4. Plant-Microbe Interfaces: PMI Data Consolidation: Estimating Data Quality in Large Datasets

Dave Ussery[1,2*] (usserydw@ornl.gov), Miriam Land,[1] Doug Hyatt,[2] Guruprasad H. Kora,[1,4] Se-Ran Jun,[1,5] Loren Hauser,[1,2,3] **Gerald A. Tuskan,[1]** and **Mitchel J. Doktycz[1]**

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge TN 37831; [2]Genome, Science and Technology Graduate Program, University of Tennessee, Knoxville TN 37996; [3]Computing & Computational Sciences, Oak Ridge National Laboratory, Oak Ridge TN 37831; [4]Department of Microbiology, University of Tennessee, Knoxville TN 37996; [5]Joint Institute for Computational Sciences, University of Tennessee, Knoxville, TN 37831

http://PMI.ornl.gov

Project Goals: The goal of the PMI SFA is to understand the genome-dependent molecular and cellular events involved in establishing and maintaining beneficial interactions between plants and microbes. *Populus* and its associated microbial community serves as the experimental system for understanding how these molecular events manifest themselves within the spatially, structurally, and temporally complex scales of natural systems. To achieve this goal, we focus on 1) characterizing host and environmental drivers for diversity and function in the *Populus* microbiome, 2) utilizing microbial model system studies to elucidate *Populus*-microbial interactions at the molecular level and dissecting the signals and pathways responsible for initiating and maintaining microbial relationships and 3) develop metabolic and genomic modeling of these interactions to aid in interpreting the molecular mechanisms shaping the *Populus*-microbial interface.

Too much data: There has been an explosion of new methods for generating massive amounts of data. We are developing methods for evaluating data quality across large biological datasets (http://genomes.ornl.gov). As a first step, we have downloaded all public prokaryotic genome sequences from 6 different databases, with a total of more than 30,000 unique sequences. Most of the genomes are of draft quality (more than 80% of the current bacterial genome sequences). Within the Plant Microbe Interface (PMI) project, there is a need to know how reliable the genome sequences are for soil bacteria being used for comparisons. We examined the microbial DNA sequences available for complete, draft, and Short Read Archive genomes in GenBank as well as three other databases (Patric, KBase, and Broad) and assigned quality scores for more than 30,000 unique prokaryotic genome sequences. Scores were assigned using four categories (sequence quality, the presence of full length rRNA genes, tRNA composition (is there at least one tRNA for each of the 20 amino acids?) and the presence of a set of 120 conserved genes in all prokaryotes, and a combined score. Genomes with a quality score above a suitable threshold are being used to construct phylogenetic trees for the PMI project.

Community Structures: There are several major clades of bacteria associated with plants and soil samples. One commonly found environmental group of bacteria is the *Pseudomonas* genus. We have extracted all 258 *Pseudomonas* genomes from DOE KBase (in November, 2013). From these, two sets of trees were generated, one based on 16S rRNA, the other on ribosomal proteins. Twenty of these genomes were sequenced as part of the PMI project, and 19 formed a cluster with other *Pseudomonas fluorescens* related genomes, and one (*P. putida* related) was distantly related. A set of 47 genomes was selected to define the phylogenetic relationships of the 20 PMI genomes. Trees based on gene content and average amino acid identity (AAI) were constructed for these 47 genomes. The

19 PMI *P. fluorescens* related genomes cluster in a group of 24 genomes, and based on the AAI, there are 20 distinct species, with the 19 PMI species representing 17 distinct species. Within this cluster, there are a total of 22,000 different gene families, and a set of 2221 core gene families. For each of the 20 species, sets of species-specific gene families and their possible function have been identified.

Data Consolidation: One of the main priorities of PMI project has been consolidation of data resources across the project. The data consolidation initiative aims to help researchers collect data from multiple, disparate sources, and integrate into a single consolidated knowledgebase. As the project grows, demand for a single point of data access, data sharing, collaboration and provenance increases. PMI Knowledgebase now has an integrated workflow subsystem (http://pmi.ornl.gov) that enables PMI researchers to aggregate their project-generated data. The system provides tools and data-driven workflows that improves access to data resources, enhances data quality along with increased data protection. It has helped pipelining and to eliminate data duplication, thus providing a platform to consolidate data assets into a central repository resulting in efficient management and use of project-generated data.

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