotrophic and hydrogenotrophic growth conditions (MeOH, $MeOH/H_2$, CO_2/H_2, TMA, TMA/H_2). Simulations of gene deletions showed that this constrained model generally predicted the cell doubling time better than the unconstrained model, vastly improving the model's predictive power.

In order to address interactions across the full genome we next constructed a regulatory network from RNA expression data for *M. acetivorans* grown in 15 different environmental conditions, which allowed the construction of gene regulatory networks and identification of novel regulators. This network captured many known interactions mediated by a phosphate uptake regulator and several previously characterized methyl-specific regulators (msr), as well as predicted the interactions of cobalt and nickel uptake regulators, and several putative global-acting transcription factors. In order to get an idea of the timescale involved in these regulatory interactions, RNA half-lives were acquired for cells growing in acetate, methanol and trimethylamine using full genome transcriptome profiles (RNAseq) at several timepoints after halting transcription. The data demonstrates drastic increases in RNA stability for growth on the low energy substrate acetate, however this effect is not uniform across different functional classes of enzymes, suggesting the organism has developed a mechanism to selectively stabilize RNA. RNAs involved in, or encoding proteins in, energy production metabolism, cell wall and lipid metabolism, coenzyme metabolism and signal transduction were generally stabilized, while those involved in translation and ribosome biogenesis were generally destabilized. These half-lives, along with data on differential gene expression is allowing the prediction of important shifts in metabolic pathway

usage. Work to integrate regulation data with the metabolic models to predict dynamic response of the organisms to their environment is ongoing.

Single-Cell Genomics of Pacific Ocean Anoxic Marine Zones Bacterial & Archaeal "Dark Matter"

A.M. Plominsky* (alvaro.munoz@imo-chile.cl)[1,2], S. Podell[3], E.E. Allen[3,4], O. Ulloa[1,2]

[1]Departamento de Oceanografia, Universidad de Concepcion, Concepcion, Region del Bio-Bio, Chile; [2]Instituto Milenio de Oceanografia, Chile; [3]Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, California, USA; [4]Division of Biological Sciences, University of California, San Diego, California, USA

Poor ventilation and high O_2 demand during the degradation of organic matter generate O_2 deficient intermediate layers in the ocean. In some cases, the O_2 concentration drops below the detection limit of the most sensitive sensors (<5 nM dissolved O_2) where -due to nitrate reduction- nitrite (NO_2^-) accumulates, creating permanent Anoxic Marine Zones (AMZs). This is the case in the Arabian Sea and in the Eastern Tropical North and South Pacific (ETNP and ETSP, respectively). AMZs present diverse microbial communities, dominated by anaerobic, ammonium- and sulfur-oxidizing bacteria at their core, and aerobic ammonia-oxidizing archaea at their boundaries (oxyclines). However, cultured isolates and sequenced genomes for both the dominant and less-abundant microorganisms inhabiting the AMZs are lacking.

The aim of our study was thus to generate reference genomes for un-described AMZ bacterial and archaeal taxa, through single-cell genome sequencing. We targeted microbial communities inhabiting the upper oxycline, secondary chlorophyll maximum within the AMZ, anoxic core (NO_2-peak) and lower oxycline of the ETNP and ETSP AMZs. Several hundred single-cell amplified genomes (SAGs) were generated, and their 16S rRNA genes were screened to select initially for taxa with no closely related sequenced genomes. The first 36 SAGs selected for whole genome sequencing consisted of 33 Bacteria and 3 Archaea. Most of these bacterial SAGs were from the phylum Proteobacteria (21 SAGs from 8 different taxa), Bacteriodetes (4 SAGs from 2 taxa) and Marinimicrobia (4 SAGs from 1 taxon), although individual SAGs from the phylum Acidobacteria and candidate divisions OP11, OP3 and OD1 were also sequenced. The Archaeal SAGs belonged to the Marine Benthic Group A (1 SAG) and Marine Group IIB (2 SAGs), of the Thaumarchaea and Euryarchaea phyla, respectively.

The distribution of these taxa across the O_2-gradient of the ETSP AMZ was assessed by the recruitment of metagenomic reads to taxonomically-binned (decontaminated) SAG contigs. This analysis confirmed that most of these taxa are rare in their natural environment, and that they partition along the O_2 and nutrient gradients. Information generated through gene prediction and annotation will serve as a baseline to assess the metabolic potential of this AMZ "Microbial Dark Matter". Future work will also include the sequencing of SAGs of microorganisms that are known to play important biogeochemical roles in AMZ, but for which reference genomes are lacking.

Exploring a high-CO_2 subsurface environment using metagenomics and single cell genomics

Alexander J. Probst* (probst@berkeley.edu)[1], Joanne Emerson[1], Brian C. Thomas[1], Jessica Jarett[2], Tanja Woyke[2], Susannah Tringe[2], and Jillian F. Banfield[2]

[1]Department of Earth and Planetary Sciences, University of California, Berkeley, CA 94720 USA; [2]DOE Joint Genome Institute, Walnut Creek, CA 94598 USA

Previous work has demonstrated substantial (phylo)genetic and metabolic novelty in Earth's subsurface, yet subsurface microbial life remains vastly unexplored. Here, we report a metagenomic study of a cold, CO_2-driven aquifer that represents a window to the deep subsurface. Microbial samples were taken by sequential filtration through different filter sizes and yielded high-quality DNA for metagenomic sequencing. Binning of assembled sequences resulted in more than 100 genomes with a completeness of at least 70%. Presented are also the results of a detailed study of a highly abundant autotrophic archaeon from this subsurface environment. Notable in this system is the high incidence of pathways involved in CO_2 fixation. In ongoing work, a parallel single cell genomics study will further expand understanding of microbial diversity and function in this environment. The data of this investigation are of substantial importance for understanding subsurface ecosystems and their impact on global carbon cycling.

Using single type cell-root hair cell to decipher soybean and sorghum epigenomic and transcriptomic response to adverse environment changes at molecular level

Zhenzhen Qiao* (zhenzhen.qiao-1@ou.edu), Jennifer Brooks, Marc Libault, Mehrnoush Rey, Michelle Wood

Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, 73019, USA.

To adapt to the diverse environmental changes, plants are capable of adjusting themselves morphologically, physiologically and at the molecular level. These changes are, for a part, dependent of the transcriptional activity of the genes and epigenomic modifications. Over the past decades, plant transcriptional and epigenomic responses to adverse environments have been characterized. However, the cellular complexity of the organs used in these studies is a limitation to precisely understand the relationships existing between epigenomic and transcriptomic changes in response to various stresses.

In our project, to overcome this difficulty, we use root hair cells from soybean and sorghum as single cell type models to characterize their epigenomic and transcriptomic regulation in response to various environmental stresses. Working with two evolutionary distant model plants will facilitate our understanding of their differential adaptation to environmental stresses. In addition, taking advantage of the characterization of soybean homeologous genes, we are able to investigate the impact of the duplication of the plant genome on gene expression and plant adaptation to stresses.

To mimic the adverse environments, we treated the plants with different stresses including nutrients deficiency (i.e., nitrogen, potassium, ion and phosphorous deprivation), high and low temperatures and pHs, drought, salinity, and elevated atmospheric CO_2 concentration using a unique growth system: an ultrasound aeroponic system. This system allows a homogeneous and controlled treatment of the plants as well as a direct access to the root hair cells. Afterword, we collected root hair cells preliminary to the isolation of total RNA, gDNA and smRNA. These nucleic acids are currently sequences using JGI high-throughput sequencing facilities.

First evidence of *Helitrons* transposable elements insertion in fungi.

Lucía Ramírez* (lramirez@unavarra.es), **Alessandra Borgognone**, and Antonio G. Pisabarro

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

Helitrons represent a superfamily of DNA transposons containing specific hallmarks, discovered by computational analysis of eukaryotic genomes. The presence of these elements was recently uncovered by computational and experimental analyses in *Pleurotus ostreatus*, a lignin-degrading basidiomycete of increasing interest for different industrial applications. Previous studies based on sequence analyses showed that helitrons can capture, amplify and express captured host genes, playing an important role on the evolution of many eukaryotes genomes. Putative autonomous helitrons carry a motif similar to the replication initiator (Rep) of plasmid rolling circle replicons, as well as a DNA helicase (Hel) domain. According to the proposed transposition method, these elements belong to Class II transposable elements and mobilize through a rolling circle replication (RC) instead of the cut-and paste mechanism used by other class II transposons. Despite the novelty of these elements, they have been poorly studied due to the lack of a system to detect their transposition. In fact, no evidence of a helitron mobilization has ever been shown under laboratory conditions. Here we describe the first evidence of a helitron somatic transposition in the chromosome I of the *P.ostreatus* strain PC15 obtained by dedikaryotisation of the N001 strain. Results obtained by PCR analysis and southern blot hybridisation and confirmed by sequencing data, show that an event of helitron insertion lacking of target site duplication was detected in the monokaryotic strain PC15. The absence of this helitron was confirmed in N001, the parental strain of PC15, as well as in its progeny were no evidence of helitron located in this chromosomal site has been detected.

These results suggest the influence of replicative factors on transposition events, the hypothetical causes that determine the mobilization of these elements and may highlight their influence on the genomic organization and gene expression.

Trancriptome profile of the interaction between arbuscular mycorrhizal fungi and common bean (*Phaseolus vulgaris* L.) roots under drought stress

Gustavo Henrique Recchia* (ghrecchia@cena.usp.br), Enéas Ricardo Konzen, Fernanda Cassieri, Siu Mui Tsai

Center of Nuclear Energy for Agriculture - University of Sao Paulo; Piracicaba/SP, Brazil.

Common bean (Phaseolus vulgaris L.) productivity has been severely affected by drought stress episodes due to low technology cultivation systems, especially in developing countries. Plants usually establish mutualistic symbiotic relationships with Arbuscular Mycorrhizal Fungi (AMFs) that are vital for nutrient and water absorption. On the other hand AMFs benefit from plants by absorbing photosynthesized organic compounds. This interaction leads to a differential process of regulation of genes that are pivotal for the metabolic pathways implicated on water absorption in roots, and as a result in drought stress responses. In the present study, the drought tolerant common bean genotype BAT477, inoculated with a mixture of three different AMF species (*Glomus clarum, Acaulospora scrobiculata* and *Gigaspora rosea*), was submitted to a 10 days watering interruption treatment during its pre-flowering stage in a greenhouse experiment. Total RNA was extracted from the roots of 96 hours drought stressed plants,

and well-watered controls, both inoculated and non-inoculated, consisting of four treatments and three biological replicates. A RNA-seq analysis was performed following the HiScan™Sq System (illumina®) platform pipeline. The transcriptomes were mapped against the reference genome of *P. vulgaris* (Pvulgaris128 - Phytozome v.9.1), and statistically compared for differentially regulated transcripts (FDR<0.05). The analysis revealed 12,087 transcripts differentially regulated in the inoculated plants under stress against control treatments, and 11,939 transcripts in non-inoculated plants against the control. From these sets, 21 transcripts were statistically significant regulated comparing the inoculated treatment against the non-inoculated treatment both under drought, being 17 of them upregulated in roots under the influence of AMFs and four downregulated. Functional annotations were obtained through the *green plant* database of the Gene Ontology Consortium. Two functional classes were highlighted: cell-wall transporters and DNA-binding transcription factors. Three sets of transcripts, aquaporins, NAC and DREBs transcription factors, totalizing 25 transcripts, were selected and had their gene expression profile analyzed through RT-qPCR comparing the four treatments and total RNA obtained from roots and leaves harvested along the five days of imposed drought (after 24, 48, 72, 96 and 120 hours). This study validated the RNA-Seq experiment and provided new insightful information, since two aquaporins transcripts [*PvPIP1;3* (31.19 fold change) and *PvPIP2;1* (4.96 fold change)] were exclusively upregulated in leaves of inoculated plants under drought stress, attesting the AMF influence is not restricted to roots. This study helps to provide a more detailed transcriptome database that not only contributes to enrich the dataset of genes involved on drought stress response in common beans, but also strengthens the understanding of complex genetic relationships established between plants and their AMF partners, and most importantly how this contributes to the improvement of drought tolerance in plants.

This work was supported by FAPESP, CAPES, CNPq.

Genomic evolution of the ascomycetous yeasts

Robert W. Riley* (rwriley@lbl.gov)[1], Sajeet Haridas[1], Asaf Salamov[1], Kyria Boundy-Mills[2], Markus Göker[3], Chris Todd Hittinger[4], Hans-Peter Klenk[5], Mariana Lopes[4], Jan P. Meier-Kolthoff[3], Antonis Rokas[6], Carlos Rosa[7], Carmen Scheuner[3], Marco Soares[4], Benjamin Stielow[8], Jennifer H. Wisecaver[6], Ken Wolfe[9], Meredith Blackwell[10], Cletus Kurtzman[11], Igor Grigoriev[1], Thomas Jeffries[12]

[1]US Department of Energy Joint Genome Institute, Walnut Creek, CA; [2]Department of Food Science and Technology, University of California Davis, Davis, CA; [3]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, GERMANY; [4]Laboratory of Genetics, Genetics/Biotechnology Center, Madison, WI; [5]School of Biology, Newcastle University, Newcastle upon Tyne, UK; [6]Department of Biological Sciences, Vanderbilt University; [7]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; [8]CBS-KNAW Fungal Biodiversity Centre, Utrecht, Netherlands; [9]UCD School of Medicine & Medical Science, Conway Institute, University College Dublin, Dublin, Ireland; [10]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA; [11]USDA, ARS, MWA, NCAUR, BFPM, Peoria, IL; [12]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI

Yeasts are important for industrial and biotechnological processes and show remarkable metabolic and phylogenetic diversity despite morphological similarities. We have sequenced the genomes of 16 ascomycete yeasts of taxonomic and industrial importance including members of Saccharomycotina and Taphrinomycotina. Phylogenetic analysis of these and previously published yeast genomes helped resolve the placement of species including Saitoella complicata, Babjeviella inositovora, Hyphopichia

burtonii, and Metschnikowia bicuspidata. Moreover, we find that alternative nuclear codon usage, where CUG encodes serine instead of leucine, are monophyletic within the Saccharomycotina. Most of the yeasts have compact genomes with a large fraction of single exon genes, and a tendency towards more introns in early-diverging species. Analysis of enzyme phylogeny gives insights into the evolution of metabolic capabilities such as xylose fermentation, methanol utilization, and assimilation of alternative carbon sources.

Assessing the adaptation of the soil fungal community to global warming

Adriana L. Romero-Olivares* (alromer1@uci.edu)[1], Steven D. Allison[1], John W. Taylor[2], Kathleen K. Treseder[1]

[1] Department of Ecology and Evolutionary Biology, University of California-Irvine, Irvine, CA 92697, USA; [2]Department of Plant and Microbial Ecology, University of California-Berkeley, Berkeley, CA 94720-3102, USA

The capacity of fungi to adapt to global warming might have important long-term consequences for global carbon cycling. For example, fungi could adapt to warming by increasing their carbon use efficiency, which could reduce the release of CO_2 into the atmosphere, resulting in a negative feedback on climate change. Nevertheless, few studies have examined evolutionary responses of fungal taxa to warming.

We examined the adaptation of the model fungus *Neurospora discreta* to moderately warm (16 °C) and warm (28 °C) temperatures in the lab. We examined its physiological changes after 1500 mitotic generations. We tested the hypothesis that adapted strains will have higher fitness than parental strains when grown in the selective environment. More specifically, adapted strains would produce less biomass, owing to tradeoffs in the allocation of resources towards sporulation; this tradeoff would be accompanied by lower production of CO_2 due to higher efficiency in carbon use metabolism.

We found partial support for our hypothesis. *N. discreta* seemed to adapt to warm temperatures based on patterns of spore and biomass production. However, contrary to our hypothesis, adaptation was not accompanied by down-regulation of CO_2 production. In fact, CO_2 production significantly increased in the adapted strains. We concluded that adaptation to warm temperatures selected for greater spore production to increase fitness in the selective environment. Since spores are energetically costly to produce, this change led to increased CO_2 respiration and a decline in growth rate. Our results did not support the idea that adaptation to warm temperatures will lead to a more efficient carbon use metabolism.

With our JGI project, we will expand our research to study the potential adaptation of the fungal community inhabiting a boreal forest that has been experimentally warmed for nine years. We will identify the mechanisms underlying adaptation by using a meta-transcriptomics approach. Specifically, we will analyze the expression profiles of genes related to decomposition of different carbon substrates, as well as genes relevant for carbon metabolism, growth, and reproduction. We will use our findings in *N. discreta*, as well as annotated genomes of other sequenced model fungi, as baseline to analyze our data. In addition, we are culturing multiple isolates from the same taxonomic family, from warmed and unwarmed soils, in order to compare their genomes and further determine if the fungal community is adapting to global warming.

Mechanisms of Vitamin Exchange in a 20 Member Microbial Consortia

Margaret Romine* (Margie.romine@pnnl.gov)[1], Lindsey Anderson[1], Premchendar Nandhikonda[1], Bill Nelson[1], Dmitry Rodionov[2], and Aaron Wright[1]

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]Sanford-Burnham Medical Research Institute, La Jolla, California

We have used metagenome sequence to assemble complete or near-complete genomes from nearly all the members (~20) of two unicyanobacterial consortia (UCC) developed from microbial mats associated with a saline ($MgSO_4$) lake in northern Washington State. These consortia have been maintained under constant light in defined media containing salts, trace minerals, and CO_2 as the primary carbon source. This defined system has allowed us to use single-organism metabolic reconstruction strategies to infer member function and metabolic requirements and to predict interactions among members and identify the genes and processes involved. We used the assembled genomes and functional predictions along with activity-based chemical probes (ABPs) to identify potential vitamin producers and consumers as well as underlying mechanisms responsible for vitamin exchange among community members.

Vitamin auxotrophy and prototrophy provides the basis for a common metabolic interaction between microbial phototrophs and heterotrophs. By analyzing our UCC we are able to provide the first insights into how auxotrophy can be maintained in a self-sustaining multi-member community. As no external source of vitamin is provided to our cultures, auxotrophs must acquire this resource from other members or employ a metabolic strategy that obviates the need for vitamins that they cannot synthesize. Genomic analysis revealed that only the cyanobacterial and Halomonad members are capable of synthesizing all coenzymes derived from B vitamins. While all members could produce FAD/FMN (B_2) and pyridoxine (B_6), auxotrophy for the remaining B vitamins are widespread. By identifying vitamin-requiring metabolic processes in each member we were able to uncover an unexpected relationship between the ability to produce B_{12} and its requirement for essential processes. In general, the auxotrophs encode more enzymes that require B_{12} than the five producers do. Although the two *Halomonas* sp. can synthesize B_{12}, they only require it for growth on ethanolamine suggesting a potential role for these organisms as producers of this commodity. Genomic analysis revealed that B_{12} likely acts as a regulatory sensor in the phototrophic *Rhodobactericeae* and *Algoriphagus marincola*, controlling photosynthesis and carotenoid biosynthesis, respectively. Its dual role as a sensor and coenzyme suggests that it may be a key determinant controlling community dynamics.

We applied a B_{12}-based ABP to the UCC isolate, *Halomonas* HL-48, to capture and identify B_{12}-interacting proteins. Proteomic analysis of captured proteins revealed an additional B_{12}-binding regulator, PhrR. Regulon analysis identified putative PhrR DNA-binding targets and suggest that PhrR controls multiple processes including biosynthesis of cyclopropane fatty acids and folate, chemotaxis, and photolyase. Besides this regulator and four known B_{12}-dependent enzymes, the B_{12}-ABP also captured enzymes that use or enable recycling of S-adenylosyl methionine (SAM) or that are involved folate-mediated transfer of one carbon units for anabolic metabolism. Collectively, these findings suggest that balanced vitamin allocation among community members is achieved through differential regulation of vitamin-requiring metabolic processes.

What Makes Cyanobacteria Tick: Understanding Promoter Regulation in *Synechococcus* sp. PCC 7002

Anne M. Ruffing* (aruffin@sandia.gov) Lucas M. Strickland, Travis J. Jensen

Bioenergy and Defense Technologies, Sandia National Laboratories, Albuquerque, NM 87123, USA

As the first oxygenic phototrophs, cyanobacteria are believed to have played a key role in shaping the Earth, and they are also the host of many proposed 'green' chemistry processes, including biofuel production. Therefore, to understand how cyanobacteria contribute to current ecosystems and also how to best engineer cyanobacteria for fuel and chemical syntheses, we must understand the mechanisms and motifs contributing to promoter regulation in cyanobacteria. In this study, we used previous RNA-seq data of the cyanobacterium *Synechococcus* sp. PCC 7002 to select promoter regions of genes showing various levels of expression (10^{-5} – 10^{-2} counts/total counts) and regulatory patterns (constitutive, linear phase, and stationary phase). A 500 bp region upstream of these genes was selected as the putative promoter region and integrated upstream of a yellow fluorescent protein (Ypet) with the transcription terminator region from the RuBisCo operon. The synthetic operons were then integrated into a putative neutral site in *Synechococcus* sp. PCC 7002. A total of 24 promoter mutants were evaluated for Ypet fluorescence under both continuous and diurnal light conditions. Additionally, the promoter sequences were analyzed for common motifs correlating with the level of observed Ypet fluorescence. We identify several putative motifs associated with diurnal light expression, regulatory patterns of expression, and overall promoter strength.

Reconstruction of ancestral ligninolytic peroxidase sequencesfrom ten Polyporales genomes

Francisco J. Ruiz-Dueñas* (fjruiz@cib.csic.es), **Ivan Ayuso**, Angel T. Martínez,

Centro de Investigaciones Biológicas, CSIC, Ramiro de Maeztu 9, E-28040-Madrid, Spain

The sequencing and analysis of 31 fungal genomes allowed the dating of the evolutionary origin of lignin degradation in the Paleozoic period (1). It was also determined the ancestral state of discrete characters in class-II peroxidases (PODs, including lignin peroxidase, manganese peroxidase, versatile peroxidase and generic peroxidase families) such as the appearance or disappearance of the oxidizing sites that determine the catalytic activities of these enzymes. Ancestral state reconstruction was subsequently focused on PODs from white-rot fungi of the order Polyporales (2) using 10 of the genomes included in the Saprotrophic Agarycomycotina Project (SAP) coordinated by David Hibbet (Clark University). However, both studies only analyzed discrete characters instead of sequences. We propose the reconstruction not only of those discrete characters, but also of the whole proteins by inferring their ancestral amino acid sequences using the information extracted from extant PODs. The evolution of gene function is a central issue in molecular evolution and ancestral sequence reconstruction is a powerful tool for these studies. It has the advantage of inferring ancient gene sequences whose properties can then be tested in laboratory by resurrection of the reconstructed sequences. During the past years, the resurrection of genes has contributed to prove several hypothesis about how life was on Earth, and how some enzymes evolved and obtained the functions they have today. Resurrected enzymes are of interest not only because of the basic information they provide about evolution of a specific family of proteins, but also because they could have a great biotechnological potential. The ancestral PODs catalyzed reactions in a planet where the conditions were very different from actual Earth, with different temperature, pH, oxidation conditions, etc. Thereby, by resurrection and characterization of those sequences in the laboratory we will be able to determine the mechanisms that led the ancient peroxidases to the functions and properties they have today, including the ability to degrade the recalcitrant lignin polymer, a key issue for development of land ecosystems. Here we present the bioinformatic analysis of the resurrected sequences of ancestral PODs, corresponding to the

nine main nodes identified during the evolution of these enzymes in Polyporales, with special interest in the oxidizing sites and their modification through the evolution.

1. Floudas, D., M. Binder, R. Riley, K. Barry, R. A. Blanchette, B. Henrissat, A. T. Martínez, R. Otillar et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336:1715-1719.

2. Ruiz-Dueñas, F. J., T. Lundell, D. Floudas, L. G. Nagy, J. M. Barrasa, D. S. Hibbett, and A. T. Martínez. 2013. Lignin-degrading peroxidases in Polyporales: An evolutionary survey based on ten sequenced genomes. Mycologia 105:1428-1444.

Comparative analysis of transcription factor families across fungal tree of life.

Asaf Salamov* (aasalamov@lbl.gov) and Igor Grigoriev

DOE Joint Genome Institute, Walnut Creek CA 94598

Transcription factors (TFs) are proteins that regulate the transcription of genes, by binding to specific DNA sequences. We have analysed the distribution and evolution of 62 known TF families in more than 300 fungal genomes from Mycocosm portal (http://jgi.doe.gov/fungi/).

We have shown that while fungal-specific TF families, like Zinc finger (Zn2Cys) constitute the largest fraction of TFs repertoire in most fungal genomes, particularly greatly expanding in Pezizomycotina clade of Ascomycota, the universal eukaryotic TFs, like HLH, Homeobox, bZIP, GATA and others, are more abundant in Mucoromycotina and other early-divergent clades. We discuss the different evolutionary pathways of individual TF families.

Evolution of *Escherichia coli* to 42 °C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations

Troy E. Sandberg* (tsandberg@ucsd.edu)[1], Margit Pedersen[2], Ryan A. LaCroix[1], Ali Ebrahim[1], Mads Bonde[2], Markus J. Herrgard[2], Bernhard O. Palsson[1,2], Morten Sommer[2,3], and Adam M. Feist[1,2]

[1]Department of Bioengineering, UCSD, La Jolla, CA; [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark; [3]Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark

Adaptive laboratory evolution (ALE) has emerged as a valuable method by which to investigate microbial adaptation to a desired environment. Here, we performed ALE to 42 °C of ten parallel populations of *Escherichia coli* grown in glucose minimal media. Tightly controlled experimental conditions allowed selection based on exponential-phase growth rate, yielding strains that uniformly converged toward a similar phenotype along distinct genetic paths. Adapted strains possessed as few as 6 and as many as 55 mutations, and of the 144 genes that mutated in total, 14 arose independently across two or more strains. This mutational recurrence pointed to the key genetic targets underlying the evolved fitness increase. Genome engineering was used to introduce the novel ALE-acquired alleles in random combinations into the ancestral strain, and competition between these engineered strains reaffirmed the impact of the key mutations on the growth rate at 42 °C. Interestingly, most of the identified key gene targets differed significantly from those found in similar temperature adaptation studies,

highlighting the sensitivity of genetic evolution to experimental conditions and ancestral genotype. Additionally, transcriptomic analysis of the ancestral and evolved strains revealed a general trend for restoration of the global expression state back toward pre- heat stressed levels. This restorative effect was previously documented following evolution to metabolic perturbations, and thus may represent a general driver of evolutionary adaptation. The widespread evolved expression shifts were enabled by a comparatively scant number of regulatory mutations, providing a net fitness benefit but causing suboptimal expression levels for certain genes, such as those governing flagellar formation, which then became targets for additional ameliorating mutations. Overall, the results of this study provide insight into the adaptation process and yield lessons important for the future implementation of ALE as a tool for biological discovery and engineering.

Microbial Responses to Extreme Disturbance: Insights from the Ongoing Centralia Coal Mine Fire

Ashley Shade* (shade.ashley@gmail.com), Tammy C. Tobin, Taylor K. Dunivin, Sang-Hoon Lee, Jackson W. Sorensen, and Joshua R. Herr

Michigan State University, Microbiology and Molecular Genetics, Biomedical Physical Sciences

The Centralia, Pennsylvania coal seam fire has been burning near-surface since 1962. It has created an extreme environment that supports coal fire-adapted microbial life. Heat, steam and combustion products (CO, CO_2, SO_x and NO_x) vent upward from the fire through the overlying soils, increasing soil surface temperatures to over 80°C. Soil chemistry is altered by both spontaneous and microbially-mediated chemical reactions. As the fire expands into new areas, it also retreats from some affected sites, which can then recover. This unusual habitat provides an opportunity to investigate the selective pressures and community processes that promote microbial community stability (resistance, resilience) in the face of extreme, ongoing disturbance. It also offers an opportunity to investigate the ecology of thermophiles (e.g., Chloroflexi, Firmicutes and Armatimonadetes) that are typically rare in temperate soils, but have become prevalent in these fire-affected soils. We explored the soil properties and microbial community structure along two chronosequences of historical fire activity and recovery. From these soils, we measured soil chemistry and temperature, extracted bulk soil DNA for sequencing, and cultivated thermophilic and arsenic tolerant microorganisms. We report results from our first sequencing endeavors, including 16S-tag sequencing along two fire fronts and one metagenome from an active surface vent. We also readily cultivated arsenic tolerant microorganisms (minimum inhibitory concentrations 25 to >200 mM), inclusive of isolates affiliated with *Bacillus*, *Enterobacter*, *Stenotrophomonas*, and *Acinetobacter* lineages that had genes for arsenite efflux pumps (*arc3(2)* and *arsB*) and could grow on arsenate-amended media (10 mM). We conclude with future directions for our ongoing research in Centralia, as supported in part by the JGI's CSP for metagenome sequencing.

Keywords
Microbial diversity, thermophiles, microbial community ecology, rare biosphere, soil microbiology, metagenomics, extreme environment, disturbance

Complex dynamics due to geochemical transitions revealed by time-series metagenomics in aquifer microbial communities

Itai Sharon* (itaish@berkeley.edu)[1], Laura A. Hug[1], Rose S. Kantor[1], Brian C. Thomas[1], Kelly C. Wrighton[1,+], Kim M. Handley[1,^] and Jillian F. Banfield[1,2]

[1] Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA 94720, USA; [2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [+] Present address: Department of Microbiology, The Ohio State University, Columbus, Ohio, USA; Present address: Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA.

Microbes that inhabit aquifer sediments are important contributors to geochemical processes taking place in these environments. An accurate understanding of microbial activity and community response to perturbations in sediments is currently mostly lacking, and could inform bioremediation efforts. Here we studied microbial community response to organic carbon (acetate) stimulation of aquifer groundwater in Rifle, CO. Thirteen metagenomic samples were collected over a time period of 63 days after stimulation at times of geochemical interest. 78 microbial genomes were reconstructed at an average genome completeness of 87% using time series binning techniques. Dozens of phage genomes were also recovered as well as hundreds of mobile elements. Complete and near complete genomes were reconstructed for members of several candidate phyla, including TM7 (Saccharobacteria), OD1 (Parcubacteria), WWE3 and SR1. Significant changes in community composition were observed during iron- and sulfate- reduction phases following acetate amendment. The most abundant community members during the preliminary phase of iron reduction belonged to Delta- and Betaproteobacteria as well as Elusimicrobia. These organisms were replaced by members of Alpha- and Deltaproteobacteria, OP9 (Atribacteria), Firmicutes, SR1, Spirochaetes and Tenericutes as the system transitioned into a sulfate reduction phase. Genomic analyses suggest that only some of these community members responded directly to the changing conditions while the abundance of others may have changed due to interactions with these and other members. *Geobacter,* previously shown to be most abundant during the iron reduction phase, was represented by several members of the genus that exhibited different abundance patterns. Results overall provide a detailed description of community response to changes in the environment.

An improved assembly algorithm for *de novo* circular genome reconstruction

Helen Shearman* (helen@biomatters.com)[2], **Christian Olsen**[1], K. Qaadri[1], R. Moir[2], M. Kearse[2], S. Buxton[2], Man Kit Cheung[2], J. Kuhn[2], S. Stones-Havas[2], C. Duran[2]

[1]Biomatters, Inc. Newark, NJ; [2]Biomatters, Ltd. Auckland, New Zealand

Circular chromosomes or genomes, such as viruses, bacteria, mitochondria and plasmids, are a common occurrence in nature, but despite the wide array of algorithms available for *de novo* assembly, the circularity of these DNA molecules is largely overlooked. There are some limitations of this oversight. Current NGS assembly algorithms assume a linear molecule and will result in a linearly represented genome, with a breakpoint in an arbitrary position. This is becoming increasingly problematic with increasing NGS read lengths resulting in fewer contig assemblies with less ambiguity. Additionally, long read sequencing technologies promise to provide sequences that may greatly span breakpoints at the expense of coverage - resulting in a relatively large quanta of information loss.

Although there are currently methods for the re-circularisation of contigs post-assembly by identifying common trailing/leading sequence motifs, we present a more robust approach of circularising during the assembly process whilst still allowing the merging of similar and sub-contigs throughout the overlap-based approach. This method can also combat the issue of chimeric sequence due to contamination, a common problem when sequencing bacterial cultures, due to decreased likelihood of conserved contig ends due to timely circularisation.

We present results from both 28.6 million read *Pan troglodytes* illumina data set and 267,491 read *Panthera leo persica* (Asiatic Lion) mitochondrial NGS library produced using an Ion Torrent sequencing machine. These results are discussed and compared to some of the more popular linear assembly algorithms in common usage today.

Geneious R8 is the first bioinformatics software package to offer a circular *de novo* assembly method. The Geneious Circular *de novo* assembler is developed by Biomatters and may be found at http://www.geneious.com

Autonomous Adaptive Sampling Of Microbial Processes In A Dynamic Estuary

Holly M. Simon* (simonh@oh,su.edu)[1,2] **Lydie Herfort**[1], Charles Seaton[1], Michael J. Wilkins[1], and António M. Baptista[1,2]

Center for Coastal Margin Observation & Prediction[1] and Institute of Environmental Health[2], Oregon Health & Science University, Portland, Oregon, USA

The Columbia River has a dynamic and fast flushing estuary impacted by strong advection and mixing of riverine and oceanic waters, and high but variable load of suspended particulate matter. Transient, but re-occurring water and nutrient fluxes from end-members impart strong spatial and temporal gradients, contributing to microbiological hotspots that play important ecological roles in the estuary. Investigations of the corresponding microbiota require precisely timed samples that are contextualized by physical and biogeochemical data. To accomplish this, we embedded a robotic microbial sampler (Environmental Sample Processor, ESP) within the operations of an interdisciplinary observation and prediction system (SATURN; www.stccmop.org/saturn). The implementation of autonomous, adaptively sampled water collection by the ESP was based on environmental conditions assessed from SATURN physical and biogeochemical sensors. Water was pumped from multiple depths to sensors and the ESP on dry land. If water met user-defined parameters, ESP sampling was automatically initiated. This strategy was tested during three deployments in summer 2013, during which operational tools for analysis and visualization were used to formulate well-constrained mission plans by providing estimates of the intensity and timing of estuarine turbidity maxima (ETM) and intrusion of oxygen-depleted ocean water. This allowed for effective characterization of the impact of these events on selected estuarine microbiota. For example, quantitative PCR analysis indicated that intrusion of upwelled oceanic water into the lower estuary resulted in higher numbers of archaeal ammonia oxidizers due, at least in part, to their subsequent entrapment in ETM. Ongoing efforts to characterize the impacts of physical forcing on microbial processes in the estuary are being addressed in a collaborative Community Sequencing Project with scientists at the DOE Joint Genome Institute. Comparative metagenomics and metatranscriptomics will be used to test the idea that episodic forcing events impact microbial population structure and activities differentially during periods of high- and low-nutrient and energy fluxes in the estuary. Initial results from extraction and purification of ESP-collected water samples show nucleic acid recoveries are within the desired ranges for sequencing and indicate feasibility of the approach.

Updates to the RNA-Seq Gene Expression Analysis pipeline at the Joint Genome Institute

Vasanth Singan* (VRSingan@lbl.gov), Anna Lipzen, and Erika A. Lindquist

DOE-Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA

Transcriptome sequencing and analysis spans all scientific programs at the JGI. Gene expression profiling of transcriptomes remains a core competency at the JGI. Recent years have seen a deluge of transcriptomics studies using Next-Generation Sequencing technologies including RNA-Seq. Concurrent with the development in technology, a variety of analysis software and algorithms are being developed. Here at JGI, we strive towards constantly updating and testing our pipeline to adapt and improve the analysis. In this poster, we discuss the analyses and changes to our pipeline and deliverable to users.

Our updated RNA-Seq gene expression analysis pipeline is designed to work with multiple libraries in parallel. It incorporates three different read aligners including BWA, TopHat and HISAT. Based on the complexity and need of the project, the aligner can be selected. The pipeline produces a read counts table and FPKM normalized counts table for each of the genes/genomic features. A comparison of various open source Differential Gene Expression (DGE) tools (Cufflinks2, DESeq2 and EdgeR) was performed and based on the results we have incorporated DESeq2 as our tool of choice.

The deliverables include BAM alignment files along with the references used, a read counts table, FPKM normalized read counts tables, a DGE_summary table along with condition pairwise results and a visualized heatmap of replicate correlation to detect contamination in samples. Optionally, we provide depth track(s) for fungal projects through Mycocosm – JGI Genome Portal (http://genome.jgi-psf.org/programs/fungi/index.jsf).

Less carbon-intensive energy sources are needed to reduce greenhouse gas emissions and their predicted role in climate change

Esther Singer* (esinger@lbl.gov), Thomas E. Juenger, Tanja Woyke

DOE Joint Genome Institute, Walnut Creek, CA, USA

There is growing interest in the potential of biofuels for meeting this need. A critical question is whether large-scale biofuel production can be ***sustainable*** over the time scales needed to mitigate our carbon debt from fossil fuel consumption. The carbon balance and ultimately the sustainability of biofuel feedstock production is the result of complex climate-coupled interactions between carbon fixation, sequestration, and release through combustion. Similarly, the long-term productivity of biofuels depends on the environmental factors limiting plant growth. These factors are often related to soil resources which involve complex interactions at the plant-microbe-soil interface impacting their availability and cycling. We have used 16S and ITS iTAG sequencing (with highly successful PNA plastid blocking) in an initial exploration of the bacterial and fungal communities on and within leaf and root surfaces. Our study included 6 biological replicates of 4 genotypes (2 upland [VS16 & Dakota] and 2 lowland ecotypes [AP13 & WBC]) grown in the presence and absence of nitrogen fertilization and sampling 4 compartments (phyllosphere, rhizosphere, leaf and root endophytic compartments). We discovered thousands of fungi and bacterial OTUs composing the microbiome with clear indicator taxa for sampled compartments. Moreover, our analysis suggests that host attributes associated with upland/lowland ecotypic divergence play an especially important role in structuring diversity and abundance in comparison to soil nitrogen abundances.

Relating the diversity, abundance, and activity of ammonia-oxidizing archaeal communities to nitrification rates in the coastal ocean

Jason M. Smith* (jsmith@mabari.org)[1], Christopher A. Francis*[2,] and Bradley B. Tolar[2]

[1]Biological Oceanography Group, Monterey Bay Aquarium Research Institute, Moss Landing, CA; [2]Environmental Earth System Science, Stanford University, Stanford, CA

Nitrification, the microbial oxidation of ammonia to nitrate, plays a critical role in the marine nitrogen cycle. In the ocean's surface layer, this process alters the distribution of inorganic nitrogen species available to phytoplankton and produces N_2O. The first step of nitrification—the oxidation of ammonia to nitrite—is now believed to be carried out primarily by ammonia-oxidizing archaea (AOA) throughout the ocean. Our previous work in Monterey Bay (MB) and the contiguous California Current System (CCS) has demonstrated a strong relationship between the abundance of archaeal ammonia monooxygenase α-subunit (*amoA*) genes and transcripts and rates of nitrification. Our ongoing CSP project seeks to determine whether variability in the abundance, genomic diversity or cellular activity of AOA relates to observed changes in nitrification rates, or the prevailing 'metabolic state' of a given water mass, in MB or the CCS. Across a composite library of 94 samples and more than 9 million 16S rRNA reads, we observed remarkable agreement between the relative abundance of Marine Group I (MGI) *Thaumarchaeota* 16S rRNA reads (as % of total) and the absolute abundance of AOA *amoA* genes determined via qPCR. In addition, relative changes in the abundance of MGI 16S rRNA genes both with depth and between stations were in close agreement with changes in nitrification rates. Combined with our previous work, these findings suggest 16S rRNA tag sequence data to be a powerful tool for the study of AOA abundance and diversity, and the distribution of nitrification rates, in the ocean. Furthermore, we used our extensive 16S rRNA dataset to select >15 samples of particular interest (collected from multiple stations and depths, ranging from 10 to 80 m), for detailed metagenomic and metatranscriptomic analysis. With much of this sequence data now in hand, we have the unprecedented opportunity to examine AOA metagenomes and metatranscriptomes in oceanic samples for which nitrification rates (and other key biogeochemical activities and parameters) have also been measured.

Plant mutants affecting root carbon allocation alter the root microbiome

Gary Stacey* (staceyg@missouri.edu)[5], **Ljiljana Paša-Tolić**[1], Fernanda P. do Amaral[5], Charles K. Ansong[1], Heather M. Brewer[1], Chan Lan Chun[4], Yaya Cui[5], Alice Dohnalkova[1], Abigail Ferrieri[1], Richard Ferrieri[2], Karina Freire d'Eça Nogueira Santos[5], Susannah Tringe[3], Angela D. Norbeck[1], Vania C. S. Pankievicz[5], Fabio Pedrosa[6], Michael Sadowsky[4], Lihui Song[5], Emanuel M. de Souza[6]

[1]Evironmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA; [2]Biological, Environmental & Climate Sciences Department, Brookhaven National Laboratory, Upton, NY; [3]Microbial Systems Group, Metagenome Program, DOE Joint Genome Institute, Walnut Creek, CA; [4]Biotechnology Institute, University of Minnesota, St. Paul, MN; [5]Divisions of Plant Science and Biochemistry, C.S. Bond Life Science Center, University of Missouri, Columbia, MO; [6]Department of Biochemistry and Molecular Biology, Federal University of Paraná, 81531-980, Curitiba, Brazil

The plant rhizosphere contains a complex microbial community that directly impacts plant growth, health, and development. This microbial community, located chiefly in the rhizosphere, is largely supported by the release of nutrients from the plant root with carbon and nitrogen sources being among the most important. Climate change factors, such as elevated CO_2 and soil temperature, are

likely to impact the rhizosphere community with perhaps detrimental effects for plant health. In this study, we utilized well characterized *Arabidopsis* mutants defective in carbon allocation to study changes in the rhizosphere microbial community.

Plants were grown hydroponically or in soil and rhizosphere and rhizoplane microbial communities were established from a soil inoculum. Roots were isolated from plants harvested at first light in the morning and just after the lights were turned off in the evening. The rhizosphere and rhizoplane communities were extracted by either gentle washing or sonication. We used a variety of analytical tools to view the interactions of plant roots with the soil microbial population including: i) ^{11}C-labeling to measure carbon allocation from shoot to roots; ii) characterization of rhizosphere populations by metagenomic and 16S rDNA amplicon sequencing; iii) imaging the rhizosphere using scanning electron microscopy (SEM); and iv) quantitative metaproteomic and metametabolomic measurements .

We compared rhizospheric soil of wild-type *A. thaliana* and mutants defective in shoot-root carbon partitioning to directly analyze the role of plant carbon allocation and root exudation in the control of rhizosphere microbial populations. SEM analyses showed that roots were heavily colonized with a well-developed biofilm, composed of a diverse microbial community. Initial results obtained from 16s rRNA sequencing of the rhizosphere microbiome confirmed the high microbial diversity and indicated significant changes in the community structure between wild-type and mutants. For metaproteomics analyses, proteins extracted from the rhizosphere soil were subjected to single-dimension LC-MS analysis. In lieu of a database derived from metagenomic sequencing, we initially searched LC-MS data against a database comprised of all bacteria, archaea, and fungal organisms in TrEMBL. This approach initially identified 170 non-redundant, non-plant proteins (at 1% FDR). In the context of the current state of the art in rhizospheric soil metaproteomics, this represents more than twice the number of non-redundant, non-plant proteins typically identified. Examination of the 170 non-plant proteins revealed that bacteria are the dominant organisms in the *Arabidopsis* rhizosphere microbiome (82%), followed by fungi (15%) and archaea (3%). The metaproteomics-based community structure is consistent with recent studies using metagenomics and furthermore provides direct protein-level evidence for active biological processes within the community. Metametabolomics analyses of the same rhizosphere soil yielded a broad profile of fatty acid methyl esters and polar metabolites. Furthermore, comparative metagenomic, metaproteomics and metametabolomics are ongoing to define changes in community function between mutant and wild type samples.

JGI Plant Gene Atlas: Progress towards complete transcriptome resources in JGI 'Flagship' Plants

Gary Stacey* (staceyg@missouri.edu)[3], **Avinash Sreedasyam**[1,2], Thomas P. Brutnell[7], Thomas E. Juenger[4], Md Shakhawat Hossain[3], Sabeeha S. Merchant[8], Todd Mockler[7], John Mullet[10], Stefan A. Rensing[9], Jeremy Schmutz[1,2], Yuhong Tang[5], Gerald A. Tuskan[6]

[1] DOE Joint Genome Institute, Walnut Creek, CA; [2] HudsonAlpha Institute for Biotechnology, Huntsville, AL; [3] Divisions of Plant Science and Biochemistry, National Center for Soybean Biotechnology, University of Missouri, Columbia, MO; [4] University of Texas at Austin, Department of Integrative Biology, Austin, TX; [5] Plant Biology Division, Samuel Roberts Noble Foundation, Ardmore, OK; [6] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [7] Donald Danforth Plant Science Center, St. Louis, MO; [8] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA; [9] University of Marburg, Karl-von-Frisch-Str. 8, 35043 Marburg, Germany; [10] Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX

The JGI Plant Gene Atlas is a large, coordinated effort to add functional information to JGI Plant Flagship Genomes and develop an updateable, transcriptomic resource that makes a substantial amount of data directly available to JGI users through the JGI Plant Portal at phytozome.jgi.doe.gov. The primary goal of this project is to develop dense RNA-seq data sets for Plant Flagships to form reference transcriptomes across common tissues and conditions and to provide the ability to compare expression from orthologous genes across the JGI Plant genomes. The secondary goal is to provide a technology test bed to further develop genomic techniques that illuminate the function of plant genes and plant regulatory and pathway elements. The initial pilot project includes samples from five major JGI Plant Flagship genomes: Chlamydomonas (algal model), Physcomitrella (moss model), Setaria (grass model), soybean (legume model and crop plant), and poplar (biomass tree crop). This pilot has now been expanded with the addition of standard tissues and conditions for three other JGI Plant Flagships: Sorghum (a C4 grass bioenergy crop and model), Brachypodium (a C3 grass model), and switchgrass (a woody perennial crop plant).

In addition to standard tissues and conditions, the Gene Atlas also includes a comparative condition across the plants focused on nitrogen metabolism. In this case, each plant was grown in sterilized medium with urea, ammonium, or nitrate. Soybean was also grown under nitrogen-fixing conditions. This experiment provides a single condition that can be compared broadly from the minimal Chlamydomonas, into the early plant model Physcomitrella, and through the dicots and monocots included in the study.

We will present ongoing work towards the Gene Atlas, including sample collection, preliminary data analysis, ongoing efforts in cross-species comparisons and data release and availability for the project.

Genome sequencing and comparative analysis of the biocontrol agent *Trichoderma harzianum sensu stricto* TR274

Andrei Steindorff* (assteindorff@lbl.gov)[1,2], Irina S. Druzhinina[3], Igor Grigoriev[1], Sajeet Haridas[1], Christian P. Kubicek[3], Alan Kuo[1], E.F. Noronha[2], Robert W. Riley[1], Asaf Salamov[1], C.J. Ulhoa[2]

[1]Joint Genome Institute – Walnut Creek – CA – USA; [2]Universidade de Brasilia – Brasilia – DF – Brazil; [3]Institute of Chemical Engineering Vienna University of Technology, Vienna, Austria

Biological control is a complex process which requires many mechanisms and a high diversity of biochemical pathways. The species of *Trichoderma harzianum* are well known for their biocontrol activity against many plant pathogens. To gain new insights into the biocontrol mechanism used by *T. harzianum*, we sequenced the isolate TR274 genome using Illumina. The assembly was performed using AllPaths-LG with a maximum coverage of 100x. The assembly resulted in 2282 contigs with a N50 of 37033bp. The genome size generated was 40.8 Mb and the GC content was 47.7%, similar to other Trichoderma genomes. Using the JGI Annotation Pipeline we predicted 13,932 genes with a high transcriptome support. CEGMA tests suggested 100% genome completeness and 97.9% of RNA-SEQ reads mapped to the genome. The phylogenetic comparison using orthologous proteins with all Trichoderma genomes sequenced at JGI, corroborates the Trichoderma (*T. asperellum* and *T. atroviride*), Longibrachiatum (*T. reesei* and *T. longibrachiatum*) and Pachibasium (*T. harzianum* and *T. virens*) section division described previously. The comparison between two *Trichoderma harzianum* species suggests a high genome similarity. Analyses of the secondary metabolites, CAZymes, transporters, proteases and transcription factors were performed. The Pachybasium section expanded virtually all categories analyzed compared with the other sections. CAFE analysis showed positive correlation between these

families and genome size. These results suggests that these proteins families has an important role on their respective phenotypes. Future analysis will improve the understanding of this complex genus and give some insights about its lifestyle and the interactions with the environment.

BIGELOW SINGLE CELL GENOMICS CENTER

Ramunas Stepanauskas* (rstepanauskas@bigelow.org), Wendy K. Bellows, Joe Brown, Elizabeth Fergusson, Jessica M. Labonté, Corianna Mascena, Nicole Poulton, Brian Thompson, Ben Tupper

Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine

Bigelow Single Cell Genomics Center (SCGC), established in 2009, is a nonprofit research and service center. Our mission is to make single cell genomics accessible to the broad research community and to serve as an engine for discoveries in microbial ecology, evolution, bioprospecting and human health. Microbial single cell DNA sequencing, pioneered by SCGC scientists, reads the genomic blueprints of the most fundamental units of life without the need for cultivation. This is a powerful approach to analyze biochemical properties and evolutionary histories of uncultured microorganisms, thought to constitute over 99% of biological diversity on Earth. Single cell genomics also provides unique insights into the microdiversity and evolutionary processes within microbial populations and within multicellular organisms.

SCGC is the first shared user facility of its kind. Since its establishment, SCGC has already developed partnerships and supported research projects at over 100 universities, research institutes, and companies in six continents. Over 1,000,000 individual cells have been processed through our high-throughput pipeline. Sources of these cells include diverse marine environments, soils, deep subsurface, gut content, and others. SCGC activities have provided unique genomic data from many major evolutionary branches of bacteria, archaea, and eukarya that resist cultivation, making a significant impact on our understanding of life on planet Earth. Most recent SCGC's technological advances include scaled-up single cell genomic sequencing and assembly, improved DNA amplification techniques, and matching of single cell optical properties with their genomic sequences.

Elucidating cyanobacterial recycling of microbial mat extracellular matrix

Rhona K. Stuart* (stuart25@llnl.gov)1, Xavier Mayali[1], Jackson Z. Lee[3,4], Craig R. Everroad[4], Mary S. Lipton[2], Amy Boaro[2], Brad M. Bebout[3], Jennifer Pett-Ridge[1], Peter K. Weber[1], Michael Thelen[1].

[1]Lawrence Livermore National Laboratory, Livermore, CA; [2]Pacific Northwest National Laboratory, Richland, WA; [3]Exobiology Branch, NASA Ames Research Center, Moffett Field, CA; [4]Bay Area Environmental Research Institute, Sonoma, CA.

Cyanobacterial carbon excretion is crucial to carbon cycling in many microbial communities, but the nature and bioavailability of the carbon excreted is dependent on physiological function, which is often unknown. Hypersaline laminated photosynthetic mats are a good model system for the study of carbon flow in a complex community because they are sustained primarily by photosynthesis in a relatively closed system.

These mats have a large reservoir of carbon in the extracellular matrix, but how cyanobacterial matrix production is regulated, and what microorganisms consume matrix material, are relatively unexplored questions. To better understand cyanobacterial carbon excretion, we examined the macromolecular composition of the extracellular matrix of microbial mats from Elkhorn Slough in Monterey Bay, California. We characterized the extracellular glycosyl composition and found a large potential reservoir of glucose, bound in 1,4 linked oligosaccharides. Using shotgun proteomics and a Elkhorn Slough metagenomic database (CSP#701), we identified the mat exoproteome. The exoproteome was predominantly composed of cyanobacterial proteins, many with predicted roles in breakdown of organic matter, which suggested cyanobacterial degradation of matrix organic matter. To further explore the regulation of these breakdown abilities, we characterized a biofilm-forming cyanobacterial isolate from Elkhorn Slough, ESFC-1.

Using this culture, we identified exoproteins that changed abundance over a diel cycle, and proteins that were induced by dark stress, suggesting light-dependent regulation of matrix material breakdown. We then used high-resolution imaging mass spectrometry (NanoSIMS) to characterize EPS-carbon and EPS-nitrogen re-uptake in ESFC-1. Our results demonstrated light-dependent, rapid uptake of EPS-associated carbon by ESFC-1 in culture. Based on these findings, we propose that mat *Cyanobacteria* store and recycle organic material from the mat extracellular matrix. *Cyanobacteria* are such a large percentage of the biomass in the upper phototrophic layer of the microbial mats, that their re-uptake of organic carbon has the potential to re-define carbon availability and turnover in these systems.

Funding was provided by the U.S. Department of Energy (DOE) Genomic Science Program under contract SCW1039, Joint Genome Institute CSP support (#701) provided for the metagenomes. and at Pacific Northwest National Lab supported by the OBER/GSP under the Pan-Omics project. Work at LLNL was performed under the auspices of the U.S. Department of Energy at Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

The constitution and functions of microbial communities: profiling metagenomes using contigs and reads

Kristin Tennessen* (ktennessen@lbl.gov), Alicia Clum, Brian Bushnell, Kostas Mavrommatis, Natalia N. Ivanova, Alex Copeland, Nikos C. Kyrpides, Amrita Pati

DOE Joint Genome Institute, 2800 Mitchell Dr, Walnut Creek CA 94598, USA

Advancements in sequencing technologies have enabled better understanding of the composition of microbial communities through sequencing and analysis. Using sequence assembly, contigs constituted from multiple reads can overcome sequencing errors. Ideally, only reads with sequencing errors will remain in the unassembled portion of a metagenome dataset. In practice, assembly can be affected by low coverage, strain variation, and repetitive sequences, resulting in unassembled reads that contain valuable representation of the microbial community. In this work, we explore functional and taxonomic profiling of contigs and reads using published metagenomes and a simulated metagenome. We demonstrate that the diversity and proportion of bases that assemble is directly correlated with the habitat from which a metagenome is sequenced. We find that unassembled reads profiles can differ from contig profiles, but contig profiles and all-reads profiles are quite similar. Thus, even though profiling is slightly less accurate in the unassembled portion of a metagenome, assembly may be omitted if deep analysis is not part of project scope.

Genomes OnLine Database (GOLD) v.5: an improved metadata management system based on a four level (meta)genome project classification

Alexander D. Thomas* (alexanderthomas@lbl.gov), T.B.K. Reddy, Michelle Isbandi, Jyothi Mallajosyula, Dimitrios Stamatis, Jonathon Bertsch, Nikos C. Kyrpides

DOE Joint Genome Institute, Walnut Creek, CA, USA

The Genomes Online Database (GOLD) is a global catalog of genome and metagenome projects. The latest version of GOLD (v.5) features a redesigned schema and new user interface. The improved schema more accurately represents the growing size and complexity of sequencing projects. The new user interface facilitates the submission of more diverse sequencing projects as well as complex analysis projects and a more seamless interface with the Integrated Microbial Genomes (IMG) family of analysis tools. GOLD is also actively curated to increase the quantity and improve the quality of sequencing project metadata. GOLD fully supports and promotes the Genomic Standards Consortium (GSC) Minimum Information standards. There are many challenges with integrating information from disparate and diverse sources. However, the collection, curation, and dissemination of sequencing project metadata is necessary for the analysis of genetic information and hypothesis development.

Structural and gene content variation among strains of the maize anthracnose fungus *Colletotrichum graminicola*.

Michael R. Thon* (mthon@usal.es), **Gabriel E. Rech**, Ramón Pacheco, Serenella A. Sukno

Instituto Hispano-Luso de Investigaciones Agrarias (CIALE), Department of Microbiology and Genetics, University of Salamanca, 37185 Villamayor, Spain

To better understand patterns of genetic variation in the maize pathogen *Colletotrichum graminicola*, we sequenced the genomes of seven field isolates showing a variable range of virulence to maize and collected from different regions of the world. We analyzed genomic structural variations and patterns of gene gain and loss using genomic sequences obtained from various assembly and gene annotation strategies. We identified sets of unique genes in each isolate, and discovered that they are significantly enriched with genes coding for small secreted proteins (putative effectors), which could represent evolutionary innovations directly involved in host specificity or environment adaptation. Pathogenicity and microscopic assays show that one of the isolates grows endophytically within the host. Four genes coding for enzymes directly involved in the degradation of plant cell walls are affected by genomic structural variations occurring only in the genome of the endophytic isolate, suggesting that the disruption of some of these genes could be responsible for the loss of pathogenicity in this isolate. In addition, five of the 72 genes are upregulated in the reference isolate M1.001 during the necrotrophic phase of infection suggesting their involvement in the pathogenic lifestyle of *C. graminicola*. Genes coding for putative effectors were found at or near breakpoints of genomic structural variations, suggesting this mechanism could be involved in promoting variability of these genes. Overall, genomic variation and patterns of gene gain/loss provide a valuable resource for selecting targets for further functional and population genetic analyses aimed at identifying genes involved in the development of maize anthracnose.

Methanogen communities across a salinity gradient in San Francisco Bay wetlands

Ye Tian* (YeTian@lbl.gov)[1], Susanna Theroux[1], Wyatt H. Hartman[1], Susannah Tringe[1]

[1]DOE Joint Genome Institute, Walnut Creek, California

Though covering only 6~9% of the earth's surface area, wetlands store around 35% of the terrestrial carbon on a global scale. Wetland restoration has the potential to remove a large amount of carbon dioxide (CO_2) from the atmosphere because the rate of aboveground primary production can outcompete the rate of biomass decomposition in anoxic soils of wetlands. However, wetland anoxic soil environments also promote the growth of methanogens, a group within the Archaea that are the only organisms known to produce methane (CH_4); as a result, wetlands are responsible for nearly 40~50% of CH_4 emitted to the atmosphere annually. To better understand the composition and activity of methanogens in restored wetlands, we sampled sixteen sites capturing a range of salinities (0-250 ppt) and restoration statuses (historic, restored, unrestored) throughout the San Francisco Bay-Delta region. Using 16S rRNA gene sequencing and shotgun metagenomics coupled with greenhouse gas monitoring, we compared phylogenetic distribution and metabolic potential of methanogens across all sampling sites. We built phylogenetic trees with 16S rRNA gene data as well as mcrA gene data (coding for methyl coenzyme M reductase alpha subunit, involved in the terminal stage of methane production). Though traditional dogma dictates that sulfate-reducing bacteria outcompete methanogens when sulfate is available, we observed methane production in regions with ample sulfate availability, suggesting the presence of methanogen species that use alternative substrates (such as trimethylamine and methanol). Our metagenome data confirmed the presence of alternative methanogenesis pathways, and distinct abundances of methanogens across salinity, organic carbon, and restoration age gradients. Our study facilitates a more comprehensive understanding of the link between belowground methanogen communities to aboveground methane production, providing insights into wetland restoration management.

Microbial species delineation using whole genome sequences

Neha J. Varghese* (njvarghese@lbl.gov)[1], Supratim Mukherjee[1], Natalia N. Ivanova[1], Konstantinos T. Konstantinidis[2], Kostas Mavrommatis[3], Nikos C. Kyrpides[1] and Amrita Pati[1]

[1] Microbial and Metagenome Superprogram, DOE Joint Genomic Institute, Walnut Creek, California, 94598, United States; [2] Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332-0355, United States; [3] Celgene Corp., San Franscisco, California, 94158, United States

Faster and cheaper sequencing of microbial genomes has revealed that prokaryotic species assignments using traditional approaches can be inconsistent with information emerging from whole genomes for a significant number of species. The long-standing need for a more systematic and scalable species assignment technique can be met by the genome-wide Average Nucleotide Identity (gANI) metric, which is widely acknowledged as a robust measure of genomic relatedness. In this work, we demonstrate that the combination of gANI and the alignment fraction (AF) between two genomes accurately reflects their genomic relatedness. We introduce an efficient implementation of AF,gANI and discuss its successful application to 86.5M genome pairs between 13,151 prokaryotic genomes assigned to 3,032 species. Subsequently, by comparing the genome clusters obtained from complete linkage clustering of these pairs to existing taxonomy, we observed that nearly 18% of all prokaryotic species suffer from anomalies in species definition. Our results can be used to explore central questions such as whether

microorganisms form a continuum of genetic diversity or distinct species represented by distinct genetic signatures. We propose that this precise and objective gANI,AF-based species definition be used to address previous inconsistencies in species classification and as the primary guide for new taxonomic species assignment, supplemented by the traditional polyphasic approach, as required.

Tired of Qiime? Use SEED - a GUI based sequence editor and pipeline for high-throughput amplicon processing

Tomáš Větrovský* (vetrovsky@biomed.cas.cz), Petr Baldrian

Institute of Microbiology of the ASCR, v.v.i., Vídeňská 1083, 14220 Praha 4, Czech Republic

There are many well established pipelines for amplicon data analyses, but most of them are command line based and thus harder to comprehend and handle by biologists without background in bioinformatics. The SEED (http://www.biomed.cas.cz/mbu/lbwrf/seed/) is a free to use GUI-based sequence editor and pipeline with internal functions and functions performed by external programs that are installed for full functionality (1). SEED was created to provide an intuitive interface for fast bioinformatic analysis of PCR amplicons on desktop computers according to the suggested workflow. While the first version was limited by 32-bit architecture, the most recent versions are 64-bit and allow comfortable work with up to 8 million sequences (~4 GB of data) on a standard personal computer with 8 GB RAM. SEED is especially designed for sequential analysis of the sequences of PCR amplicons obtained by Illumina or 454-pyrosequencing including sequences of fungal ITS regions, bacterial 16S rDNA or other target genes. The program has a wide array of functions including editing of sequences and their titles, sorting, quality trimming, pair-end joining, grouping of sequences based on sequence motifs or sequence titles, batch processing of sequence groups, denoising, chimera removal, ITS extraction, sequence alignments and clustering, OTU table construction, construction of consensus sequences, creation of local databases for BLAST and searching either them or the whole NCBI, retrieval of taxonomical classification from the NCBI, calculation of diversity parameters and many more.

(1) Větrovský, T. and P. Baldrian (2013). "Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED." Biology and Fertility of Soils 49: 1027-1037.

Using the model perennial grass *Brachypodium sylvaticum* to engineer resistance to multiple abiotic stresses

John P. Vogel* (jpvogel@lbl.gov)[1], **Sean P. Gordon**[1], Maria Reguera[2], Nir Sade[2], Amy Cartwright[1], Christian Tobias[2], Roger Thilmony[2], Eduardo Blumwald[2]

[1]DOE Joint Genome Institute, Walnut Creek, CA 94598; [2]Dept. of Plant Sciences, University of California, Davis, CA 95616; [3]USDA-ARS Western Regional Research Center, Albany CA94710

We are using the perennial model grass *Brachypodium sylvaticum* to identify combinations of transgenes that enhance tolerance to multiple, simultaneous abiotic stresses. The most successful transgene combinations will ultimately be used to create improved switchgrass (*Panicum virgatum* L.) cultivars. To further develop *B. sylvaticum* as a perennial model grass, and facilitate our planned transcriptional profiling, we are sequencing and annotating the genome. We have generated ~40x genome coverage using PacBio sequencing technology of the largest possible size selected libraries (18,

22, 25kb). Our initial assembly using only long-read sequence contained 320Mb of sequence with an N50 contig length of 315kb and an N95 contig length of 40kb. This assembly consists of 2,430 contigs, the largest of which was 1.6Mb. The estimated genome size based on c-values is 340Mb indicating that about 20Mb of presumably repetitive DNA remains yet unassembled. Significantly, this assembly is far superior to an assembly created from paired-end short-read sequence, ~100x genome coverage. The short-read-only assembly contained only 226Mb of sequence in 19k contigs. To aid the assembly of the scaffolds into chromosome-scale assemblies we produced an F_2 mapping population and have genotyped 480 individuals using a genotype by sequence approach.

One of the reasons for using *B. sylvaticum* as a model system is to determine if the transgenes adversely affect perenniality and winter hardiness. Toward this goal, we examined the freezing tolerance of wild type *B. sylvaticum* lines to determine the optimal conditions for testing the freezing tolerance of the transgenics. A survey of seven accessions noted significant natural variation in freezing tolerance. Seedling or adult Ain-1 plants, the line used for transformation, survived an 8 hour challenge down to -6 deg C and 50% survived a challenge down to -9 deg C. Thus, we will be able to easily determine if the transgenes compromise freezing tolerance.

In the effort to develop biotechnological tools for perennial grass improvement, we have completed the transformation of *B. sylvaticum* with constructs containing 20 genes shown to be associated with enhanced abiotic stress tolerance in monocots. In addition, we have transformed plants with constructs containing a combination of genes (i.e. SARK::IPT- Ubi::HSR1::Ubi::NHX1) in order to simultaneously overexpress genes associated with drought + heat tolerance + salt tolerance. We generated single copy insert T_1 lines for all constructs and the generation and bulking of homozygous T_2 lines is well underway. In addition to our *B. sylvaticum* transgenics, we transformed *B. distachyon* with many of the same genes. Some of the transgenic *B. distachyon* plants subjected to a combined stress of both drought and salinity were able to produced higher yields than wild type plants. Our results indicate a great potential for the development of grasses with improved performance and yield in water-limited areas.

Broadly Applicable Method for Specific Transcript Depletion from RNA-Seq Libraries without Impacting Representation of Non-targeted Transcripts

I-Ching Wang* (icwang@nugen.com), Doug Amorese, Marie Eide, Lin Pham

NuGEN Technologies, Inc (201 Industrial Rd., San Carlos, CA 94070)

When preparing RNA-Seq libraries, many transcriptomes are dominated by rRNA and other transcripts that are not particularly informative. During sequencing, these uninformative transcripts take up a high percentage of the reads in a run resulting in higher cost to obtain the coverage needed for informative reads.

There are various methods to reduce the representation of rRNA from RNA-Seq libraries, each with its own draw-backs. The simplest method using Oligo dT priming suffers from 3' bias and fails to represent non-polyadenylated transcripts, which can shed light on gene regulation and serve as biomarkers for disease conditions. Another approach using targeted oligonucleotides annealing to specific transcripts within RNA populations followed by RNase H digestion can result in partial hydrolysis of targeted transcripts. However, the result is also a large quantity of both RNA and DNA fragments which must be efficiently removed to prevent them from being included in subsequent libraries. Finally, hybridization based capture methods employing biotinylated oligos can be effective in removing unwanted

transcripts, but require additional steps and have been observed to affect non-targeted transcripts resulting in distorted libraries.

Here we describe a novel method, Insert Dependent Adapter Cleavage (InDA-C), for effective removal of specified transcripts with minimal impact on non-targeted transcripts. InDA-C employs specific and robust enzymatic steps to greatly reduce the representation of any undesired transcripts during the construction of stranded RNA-Seq libraries. The InDA-C workflow is fast, easy, and readily customized to address unwanted transcripts of almost any kind. Using the InDA-C method, we report the removal of rRNA, globin RNA, and chloroplast RNA from RNA-Seq libraries across a wide variety of sample types and organisms (mammals, insects, plants, and prokaryotes). We have also applied this method to libraries generated from RNAs removed by capture based methods to reveal the removal of unintended, daisy-chained transcripts.

A Functional Encyclopedia for Evidence-Based Annotation of Diverse Bacteria

Kelly M. Wetmore* (kmwetmore@lbl.gov)[1], M. N. Price[1], R. J. Waters[2], J. Lamson[1], M. Callaghan[1], Y. Suh[1], J. Kuehl[1], B. Rubin[3], J. Ray[1], Cindy A. Hoover[2], Matthew J. Blow[2], J. Bristow[2], Adam P. Arkin[1], Adam. M. Deutschbauer[1]

[1] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley CA; [2] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek CA; [3]Division of Biological Sciences, University of California, San Diego

There is a great discrepancy between our ability to sequence microbial genomes and to characterize the genes encoded by these genomes. To meet this challenge, we developed a flexible microbial functional annotation pipeline consisting of high-throughput culturing and transposon mutagenesis. Here, we describe the application of this pipeline for the evidence-based annotation of gene function in diverse bacteria including those of relevance to DOE missions of bioenergy, carbon sequestration, and bioremediation. In particular, to enable the inference of gene function on a global scale in diverse bacteria, we previously developed a method, random barcode TnSeq (RB-TnSeq), to increase the throughput of mutant fitness profiling by incorporating random DNA barcodes into transposons and monitoring mutant fitness using barcode sequencing (BarSeq). Here, we present RB-TnSeq results demonstrating the reproducibility and biological meaningfulness of the method using 17 diverse bacteria from 11 different genera and 5 different divisions. We identified phenotypes for 18,453 of the 56,518 protein-coding genes that we tested, including phenotypes for 5,859 poorly annotated proteins. We identified strong cofitness or conserved phenotypes – which often lead to functional annotations – for 3,316 poorly annotated proteins and for representatives of 309 domains of unknown function (PFam, DUFs and UPFs). Across all sequenced bacterial genomes, our data provides phenotypes for potential orthologs of 14% of poorly annotated proteins and informative phenotypes for 8% of them.

Functional genomics of moss-cyanobacteria interactions in boreal forest ecosystems

Philip D. Weyman* (pweyman@jcvi.org)[3], **Denis Warshan[1]**, Jeroen T. F. Gillard[2], Christopher L. Dupont[2], Ulla Rasmussen[1]

[1]Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden; [2]Microbial and Environmental Genomics Group, J. Craig Venter Institute, La Jolla, CA.; [3]Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, CA.

Boreal forests are important ecosystems involved in global carbon cycling. The nitrogen cycling in these ecosystems is largely determined by a symbiotic association between feather mosses and diazotrophic cyanobacteria that fix majority of nitrogen flowing into boreal ecosystems. Recent secondary ion mass spectrometry (SIMS) data shows accumulation of cyanobacterial fixed nitrogen in mosses colonized by *Nostoc* cyanobacteria, establishing the flow of nitrogen from symbiont to host. Because nitrogen is often limiting in boreal forests, the interaction between the cyanobacteria and the mosses greatly affects the productivity of this ecosystem that makes up almost 30% of Earth's forested land. We seek to understand the genetic diversity of the cyanobacteria associated with the mosses, the molecular steps leading to the moss-cyanobacterial association, and the molecular currencies of the symbioses.

We present current progress in a Joint JGI/EMSL Community Science Project combining genomics, transcriptomics, and proteomics to characterize the feather moss-cyanobacteria association that covers the understory in boreal forests worldwide. Recent data from phylogenetic analysis of moss-associated cyanobacteria and from moss-cyanobacteria reconstitution experiments has informed our two central hypotheses: I) the moss-associated cyanobacteria are very diverse and different moss species host unique cyanobacterial communities, and II) the moss and cyanobacteria exchange and sense diffusible signals to establish and maintain a successful nitrogen-fixing association. Our approach is to test these hypotheses targets all levels of the central dogma and will culminate in genetic testing of hypotheses using targeted gene knockout in the symbiont of key genes derived from the results.

We are currently sequencing the genomes of five different *Nostoc* spp. that are able to form symbiotic associations with two feather moss species(*Pleurozium schreberi* and *Hylocomium splendens*). Additionally, several non-cultivable *Stigonema* spp. that make up a significant fraction of the nitrogen-fixing cyanobacteria found on feather mosses will also be sequenced after micromanipulation and multiple displacement amplification (MDA). As a control, we will sequence the genome of one *Nostoc* spp. that is unable to form symbioses with the mosses. Comparative genome analysis of these cyanobacterial species will allow us to probe the genomic diversity of feather moss-associated *Nostoc* strains and identify major genetic differences and similarities that constitute the symbiotic association with mosses. Hypotheses about gene function generated by these 'omics approaches, will be tested through a reverse genetics approach, using JGI's DNA synthesis capacity to assemble gene knockout cassettes of candidate cyanobacterial genes essential for successful moss-cyanobacteria associations. Knock-out mutants resulting from this proposal will thus be assayed for their capacity to form nitrogen-fixing associations.

Currently, we are also obtaining transcriptomic and proteomic data for both the feather moss and *Nostoc* partners grown in isolation, together, or with chemical contact only through filter separation. The generated genome sequences will provide the sequence database necessary for these studies: comparing the transcriptomic and proteomic responses of both partners during symbiotic competent and non-competent interactions will allow us to identify potential mechanisms used to establish and maintain the symbiosis and design additional targets for gene knockout and testing of symbiotic association.

Comparative Genomics of Powdery Mildews and Associated Plants

Mary C. Wildermuth* (mwildermuth@berkeley.edu), William Underwood, Jyoti Taneja, Kate Scheibel, Andrew Liu, and Shauna C. Somerville

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

Powdery mildews are widespread obligate biotrophic fungi that infect agricultural species, ornamentals, bioenergy plants (e.g., poplar, eucalyptus, camelina, grasses, soybean, sugar beet) and the model plants tomato and *Arabidopsis thaliana*. Collectively, powdery mildews are the most significant cause of plant disease worldwide. More than 900 species of powdery mildew have been described and grouped into five Tribes within the Family Erysiphaceae. Despite their importance, only one powdery mildew genome, the barley powdery mildew *Blumeria graminis* f.sp. *hordei* (*Bgh*), has been fully sequenced and assembled. Our recently selected Community Science Program Project will transform community knowledge and understanding of the powdery mildews through sequencing of genomes representing all five Tribes within the Erysiphaceae, including three species that infect designated bioenergy crops (i.e., camelina, eucalyptus, poplar). These powdery mildews have adapted to distinct plant hosts and all but one of the nine host plants used in this project have fully sequenced genomes. Transcriptome data for this diverse set of powdery mildews and their associated plant hosts at three infection stages (germination, establishment, proliferation) will also be acquired to elucidate the complex interactions that lead to successful infection.

Powdery mildews likely share features with beneficial fungi in that they must be highly adapted to their hosts, minimize host damage and defense activation, and manipulate hosts into providing nutrients. As obligate biotrophs that only grow and reproduce on living plant tissue, their metabolism is highly integrated and dependent on that of the host plant. Comparative genomics of the diverse powdery mildews on their associated hosts will allow for identification of core components necessary for the (obligate) biotrophic lifestyle as well as specific factors that act as determinants of powdery mildew host range. For example, comparisons across powdery mildew species' cell wall degrading enzymes (CWDE) are expected to identify a conserved core set of CWDEs, as well as novel plant species-specific CWDEs and co-factors. These latter findings could facilitate bioenergy crop-specific targeted approaches to biomass degradation. Additional powdery mildew genes and traits underlying host specialization (e.g., effector diversity, ability to metabolize host secondary metabolites) are also anticipated. The generated resources will serve a large powdery mildew community and accelerate diverse studies including outcomes that enhance the resistance of economically important crops to powdery mildew infection, limit fungicide use, allow more sustainable agricultural practices, and provide new insights into bioenergy crop growth, metabolism, disease protection and degradation. In addition, the project data resources placed in evolutionary context can inform our understanding of the evolution of obligate biotrophy and allow comparison with parasitic biotrophic microbes (e.g. rust and smut fungi, Oomycetes) and mutualistic biotrophs such as arbuscular mycorrhizal fungi to identify shared and divergent strategies of plant biotrophs.

Phylogenetic Profiling of bacterial and archaeal Pfam Families at Different Taxonomic levels

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

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

Phylogenetic profiling is a non-homology based gene functional prediction method. By studying the joint presence or absence of gene families in different taxa, phylogenetic profile analysis has been used to infer meaningful biological connections between gene families such as shared biological pathways and protein complexes. Previous phylogenetic profiling studies have been complicated by sample biases such as the phylogenetic relationships of available sequenced genomes. We and others have previously attempted to adapt methods such as independent contrast to address phylogenetic issues in phylogenetic profiling with limited success. Here we report a new approach: phylogenetic profiling of bacterial and archaeal families at different taxonomic levels. Phylogenetic tree based Operational Taxonomic Units (OTUs) were built using the TreeOTU method for different taxonomic levels. Ancestral states of gene family presence/absence were calculated for each OTU using Felsenstein's algorithm for 'Phylogenetic Independent Contrasts' (Felsenstein, 1985). Phylogentic profiling was performed for the ancestral states of all the OTUs at the same taxonomic level. Here we report phylogenetic profiling of 9432 Pfam families across 7850 at different taxonomic levels (strain, species, genus, family, class, order and phylum), as well as clusters of gene families that share presence/absence patterns across all the taxonomic levels.

Grass to Gas: Discovering the Transcriptional Regulatory Networks in *Neurospora crassa* Involved in Plant Cell Wall Degradation

Vincent Wu* (vwu104@berkeley.edu)[1], David Kowbel[1], Anna Lipzen[2], Vasanth Singan[2], Igor Grigoriev[1,2] and N. Louise Glass[1]

[1]Department of Plant and Microbial Biology UC Berkeley, [2]Department of Energy Joint Genome Institute

Conversion of cellulosic plant biomass to biofuels holds great potential for alleviating our reliance to fossil fuels. A major goal of this research is to develop fungi to be suitable for secreting large volumes of plant cell wall degrading enzymes (CWDEs). In order to productively direct our efforts, a solid understanding of how these fungi regulate the expression of CWDEs must first be established. Using *Neurospora crassa* as a model, we have begun to characterize how filamentous fungi respond to different carbon sources. By comparing transcriptional profiles of *N. crassa* grown on a variety of carbon sources, we have determined that the fungus responds uniquely to different polysaccharides. Furthermore, we have concluded that *N. crassa* uses simple sugar molecules as a way to sense what is in its environment. Unique CWDEs are transcriptionally upregulated in the presence of low concentrations of these simple sugar dimers and monomers. To move forward we are investigating predicted transcription factors that are upregulated in concert with different sets to CWDEs in order to determine if they are involved in regulating transcription of these CWDEs. Finding these transcription factors is key to understanding the regulatory pathways of CWDE expression and will establish the grounds from which we can conduct directed engineering of filamentous fungi for protein production.

MaxBin 2.0: Integrating multiple metagenomic datasets to recover more high quality individual genomes

Yu-Wei Wu* (ywwei@lbl.gov)[1,2], Blake Simmons[1,3], and Steven Singer[1,4]

[1]Joint BioEnergy Institute, Emeryville, CA 94608, USA; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [3]Biological and Materials Science Center, Sandia National Laboratories, Livermore, CA, USA; [4]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

Automated recovery of individual genomes from metagenomes remains to a challenging problem; however it is a necessary step to understand the phylogeny and ,metabolic potential of uncultivated microbial populations in natural environments. We have developed MaxBin, an automated metagenomic binning software to classify genomes from metagenomes in a purely automatic way. Based on an Expectation-Maximization algorithm, MaxBin estimates the number of genomes and groups assembled contigs or scaffolds into corresponding genomic bins. By using MaxBin, researchers can gather information from the metagenomes and study the microbial populations more easily. In order to reconstruct genomes more comprehensively and accuratley, we developed MaxBin 2.0, which utilized information from multiple metagenomic datasets to achieve better performance. MaxBin 2.0 was tested on simulated datasets and achieved the highest precision and comparable sensitivity among existing automated metagenomic binning tools such as CONCOCT, MetaBAT, and GroopM. By testing MaxBin 2.0 on real metagenomes we found that we are able to recover more genomes as compared to the original MaxBin. MaxBin 2.0 is available at:
http://downloads.jbei.org/data/microbial_communities/MaxBin/MaxBin.html or the sourceforge website http://sourceforge.net/projects/maxbin2.

Whole genome sequencing of aquatic fungi responsible for the degradation of recalcitrant substrates in aqueous environments

Christian Wurzbacher* (christian@wurzbacher.cc) **[1, 2],** Michael T. Monaghan[1, 2], Shiri Freilich[3], Andrey Yurkov[4], Dietmar Schlosser[5]

[1]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany (igb-berlin.de); [2]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany (begendiv.de); [3]Agricultural Research Organization (ARO), Newe-Ya'ar, Israel (agri.gov.il); [4]Leibniz-Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany (dsmz.de); [5]Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany (ufz.de)

Aquatic true fungi perform several important functions in natural ecosystems. Similarly to terrestrial fungi, the predominant role of aquatic fungi is the decomposition of organic carbon. Despite being key players in organic matter decomposition and detoxification in many aquatic habitats, their taxonomic diversity and metabolic potential remain poorly known. Polymeric organic carbon in freshwaters is both of terrestrial and aquatic origin, whereby quality and quantity derived from terrestrial plants can substantially differ from that of aquatic macrophytes and algae. The metabolic versatility of aquatic fungi enables them to transform a wide range of natural and anthropogenic material in aquatic food webs at a very fast pace, which provides a crucial role in carbon and nutrient cycling linked to ecosystem health on a global scale. In this CSP we propose whole-genome sequencing of 25 isolates of aquatic fungi from a representative range of ecosystem types including streams, lakes, wetlands, marine and brackish waters). They encompass a broad evolutionary spectrum (Ascomycetes, Basidiomycetes, Zygomycetes and Chytridiomycetes) that comprises a large variety of metabolic processes. This functional breadth is designed to assess both conservative and innovative enzyme sets acquired by fungi during their evolution to adapt to life in the aqueous phase. Further RNASeq coupled to functional assays will decipher major metabolic pathways. These enzyme sets may hold key features for efficient biodegradation, carbon mineralization and detoxification processes.

The functional potential of *Arabidopsis thaliana* root endophyte communities

Scott Yourstone* (scott.yourstone81@gmail.com)[1], Derek S. Lundberg[2], Sur Herrera Paredes[1], Hunter Cameron[1], Scott Clingenpeel[3], Tanja Woyke[3], Susannah Tringe[3], Corbin D. Jones[1], and **Jeffery L. Dangl**[1,4]

[1] University of North Carolina, Chapel Hill, NC; [2] Max Planck Institute for Developmental Biology, Tübingen, Germany; [3] Joint Genome Institute, Walnut Creek, CA; [4] Jeffrey Dangl is an HHMI-GBMF Plant Science Investigator

Bacterial communities living inside of plant roots (endophytes) provide beneficial phenotypes to host plants such as disease resistance, increased growth, or resistance to abiotic stress. How the functional potential of these endophytic communities shapes traits remains unclear. We are using whole metagenome sequencing to determine the functional potential of theses communities. To circumvent the challenge of host plant DNA contamination in the metagenomic data, we are combining a variety of approaches to separate microbes from host. These include 1) whole genome sequencing of cultured clonal bacterial isolates, 2) *en masse* sequencing of all culturable material via metagenomic DNA made from "plate scrapes", and 3) metagenomic DNA made from pools of bacteria separated from host material via cell sorting. Here we describe these sequencing efforts and compare functional profiles from endophyte communities sampled from roots of two *Arabidopsis thaliana* genotypes grown in two distinct soils.

Mini-Metagenomic Approaches Using FACs and Microfluidics

Feiqiao (Brian) Yu* (brianyu@stanford.edu)[1], Danielle Goudeau[2], Rex Malmstrom[2], Tanja Woyke[2], Mark A. Horowitz[1], Stephen R. Quake[1]

[1] Stanford University, Stanford, CA, USA; [2] Joint Genome Institute, Walnut Creek, CA, USA

Metagenomics deal with the analysis of genetic materials recovered from the environment. The ultimate goal is to reconstruct entire genomes of unknown microbial species. Two popular methods currently exist for metagenomic studies. Traditional shotgun approach sequence DNA extracted from large sample volumes containing hundreds of species. Even though experimental process is simple, determining which sequencing reads originated from the same cell is a complex if not impossible task. Another approach isolates single cells into microfluidic chambers or individual wells using FACs (Fluorescent Activated Cell Sorting) or an optical tweezers. This method guarantees that all sequencing reads originate from the same cell but suffers from low throughput, making it difficult to detect important but rare species.

We present a new metagenomic concept, termed mini-metagenomics, which combines the advantages of both shotgun and single cell methods. Instead of single microbial cells, we isolate many small populations of around 10 cells. Each isolate represents a mini-metagenomic sample. Lysis is performed on each isolate followed by MDA (Multiple Displacement Amplification) using Qiagen's Repli-g Single Cell Kit. Amplified genomic DNA is sequenced and assembled using standard bioinformatics pipelines. Theoretically, our mini-metagenomic method offers higher throughput than single cell methods and lower bioinformatics complexity compared to shotgun methods.

In terms of implementation, we carry out mini-metagenomic sample preparation on a FACs-based pipeline at JGI and a microfluidic-based platform at Stanford University. Both methods use an artificially mixed bacterial sample containing 5 different species whose genomes and relative abundances are known. For the FACs-based approach, 10 cells are sorted into each well. Lysis and MDA are performed in the same well. For the microfluidic approach, we modify a Fluidigm© C1 microfluidic platform. Cell mixtures of the desired concentration (yielding ~10 cells per reaction) is flown into the microfluidic chip and randomly dispersed into 96 chambers. Lysis and MDA are performed on-chip. Amplified materials are harvested into a 96 well plate. Nextera libraries are prepared from MDA products of both approaches together to avoid any library preparation related biases. Finally, DNA libraries are sequenced on the Illumina Nextseq platform.

We mapped sequencing reads to reference genomes and observed that most reads uniquely align to one of the five species with < 2% chimeras and unaligned reads. Compared to single cell sequencing, results from FACs and microfluidic mini-metagenomic approaches display improved genome coverage, suggesting that, for our mini-metagenomic method, less sequencing cost is needed to obtain the same genomic assembly performance. The FACs-based method has more precise control over cell number in each well but requires expensive machinery and elaborate setups to automate the sample prep process. On the other hand, the microfluidic-based sample preparation is easier and cheaper to perform but suffers from the Poisson distribution of per chamber cell numbers. Finally, using redundant information from multiple samples we can significantly improve assembly results while maintaining single cell resolution. Thus, the proposed mini-metagenomic approach offers significant benefits over existing metagenomic sample prep method and can benefit the exploration of unknown environmental microbiome.

Identifying metabolic processes underlying patterns of rhizosphere microbial succession and organic matter priming

Kateryna Zhalnina* (kzhalnina@lbl.gov)[1,2], Heejung Cho[1], Ulas Karaoz[2], Stefan Jenkins[2], Richard A. White III[3], Mary S. Lipton[3], Jizhong Zhou[4], Jennifer Pett-Ridge[5], Trent R. Northen[2], Mary K. Firestone[1,2], and **Eoin L. Brodie**[1,2]

[1]University of California, Berkeley, California; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Pacific Northwest National Laboratory, Richland, Washington; [4]University of Oklahoma, Norman, Oklahoma; [5]Lawrence Livermore National Laboratory, Livermore, California

Interactions between plants, soil microorganisms and soil minerals underlie both the stabilization and decomposition of soil organic carbon (SOC). Plants increase the concentration of labile C in soil through exudation, providing readily assimilated substrates for rhizosphere microbiota and stimulation or repression of SOC mineralization may result (i.e. positive or negative priming). The soil microbiome displays strong patterns of succession in response to living plant roots and although both positive and negative rhizosphere priming have been widely observed, the metabolic processes underlying these successional patterns and SOC priming are not well defined.

Using combined modeling and experimentation we are working to define key traits of soil microorganisms relevant to their fitness in the rhizosphere and transformation of carbon. From a library of approximately 300 bacterial isolates we sequenced 38 heterotrophic bacteria representative of the dominant organisms identified in metagenomes of this Mediterranean grassland soil. Analysis of genome properties and content allowed us to predict important physiological features of the isolates, such as minimum generation times (MGTs), optimum temperature and substrate utilization capabilities.

MGTs inferred directly from genomic sequences agreed well with *in vitro* observations. All analyzed genomes revealed a repertoire of features relevant to life in the rhizosphere, such as high gene copy number of sugar and organic acid transporters and polymer degrading enzymes (glycoside hydrolases and polyphenoloxidases). Isolates from the Alphaproteobacteria and Actinobacteria had the highest occurrence of these features in their genomes. To test genomic predictions we have performed a number of functional analyses using enzymatic, exometabolomic and proteomic assays.

We have developed and curated draft genome-scale metabolic models of select isolates and have used exometabolomic data to perform gap filling. These metabolic models will be used to evaluate via simulation the metabolic response of soil heterotrophic bacteria to root exudates and their impact on rhizosphere succession and SOC mineralization. Explicit features of the genome-scale models (electron donors and acceptors, C use efficiency, metabolic feedbacks and trade-offs) together with experimental data (gene expression, protein secretion, metabolite uptake and release) will be used in future work to parameterize genome-informed trait-based models of the soil microbial community aimed at predicting succession and C transformation in the rhizosphere.

Comparative genomic study towards two *Aspergillus oryzae* strains

Yiyi Zhong* (yiyizhong@cuhk.edu.hk), Wenyan Nong, Hoi Shan Kwan

Food and Nutritional Science, School of Life Science, The Chinese University of Hong Kong, Hong Kong

With a great capacity of enzyme production for macronutrient digestion, *Aspergillus oryzae* is important in food processing industry, especially soy sauce fermentation. Chinese commercial strain *A. oryzae* RD2 can produce soy sauce with high quality. The strain TS2 was isolated after repetitive use of RD2, which has limited enzyme production and produces poor-quality soy sauce. We aimed to conduct comparative genomic study to discover the reason behind the phenotypic differences. We first sequenced the genomes of RD2 and TS2 with Ion-torrent sequencing technology. By using the conservative genes including beta tubulin (BT2), calmodulin (CF) and RNA polymerase II (RPB2), we built up the phylogenetic trees of RD2 and TS2 with other 19 *Aspergillus* species. From the phylograms, the evolutionary distance between RD2 and TS2 is shortest. And the Japanese working strain RIB40 and another Chinese working strain 3.042 are also closely related to both strains. After that, we mapped RD2 and TS2 sequences to RIB40 genome separately. We focused on single nucleotide polymorphism (SNP) and deletion on the genes encoding amylases, proteases, nucleotidases and lipases. The result reveals TS2 has unique SNPs and deletion in the coding sequence of alpha-amylase gene and amylo-alpha-glucosidase gene. As for protease, genes of TS2 encoding neutral protease and proteases related to cysteine, glutamate and aspartic acid metabolism are found with mutation at multiple positions of nucleotides. In addition, TS2 also have unique SNPs and deletions on genes encoding IMP-specific nucleotidase, biphosphate nucleotidase and phospholipase. Though no evidence shows that one deletion causes the total loss of a single gene, these unique genetic variations are likely to cause the disruption of gene function like macronutrients digestion and flavor formation. Based on the comparison, we can attribute the phenotypic differences especially enzyme production between RD2 and TS2 to genetic variation. Further, we infer that the unique SNPs and deletions on TS2 genome cause the limited enzymes secreted and poor quality of soy sauce. More significantly, this study can build a genomic foundation for transcriptomic and proteomic study on these two strains. In this way, we can rationalize the pathways of metabolism involving hydrolytic digestion which can help to deepen our understanding towards the industrially important *Aspergillus* species.

Engineering N_2-fixing Cyanobacteria to Synthesize Perfumery Linalool from Air and Water

Ruanbao Zhou* (Ruanbao.zhou@sdstate.edu), Liping Gu, Charles Halfmann, Xianling Xiang, and William Gibbons

Department of Biology and Microbiology, South Dakota State University, Brookings, SD 57007

Current bio-ethanol plants release one third of carbon as CO_2 during fermentation, as well as significant amounts of low grade heat. Ideally, a photosynthetic organism could be engineered to directly convert these unused resources into high value chemicals. Linalool ($C_{10}H_{18}O$) is a naturally-occurring terpene alcohol, emitted as a volatile compound from many flowers and over 200 species of plants. In higher plants, linalool is produced from a universal isoprenoid intermediate geranyl diphosphate (GPP) through the MEP pathway. Linalool has many commercial applications in perfumed hygiene products, flavor and fragrances, pharmaceuticals, and drop-in biofuels. However, high production costs and raw material shortages have restrained the linalool production industry. Engineering cyanobacteria to photosynthetically produce linalool may have potential to simultaneously solve these two problems.

Cyanobacteria, like plants, have native metabolic pathways (Calvin cycle and MEP pathway) to photosynthetically convert CO_2 and water into a variety of reduced carbon compounds including GPP, the precursor for linalool. However, cyanobacteria lack the linalool synthase that plants use to convert GPP into linalool. In this research, the linalool synthase gene from *Norway Spruce* was fused to a synthetic, dual cyanobacterial P_{nir}–P_{psbA1} promoter and subcloned into a shuttle vector for transformation of an N_2-fixing cyanobacterium *Anabaena sp.* PCC7120*.* The first generation of transgenic *Anabaena* (with a single plant linalool synthase gene) was confirmed by GC-MS to continuously synthesize and secrete linalool via photosynthesis. To increase linalool production, a synthetic operon coding for three key enzymes (DXS-IDI-GPPS) from the MEP pathway were over-expressed in *Anabaena*. This second generation strain produced 2-3 fold more linalool than the first generation strain. Under 50 $\mu E \cdot m^{-2} \cdot s^{-1}$ light conditions, the maximum linalool productivity was ~1.7μg/L/h when used nitrate as sole nitrogen source, and ~1.3μg/L/h when used atmospheric N_2 gas as sole nitrogen source.
In summary, our engineered cyanobacteria can directly convert CO_2 and water into linalool, which is only driven by solar energy. Our data revealed that linalool produced by the engineered cyanobacterium was secreted into media and volatilized into the flask headspace, allowing for easy separation of linalool from the culture biomass. Thus, the engineered N_2-fixing cyanobacterium can serve as a living cellular factory to continuously produce and emit a wide range of commodity chemicals/fuels using only atmospheric gases (CO_2 & N_2), mineralized H_2O, and sunlig

Attendees

Timothy Alba
UNLV
albat@unlv.nevada.edu

William Alexander
University of Wisconsin-Madison/Great Lakes Bioenergy Research Center
walexander@wisc.edu

Oleg Alexandrov
DOE Joint Genome Institute
oalexandrov@lbl.gov

Ravi Alla
UC Berkeley/QB3
ravi.alla@berkeley.edu

Eric Alsop
DOE Joint Genome Institute
ebalsop@lbl.gov

Mitchell Altschuler
EBI
mitchellalt@hotmail.com

Karthik Anantharaman
UC Berkeley
karthik.a@berkeley.edu

Mikael Anderson
Technical University of Denmark
jesmith@lbl.gov

Leif Andersson
Uppsala University
leif.andersson@imbim.uu.se

Mauricio Antunes
Colorado State University/ Dept. of Biology
Mauricio.Antunes@colostate.edu

Adam Arkin
LBNL
aparkin@lbl.gov

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

Petr Baldrian
Institute of Microbiology of the ASCR, v. v. i.
baldrian@biomed.cas.cz

Massie Ballon
DOE Joint Genome Institute
mlballon@lbl.gov

Mario Banuelos
University of California, Merced
mbanuelos4@ucmerced.edu

Jason Barker
University of British Columbia
barkerjas@gmail.com

José María Barrasa
Alcalá University
josem.barrasa@uah.es

Kerrie Barry
DOE Joint Genome Institute
kwbarry@lbl.gov

Laura Bartley
University of Oklahoma
lbartley@ou.edu

Michael Barton
DOE Joint Genome Institute
mbarton@lbl.gov

Jacob Beam
Montana State University
jprbeam@gmail.com

David Beck
University of Washington
dacb@uw.edu

Joan W Bennett
Rutgers University
profmycogirl@yahoo.com

Randy Berka
Novozymes
jesmith@lbl.gov

Mark Blenner
Clemson University
blenner@clemson.edu

Matthew Blow
DOE Joint Genome Institute
mjblow@lbl.gov

Harvey Bolton
EMSL/PNNL
harvey.bolton@pnnl.gov

Gregory Bonito
Michigan State University
bonito@msu.edu

Alessandra Borgognone
Genetics and Microbiology Research Group
alessandra.borgognone@unavarra.es

Robert Bowers
DOE Joint Genome Institute/LBL
rmbowers@lbl.gov

Jennifer Bragg
JBEI
jnbragg@lbl.gov

James Brayer
Oxford Nanopore Technologies Ltd
leah.partridge@nanoporetech.com

Steven Briggs
UC San Diego
sbriggs@ucsd.edu

Jim Bristow
DOE Joint Genome Institute
Jbristow@lbl.gov

Luciano Brocchieri
University of Florida/MGM
lucianob@ufl.edu

Eoin Brodie
Lawrence Berkeley National Lab
elbrodie@lbl.gov

Joe Brown
Bigelow Laboratory
jmbrown@bigelow.org

Natasha Brown
DOE Joint Genome Institute
nzvenigorodsky@lbl.gov

Chris Brown
UC Berkeley
ctb@berkeley.edu

David Bruce
Los Alamos National Laboratory
dbruce@lanl.gov

Thomas Brutnell
Donald Danforth Plant Science Center
tbrutnell@danforthcenter.org

Yury Bukhman
GLBRC, University of Wisconsin-Madison
ybukhman@glbrc.wisc.edu

Mark Burk
Genomatica
mburk@genomatica.com

Jim Burrows
Pacific Biosciences
jburrows@pacificbiosciences.com

David Burstein
UC Berkeley
davidbur@berkeley.edu

Posy Busby
Duke University
busby@post.harvard.edu

Brian Bushnell
DOE Joint Genome Institute
bbushnell@lbl.gov

Atul Butte
Stanford University
abutte@stanford.edu

Cristina Butterfield
UC Berkeley
butterfc@berkeley.edu

Andrew Cal
USDA-ARS
andrew.cal@ars.usda.gov

Barbara Campbell
Clemson University
bcampb7@clemson.edu

Manu Capoor
Rockefeller University
mcapoor@rockefeller.edu

Joseph Carlson
DOE Joint Genome Institute
JWCarlson@lbl.gov

Amy Cartwright
DOE Joint Genome Institute
acartwright@lbl.gov

Raúl Castanera
Genetics and Microbiology Research Group
raul.castanera@unavarra.es

David Catcheside
Flinders University
David.Catcheside@flinders.edu.au

David Cavalier
Great Lakes Bioenergy Center
cavalie8@msu.edu

Rick Cavicchioli
UNSW
r.cavicchioli@unsw.edu.au

Javier A. Ceja-Navarro
LBNL
jcnavarro@lbl.gov

Meghan Chafee
Max Planck Institute for Marine Microbiology
mchafee@mpi-bremen.de

Nadia Chaidir
DOE Joint Genome Institute
nichaidir@lbl.gov

Patrick Chain
LANL
pchain@lanl.gov

Michelle Chang
UCB
mcchang@berkley.edu

Jinhui Chang
The Chinese University of Hong Kong
tjchangjinhui@gmail.com

Jui-Jen Chang
Department of Medical Research,
China Medical University Hospital, Taiwan
lancecjj@gmail.com

Cindy Chen
DOE Joint Genome Institute
cindychen@lbl.gov

Zhongqiang (John) Chen
DUPONT
zhongqiang.chen@dupont.com

Jay Chen
Oak Ridge National Laboratory
chenj@ornl.gov

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

Xuanjin Cheng
The Chinese University of Hong Kong
jeanch12345@gmail.com

Jennifer Chiniquy
Joint Genome Institute
JLChiniquy@lbl.gov

Ludmila Chistoserdova
University of Washington
milachis@uw.edu

Mansi Chovatia
DOE Joint Genome Institute
mrchovatia@lbl.gov

Beat Christen
ETH Zürich
beat.christen@imsb.biol.ethz.ch

Matthias Christen
ETH Zürich
matthias.christen@imsb.biol.ethz.ch

Tyson Clark
Pacific Biosciences
tclark@pacificbiosciences.com

Scott Clingenpeel
DOE Joint Genome Institute
srclingenpeel@lbl.gov

Alicia Clum
DOE Joint Genome Institute
aclum@lbl.gov

Cameron Coates
DOE Joint Genome Institute
rccoates@lbl.gov

Benjamin Cole
DOE Joint Genome Institute
bjcole@lbl.gov

Jessica Cole
Pacific Northwest National Laboratory
jessica.cole@pnnl.gov

John Cole
University of Illinois Urbana-Champaign
cole15@illinois.edu

Devin Coleman-Derr
USDA-ARS
devin.coleman-derr@ars.usda.gov

Robert Cottingham
Oak Ridge National Lab
cottinghamrw@ornl.gov

Kelly Craven
The Noble Foundation
kdcraven@noble.org

Cameron Currie
UW-Madison
currie@bact.wisc.edu

Chris Daum
DOE Joint Genome Institute
cgdaum@lbl.gov

Jason Davis
University of California, Merced
jdavis8@ucmerced.edu

Ronald De Vries
CBS-KNAW Fungal Biodiversity Center
jesmith@lbl.gov

Anne Dekas
Lawrence Livermore National Laboratory
dekas1@llnl.gov

Ed DeLong
Univeristy of Hawaii
edelong@hawaii.edu

Shweta Deshpande
DOE Joint Genome Institute
sdeshpande@lbl.gov

Sam Deutsch
DOE Joint Genome Institute
sdeutsch@lbl.gov

Adam Deutschbauer
Lawrence Berkeley National Laboratory
AMDeutschbauer@lbl.gov

Anthony Devine
FMC Corporation
Anthony.Devine@fmc.com

Fred Dietrich
Duke University
fred.dietrich@duke.edu

Rhyan Dockter
DOE Joint Genome Institute
rbdockter@lbl.gov

Sharon Doty
University of Washington
sldoty@uw.edu

Devin Doud
DOE Joint Genome Institute
dfdoud@lbl.gov

Meghan Drake
Oak Ridge National Laboratory
drakemm@ornl.gov

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

Irina Druzhinina
Vienna University of Technology
irina.druzhinina@tuwien.ac.at

Inna Dubchak
DOE Joint Genome Institute
ildubchak@lbl.gov

Kecia Duffy
DOE Joint Genome Institute
kmduffy@lbl.gov

Sebastien Duplessis
INRA
duplessi@nancy.inra.fr

Rob Egan
DOE Joint Genome Institute
rsegan@lbl.gov

Emiley Eloe-Fadrosh
DOE Joint Genome Institute
eaeloefadrosh@lbl.gov

Aymerick Eudes
LBNL / JBEI
ageudes@lbl.gov

Cameron Everson
San Francisco State University
ceverson@mail.sfsu.edu

Jan Fedorko
Global Change Research Centre, Academy of Sciences of the Czech Republic, v. v. i.
j.fedorko.usire@gmail.com

Michael Fero
TeselaGen Biotechnology
mike.fero@teselagen.com

Bryce Foster
DOE Joint Genome Institute
bryce.foster@lbl.gov

Brian Fox
UW-Madison/GLBRC
bgfox@biochem.wisc.edu

Christopher Francis
Stanford University
caf@stanford.edu

Roxanne Fries
DOE Joint Genome Institute
rfries@lbl.gov

Masanori Fujimoto
University of Michigan
masafuji@umich.edu

Philip Gach
Joint BioEnergy Institute
pcgach@lbl.gov

George Garrity
Michigan State University
garrity@msu.edu

Audrey Gasch
University of Wisconsin, Madison
jesmith@lbl.gov

Romina Gazis
Clark University
rgazis@clarku.edu

Gina Geiselman
UC Berkeley
geiselman.g@berkeley.edu

Lida Gifford
Lawrence Berkeley National Laboratory
lkgifford@lbl.gov

David Gilbert
DOE Joint Genome Institute
degilbert@lbl.gov

Jack Gilbert
Argonne National Laboratory
gilbertjack@gmail.com

N.Louise Glass
UC Berkeley
jesmith@lbl.gov

Tijana Glavina del Rio
DOE Joint Genome Institute
tglavina@lbl.gov

Cinta Gomez
UC Berkeley
cgomezsilvan@lbl.gov

Tania Gonzalez
UC Berkeley
tlgonzalez@berkeley.edu

David Goodstein
DOE Joint Genome Institute
dmgoodstein@lbl.gov

Stephen Goodwin
Purdue University/USDA
jesmith@lbl.gov

Sean Gordon
DOE Joint Genome Institute
seangordon07@gmail.com

Danielle Goudeau
DOE Joint Genome Institute
dgoudeau@lbl.gov

Danielle Graham
DOE Joint Genome Institute
dgraham@lbl.gov

Bastian Greshake
Goethe University Frankfurt am Main
bgreshake@googlemail.com

Igor Grigoriev
DOE Joint Genome Institute
ivgrigoriev@lbl.gov

Michalis Hadjithomas
DOE Joint Genome Institute
michalis@lbl.gov

Aria Hahn
University of British Columbia
hahnaria@gmail.com

Richard Hall
Pacific Biosciences
rhall@pacb.com

Steven Hallam
University of British Columbia
shallam@mail.ubc.ca

Peter Hallin
Novozymes A/S
pfha@novozymes.com

Richard Hamelin
The University of British Columbia/Natural Resources Canada
richamelin@gmail.com

Joshua Hamilton
University of Wisconsin-Madison
jjhamilton2@wisc.edu

Nancy Hammon
DOE Joint Genome Institute
nmhammon@lbl.gov

Sajeet Haridas
DOE Joint Genome Institute
sharidas@lbl.gov

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

Wyatt Hartman
DOE Joint Genome Institute
whhartman@lbl.gov

Emily Hatas
Pacific Biosciences
ehatas@pacb.com

Richard Hayes
DOE Joint Genome Institute
rdhayes@lbl.gov

Brian Hedlund
University of Nevada, Las Vegas
brian.hedlund@unlv.edu

Monica Heger
GenomeWeb
mheger@genomeweb.com

Jan Hennessey
Berkeley Lab
jkhennessey@lbl.gov

Carlos Hernandez-Garcia
JBEI / LBNL
cmhernandezgarcia@lbl.gov

Sur Herrera Paredes
UNC-CH
sur00mx@gmail.com

Matthias Hess
University of California Davis
mhess@ucdavis.edu

David Hibbett
Clark University
jesmith@lbl.gov

Nathan Hillson
DOE Joint Genome Institute
njhillson@lbl.gov

Chris Hittinger
University of Wisconsin - Madison
cthittinger@wisc.edu

Beth Hofstad
Pacific Northwest National Lab
beth.hofstad@pnnl.gov

Cindi Hoover
DOE Joint Genome Institute
cahoover@lbl.gov

Laura Hug
UC Berkeley
laura.hug@berkeley.edu

Phil Hugenholtz
University of Queensland
p.hugenholtz@uq.edu.au

Hope Hundley
DOE Joint Genome Institute
HNTice@lbl.gov

Marcel Huntemann
DOE Joint Genome Institute
mhuntemann@lbl.gov

Sam Ian
faculty law
info.kangsam@yahoo.com

Kateryna Ievdokymenko
UBC
ievdokymenko@gmail.com

Bill Inskeep
Montana State University
binskeep@montana.edu

Michelle Isbandi
DOE Joint Genome Institute
misbandi@lbl.gov

Dan Jacobson
Oak Ridge National Laboratory
jacobsonda@ornl.gov

Mei-yeh Jade
Biodiversity Research Center, Academia Sinica
meiyehlu@gate.sinica.edu.tw

Timothy James
U. Michigan
tyjames@umich.edu

Jessica Jarett
DOE Joint Genome Institute
jkjarett@lbl.gov

Barbara Jasny
Science/AAAS
bjasny@aaas.org

Zackary Jay
Montana State University
zackary.jay@montana.edu

Jerry Jenkins
HudsonAlpha Institute
jjenkins@hudsonalpha.org

Yongqin Jiao
LLNL
jiao1@llnl.gov

Beibei Jing
The Joint BioEnergy Institute (JBEI)
Bjing@lbl.gov

Nicole Johnson
1981
nicolevjohnson@lbl.gov

Sissel Juul
Oxford Nanopore Technologies, Inc
leah.partridge@nanoporetech.com

Shawn Kaeppler
University of Wisconsin-Madison; GLBRC
smkaeppl@wisc.edu

Ananth Kalyanaraman
Washington State University
ananth@eecs.wsu.edu

Janine Kamke
AgResearch
janine.kamke@agresearch.co.nz

Sophien Kamoun
The Sainsbury Lab
Sophien.Kamoun@sainsbury-laboratory.ac.uk

Don Kang
DOE Joint Genome Institute
ddkang@lbl.gov

Lennart Kappelmann
Max Planck Society
lkappelm@mpi-bremen.de

ULAS KARAOZ
LBNL
ukaraoz@lbl.gov

Anne-Kristin Kaster
Leibniz Institute DSMZ
a.kaster@dsmz.de

Andrew Katz
Affiliate Technician
katz@ucdavis.edu

Megan Kennedy
DOE Joint Genome Institute
mckennedy@lbl.gov

Mahsa Khorasani
University of Washington
mkh2612@uw.edu

Soohong Kim
Broad Institute of MIT and Harvard
soohong@broadinstitute.org

Kristi Kim
Pacific Biosciences
kspittle@pacb.com

Jeffrey Kimbrel
Joint BioEnergy Institute
jakimbrel@lbl.gov

Nobutada Kimura
National Institute of Advanced Industrial Science and Technology
n-kimura@aist.go.jp

Jonathan Klassen
University of Connecticut
jonathan.klassen@uconn.edu

Jason Klein
University of Washington
jcklein@uw.edu

Manuel Kleiner
University of Calgary
manuel.kleiner@ucalgary.ca

Hans-Peter Klenk
Newcastle University
hans-peter.klenk@online.de

Annegret Kohler
INRA Nancy
kohler@nancy.inra.fr

Eneas Konzen
University of Sao Paulo
eneas_florestal@yahoo.com.br

David Koppenaal
Pacific Northwest National Laboratory
david.koppenaal@pnnl.gov

Joel Kostka
Georgia Institute of Technology
joel.kostka@biology.gatech.edu

Joseph Koussa
NYUAD
jk150@nyu.edu

Max Krall
USDA-ARS
max.krall1@gmail.com

Karen Krüger
Max Planck Institute for Marine Microbiology
kkrueger@mpi-bremen.de

Dave Kudrna
University of Arizona
dkudrna@email.arizona.edu

Ursula Kues
University of Göttingen
ukuees@gwdg.de

Elizabeth Kujawinski
WHOI
ekujawinski@whoi.edu

Sunita Kumari
Cold Spring Harbor Labs
kumari@cshl.edu

Victor Kunin
Taxon Biosciences
kunivi@gmail.com

Rita Kuo
DOE Joint Genome Institute
rckuo@lbl.gov

Alan Kuo
DOE Joint Genome Institute
akuo@lbl.gov

Hoi Shan Kwan
The Chinese University of Hong Kong
hoishankwan@cuhk.edu.hk

Nikos Kyrpides
DOE Joint Genome Institute
NCKyrpides@lbl.gov

Jessy Labbe
ORNL
Labbejj@ornl.gov

Piyush Labhsetwar
University of Illinois Urbana-Champaign
labhset1@illinois.edu

David Lai
DOE Joint Genome Institute
dlai@lbl.gov

Kathleen Lail
DOE Joint Genome Institute
klail@lbl.gov

Regina Lamendella
Juniata college
lamendella@juniata.edu

Olga Lastovetsky
Cornell University
ol57@cornell.edu

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

Cynthia Lawley
Illumina
clawley@illumina.com

Sarah Lebeis
University of Tennessee
slebeis@utk.edu

Marc-Henri LEBRUN
Marc-Henri
marc-henri.lebrun@versailles.inra.fr

Ji Lee
DOE Joint Genome Institute
jel@lbl.gov

Janey Lee
DOE Joint Genome Institute
jlee2@lbl.gov

Scott Lee
DOE Joint Genome Institute
sjlee@lbl.gov

Ethan Levesque
Aloha Medicinals Inc
ethanlevesque@gmail.com

Asaf Levy
DOE Joint Genome Institute
alevy@lbl.gov

Ivan Liachko
University of Washington
il34@uw.edu

Shoudan Liang
DOE Joint Genome Institute
sliang@lbl.gov

Junyan Lin
DOE Joint Genome Institute
junyanlin@lbl.gov

Stephen Lindemann
Pacific Northwest National Laboratory
stephen.lindemann@pnnl.gov

Tomas Linder
Swedish University of Agricultural Sciences
tomas.linder@slu.se

Erika Lindquist
DOE Joint Genome Institute
ealindquist@lbl.gov

Alexandra Linz
University of Wisconsin Madison
amlinz@wisc.edu

Xiaoyu Liu
Thermo Fisher Scientific
Xiaoyu.liu@thermofisher.com

Po-Yu Liu
DOE Joint Genome Institute
pliu@lbl.gov

Brian Looney
University of Tennessee
blooney@vols.utk.edu

Joyce Loper
USDA-ARS and Oregon State University
loperj@science.oregonstate.edu

Taina Lundell
University of Helsinki
taina.lundell@helsinki.fi

Susan Lynch
University of California San Francisco
susan.lynch@ucsf.edu

Rex Malmstrom
DOE Joint Genome Institute
rrmalmstrom@lbl.gov

Nasim Mansoori
Lawrence Berkeley Lab-JBEI
nmansoorizangir@lbl.gov

Xinwei Mao
UC Berkeley
xmao1@berkeley.edu

Francis Martin
INRA
fmartin@nancy.inra.fr

Angel T. Martínez
CIB-CSIC
atmartinez@cib.csic.es

Corianna Mascena
Bigelow Laboratory for Ocean Sciences
cmascena@bigelow.org

Rebekah Mathews
DOE Joint Genome Institute
rpmathews@lbl.gov

Tyler McCubbin
Afingen
tmccubbin@lbl.gov

Ran Mei
University of Illinois
mrzhsy@gmail.com

Xiandong Meng
DOE Joint Genome Institute
xiandongmeng@lbl.gov

Laura Meredith
Stanford University
predawn@stanford.edu

Therese Mitros
DOE Joint Genome Institute
tkmitros@lbl.gov

Stephen Mondo
DOE Joint Genome Institute
sjmondo@lbl.gov

Scott Monsma
Lucigen Corp.
smonsma@lucigen.com

Abraham Morales-Cruz
UC Davis
abmora@ucdavis.edu

Hiroshi Mori
Tokyo Institute of Technology
hmori@bio.titech.ac.jp

Gary Muehlbauer
University of Minnesota
muehl003@umn.edu

Amanda Muehlbauer
University of California Berkeley
muehlbauer@berkeley.edu

Supratim Mukherjee
DOE Joint Genome Institute
supratimmukherjee@lbl.gov

Aindrila Mukhopadhyay
Lawrence Berkeley National Laboratory
amukhopadhyay@lbl.gov

Avaro Munoz Plominsky
University of Concepcion
alvaro.munoz@imo-chile.cl

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

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

Christine Naca
DOE Joint Genome Institute
clnaca@lbl.gov

Sarah Nalven
Oregon State University
sarahnalven@gmail.com

Sangeeta Nath
DOE Joint Genome Institute
snath@lbl.gov

Ali Navid
LLNL
navid1@llnl.gov

Stephen Nayfach
UCSF
snayfach@gmail.com

Simona Necula
DOE Joint Genome Institute
sfnecula@lbl.gov

William Nelson
Pacific Northwest National Laboratory
william.nelson@pnnl.gov

David Nelson
New York University Abu Dhabi
drn2@nyu.edu

Vivian Ng
DOE Joint Genome Institute
vng@lbl.gov

Chew Yee Ngan
DOE Joint Genome Institute
cyngan@lbl.gov

Torben Nielsen
DOE Joint Genome Institute
torben@lbl.gov

Masaru Nobu
University of Illinois
masarunobu@gmail.com

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

Trent Northen
DOE Joint Genome Institute
trnorthen@lbl.gov

Dan Northrup
Booz Allen Hamilton
northrup_daniel@bah.com

Jeanette Norton
Utah State University
jeanette.norton@usu.edu

Erin Nuccio
Lawrence Livermore National Laboratory
nuccio1@llnl.gov

Ernst Oberortner
LBNL - DOE Joint Genome Institute
EOberortner@lbl.gov

Seungdae Oh
University of Illinois at Urbana-Champaign
seungdae@illinois.edu

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

Ai Oikawa
AFINGEN, Inc.
aoikawa@lbl.gov

Jeffrey P. Osborne
Manchester University
jposborne@manchester.edu

Robert Otillar
DOE Joint Genome Institute
rpotillar@lbl.gov

David Paez
DOE Joint Genome Institute
adpaezespino@lbl.gov

Marton Palatinszky
Univ. of Vienna
palatinszky@microbial-ecology.net

Chongle Pan
Oak Ridge National Laboratory
panc@ornl.gov

Dale Pelletier
Oak Ridge National Laboratory
pelletierda@ornl.gov

Paul Peluso
Pacific Biosciences, Inc.
ppeluso@pacificbiosciences.com

Ze Peng
DOE Joint Genome Institute
zpeng@lbl.gov

Len Pennacchio
DOE Joint Genome Institute
LAPennachio@lbl.gov

Christa Pennacchio
DOE Joint Genome Institute
cppennacchio@lbl.gov

Rene Perrier
DOE Joint Genome Institute
raperrier@lbl.gov

Joseph Peterson
University of Illinois Urbana-Champaign
jrptrsn3@illinois.edu

Jennifer Pett-Ridge
LLNL
pettridge2@llnl.gov

Jeremy Phillips
DOE Joint Genome Institute
jlphillips@lbl.gov

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

Bharath Prithiviraj
The Noble Foundation
bprithiviraj@noble.org

Alexander Probst
University of California, Berkeley
probst@berkeley.edu

Zhenzhen Qiao
University of Oklahoma
zhenzhen.qiao-1@ou.edu

Weigang Qiu
City University of New York, Hunter College
weigang@genectr.hunter.cuny.edu

Pablo Rabinowicz
U.S. Department of Energy
pablo.rabinowicz@science.doe.gov

David Rank
Pacific Biosciences
drank@pacificbiosciences.com

Sonja Raub
Monash University
scraub@gmail.com

Prasun Ray
The Noble Foundation
pray@noble.org

Gustavo Henrique Recchia
CENA-USP
gust_henr@yahoo.com.br

Tatiparthi Reddy
DOE Joint Genome Institute
tbreddy@lbl.gov

Sue Rhee
Carnegie Institution for Science
srhee@carnegiescience.edu

Robert Riley
DOE Joint Genome Institute
rwriley@lbl.gov

Errol Robinson
EMSL\ PNNL
errol.robinson@pnnl.gov

David Rodriguez
ThermoFisher Scientific
david.rodriguez2@thermofisher.com

Antonis Rokas
Vanderbilt University
antonis.rokas@vanderbilt.edu

Dan Rokhsar
DOE Joint Genome Institute
DSRokhsar@lbl.gov

Adriana Romero-Olivares
UC Irvine
alromer1@uci.edu

Margaret Romine
Pacific Northwest National Laboratory
margie.romine@pnnl.gov

Eddy Rubin
DOE Joint Genome Institute
emrubin@lbl.gov

Anne Ruffing
Sandia National Laboratories
aruffin@sandia.gov

Francisco Javier Ruiz-Dueñas
CIB, CSIC
fjruiz@cib.csic.es

Elizabeth Ryan
DOE Joint Genome Institute
eryan@lbl.gov

Asaf Salamov
DOE Joint Genome Institute
aasalamov@lbl.gov

Kourosh Salehi-Ashtiani
New York University Abu Dhabi
ksa3@nyu.edu

Troy Sandberg
UCSD
tsandberg@ucsd.edu

Hila Sberro
Weizmann Institute of Science
hsberro@gmail.com

Wendy Schackwitz
DOE Joint Genome Institute
wsschackwitz@lbl.gov

Jeremy Schmutz
DOE Joint Genome Institute
jschmutz@hudsonalpha.org

Duncan SCOTT
DOE Join Genome Institute
dnscott@lbl.gov

Volkan Sevim
DOE Joint Genome Institute
vsevim@lbl.gov

Ashley Shade
Michigan State University
shadeash@msu.edu

Atif Shahab
DOE Joint Genome Institute
ashahab@lbl.gov

Nicole Shapiro
DOE Joint Genome Institute
nrshapiro@lbl.gov

Aditi Sharma
DOE Joint Genome Institute / LBNL
aditisharma@lbl.gov

Itai Sharon
UC Berkeley
itaish@berkeley.edu

Helen Shearman
Biomatters
helen@biomatters.com

Shengqiang Shu
DOE Joint Genome Institute
sqshu@lbl.gov

Christine Shulse
DOE Joint Genome Institute
cnshulse@lbl.gov

Justin Siegel
UC Davis
jbsiegel@ucdavis.edu

Blake Simmons
Joint BioEnergy Institute
basimmons@lbl.gov

Holly Simon
Oregon Health & Science University
simonh@ohsu.edu

Vasanth Singan
DOE Joint Genome Institute
VRSingan@lbl.gov

Esther Singer
DOE Joint Genome Institute
esinger@lbl.gov

Jeffrey Skerker
University of California, Berkeley
skerker@berkeley.edu

Steve Slater
Midwestern BioAg
steve.slater@midwesternbioag.com

Tatyana Smirnova
DOE Joint Genome Institute
tsmirnova@lbl.gov

Jacquelyn Smith
LBNL
jesmith@lbl.gov

Jason Smith
MBARI
jsmith@mbari.org

Kim Smuga-Otto
UC Santa Cruz
kimsmuga@gmail.com

Shauna Somerville
University of California at Berkeley
ssomerville@berkeley.edu

Rotem Sorek
Weizmann Institute of Science
rotem.sorek@weizmann.ac.il

Joseph Spatafora
Oregon State University
jesmith@lbl.gov

Avinash Sreedasyam
HudsonAlpha Institute for Biotechnology
asreedasyam@hudsonalpha.org

Gary Stacey
University of Missouri
staceyg@missouri.edu

Jillian Steele
DOE Joint Genome Institute
jksteele@lbl.gov

Andrei Steindorff
DOE Joint Genome Institute
assteindorff@lbl.gov

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

Roman Stocker
MIT
romans@mit.edu

Rhona Stuart
LLNL
stuart25@llnl.gov

Sarah Su
USDA-ARS
ssu4@mail.ccsf.edu

Seung-Jin Sul
Lawrence Berkeley National Lab
ssul@lbl.gov

Anandkumar Surendrarao
UC-Davis
aksrao@ucdavis.edu

Diane Joyce Sutton
state of california
djsutton@mac.com

Taichi Takasuka
Hokkaido University
tendo.nd@gmail.com

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

Angela Tarver
DOE Joint Genome Institute
amtarver@lbl.gov

Neslihan Taş
LBNL
ntas@lbl.gov

Tootie Tatum
DOE Joint Genome Institute
oltatum@lbl.giv

Kristin Tennessen
DOE Joint Genome Institute
ktennessen@lbl.gov

Susanna Theroux
DOE Joint Genome Institute
stheroux@lbl.gov

Alexander Thomas
DOE Joint Genome Institute
alexanderthomas@lbl.gov

Michael Thon
University of Salamanca
mthon@usal.es

Ye Tian
DOE Joint Genome Institute
YeTian@lbl.gov

Christian Tobias
USDA ARS
christian.tobias@ars.usda.gov

Susannah Tringe
DOE Joint Genome Institute
sgtringe@lbl.gov

Andrew Tritt
DOE Joint Genome Institute
ajtritt@lbl.gov

Adrian Tsang
Concordia University
jesmith@lbl.gov

Tim Tschaplinski
Oak Ridge National Laboratory
tschaplinstj@ornl.gov

Elizabeth Tseng
Pacific Biosciences
etseng@pacificbiosciences.com

Ray Turner
DOE Joint Genome Institute
crturner@lbl.gov

Jerry Tuskan
ORNL
tuskanga@ornl.gov

Daniel Van der Lelie
FMC Corporation
daniel.vanderlelie@fmc.com

Neha Varghese
DOE Joint Genome Institute
njvarghese@lbl.gov

Tammi Vesth
Technical University of Denmark
tcve@bio.dtu.dk

Tomas Vetrovsky
Institute of Microbiology of the ASCR, v.v.i.
kostelecke.uzeniny@seznam.cz

Axel Visel
DOE Joint Genome Institute
avisel@lbl.gov

John Vogel
DOE Joint Genome Institute
jpvogel@lbl.gov

Carey Waage
DOE Joint Genome Institute
cvwaage@lbl.gov

Soichi Wakatsuki
SLAC/Stanford University
soichi.wakatsuki@stanford.edu

I-Ching Wang
NuGEN
icwang@nugen.com

Gaoyan Wang
DOE Joint Genome Institute
gaoyanwang@lbl.gov

Zhong Wang
DOE Joint Genome Institute
zhongwang@lbl.gov

Clay Wang
USC
clayw@usc.edu

Mei Wang
DEO Joint Genome Institute/LBL
mwang@lbl.gov

Chia-Lin Wei
DOE Joint Genome Institute
Cwei@lbl.gov

Kelly Wetmore
Lawrence Berkeley National Laboratory
kmwetmore@lbl.gov

Philip Weyman
J. Craig Venter Institute
pweyman@jcvi.org

Mary Wildermuth
University of California, Berkeley
mwildermuth@berkeley.edu

Timothy Williams
DOE Joint Genome Institute
tcwilliams@lbl.gov

Alex Worden
MBARI
azworden@mbari.org

Tanja Woyke
DOE Joint Genome Institute
twoyke@lbl.gov

Stephen Wright
University of Toronto
stephen.wright@utoronto.ca

Thomas Greg Wright
USDA-ARS
tgwright.uh@gmail.com

Dongying Wu
University of California Davis
dywu@lbl.gov

Vincent Wu
UC Berkeley
vwu104@berkeley.edu

G. Albert Wu
DOE Joint Genome Institute
gwu@lbl.gov

Yu-Wei Wu
Joint BioEnergy Institute/Lawrence Berkeley National Lab
ywwei@lbl.gov

Christian Wurzbacher
Leibniz-IGB
cwurzbacher@igb-berlin.de

Denise Yadon
DOE Joint Genome Institute
dyadon@lbl.gov

Juying Yan
DOE Joint Genome Institute
juyingyan@lbl.gov

Mi Yan
DOE Joint Genome Institute
miyan@lbl.gov

Shijie Yao
DOE Joint Genome Institute
syao@lbl.gov

Suzan Yilmaz
Sandia National Labs
syilmaz@sandia.gov

Yasuo Yoshikuni
DOE Joint Genome Institute
yyoshikuni@lbl.gov

Scott Yourstone
UNC-Chapel Hill
scott.yourstone81@gmail.com

Feiqiao Yu
Stanford University
brianyu@stanford.edu

Matthew Zane
DOE Joint Genome Institute
mczane@lbl.gov

Kamrun Zargar
Joint Bioenergy Institiute
kzargar@lbl.gov

Kateryna Zhalnina
Lawrence Berkeley National Lab
kzhalnina@lbl.gov

Jinxiang Zhang
Joint BioEnergy Institute/Sandia National lab
jinxiangzhang@lbl.gov

Yu Zhang
UC Berkeley
yuzhang@berkeley.edu

Jean Zhao
DOE Joint Genome Institute
zyzhao@lbl.gov

Yiyi Zhong
The Chinese University of Hong Kong
yiyi11zhong@gmail.com

Jizhong Zhou
University of Oklahoma
jzhou@ou.edu

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

Naupaka Zimmerman
University of Arizona
naupaka@email.arizona.edu

Author Index

15-JGI-1479

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science