

## 206. Plant-Microbe Interfaces: Correlotypes: Discovering complex, heterogenous genotypes in the Populus pan genome responsible for phenotypes and microbiomic associations.

Debra A. Weighill<sup>1,2</sup>, Gerald A. Tuskan<sup>1</sup>, David W. Ussery<sup>1</sup> and Dan A. Jacobson<sup>1\*</sup>  
([jacobsonda@ornl.gov](mailto:jacobsonda@ornl.gov))

<sup>1</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; <sup>2</sup>Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN.

<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.

Genome-wide association studies (GWAS) have focused on the analysis of individual single nucleotide polymorphisms (SNPs) in an attempt to find single alleles responsible for a phenotype. Although this has proven useful, it does not account for the fact that many phenotypes are the result of a combination of a broad range of genomic variants and cannot be simply described or controlled by a single gene or variant. Here we propose to develop a method with which to find the complex, heterogeneous collections of genomic variants responsible for many phenotypes. A number of different types of genome variants can all affect the phenotype of an organism including SNPs, small insertions/deletions (INDELs) and larger INDELs that equate to gene knockouts or replications.

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 genome-wide association methods that effectively only consider one SNP at a time or very localized genomic regions such as haplotypes. Genome sequences have previously been generated for over 1000 genotypes of Populus. We are mapping the resulting 100 billion reads to the reference genome 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 meta-assembly 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 the correlations between all pairs of genome variants. The correlations above 0.68 will be used to create a correlation network. Breadth first searches will be done on this correlation network 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.

To our knowledge, genome variant correlotypes have never been used at this scale before. One set of correlations from 4 million SNPs requires 32 trillion correlation coefficients to be calculated which would require a Petabyte of disk space should we chose to store them all. We will be generating many different sets of correlotypes based on global and local (phenotypic partitioning) correlations across the entire *Populus* pan genome. Thus, this will present the need for significant high performance computing and storage resources. The networks that result from these comparisons will contain millions of nodes and probably billions of edges.

The methods being developed here are designed to accelerate advances in plant-based bio-energy feed stocks, crop improvement and to further elucidate plant-microbial interactions. Although this approach will initially be used on the *Populus* 1000 genomes dataset the method itself is species agnostic and can be used in any project wishing to tie complex phenotypes to profiles of heterogeneous genomic variants. Not only have genomic variant correlotypes never been attempted at this scale but, they have also never been tested for significance with the sophisticated combinatorial GSA-based approach that we are developing.

*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. 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.*