PMI Performance Metric for FY18: Using genomics-based techniques, develop an approach to explore the functioning of plant-microbe interactions.

Q4 Metric: Describe the latest computational approaches needed to interpret the functioning of plant-microbe interactions.

Introduction

The Plant-Microbe Interfaces (PMI) project is a Scientific Focus Area directed towards understanding the dynamic interface that exists between plants, microbes and their environment. Project efforts are focused on characterizing and interpreting systems comprising the poplar tree (*Populus*) and its microbial community, in the context of favorable plant microbe interactions. We seek to define the relationships among these organisms in natural settings, dissect the molecular signals and gene-level responses of the organisms using natural and model systems, and rebuild the complexity of these systems using sequence characterized plants and microbes. *Populus* is an ideal host system for examining interfaces between plants and microbes and a leading candidate for bioenergy production. It is a dominant perennial component of many North American temperate forests and among only a few plant species that host both endo- and ectomycorrhizal fungal associates. Numerous other types of microorganisms can be found within, or closely associated with, various *Populus* tissues, and these organisms may range from highly beneficial to pathogenic with respect to effect on host fitness. Ultimately, an improved fundamental understanding of plant-microbe interfaces will enable the use of indigenous or engineered systems to address challenges as diverse as bioenergy production, environmental remediation, and carbon cycling and sequestration.

A deep understanding of the molecular and cellular events involved in establishing and maintaining these relationships throughout the host's life cycle requires a long-term commitment, expertise in multiple disciplines, and the continuous development of advanced experimental and computational tools to analyze and reconstruct these complex relationships. Computational analyses that take advantage of ORNL's advanced computing facilities are being leveraged towards the goal of understanding the complexity of the plant-microbe interface. This powerful approach is helping to detect genes, genotypes, environmental drivers and microbiome taxa that control these interactions. (Figure 1).

Data and computational resources

Microbial data resources. Understanding plant host-microbial selectivity and function is being accelerated by large scale microbiome isolate collections. In contrast to studies on crop or model plant species, *Populus* also hosts ecto- and endomycorrhizal fungi,[1-4] which can be more challenging to isolate and culture. We have developed extensive genomic resources for bacteria and fungi to allow detailed mechanistic studies in single inoculation and complex community experiments.

Our culture collection exceeds 5000 isolates that represent *Populus'* microbial community: bacterial (>3000 isolates) and fungal (>2000 isolates) groups associated with wild and field grown trees, as defined from rRNA, ITS amplicon, and metagenomics data. Many bacterial strains represent relatively abundant species in the *Populus* rhizosphere and endosphere (e.g., *Pseudomonas*, *Variovorax*, *Paraburkholderia*, *Rhizobium, Streptomyces,* and *Bacillus*), whereas

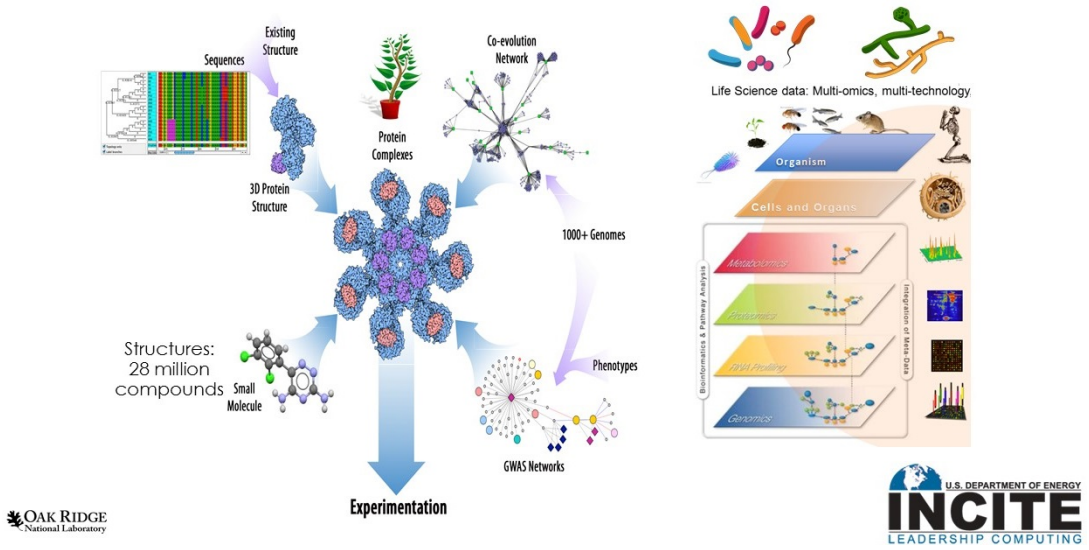

Figure 1. Computational analyses being leveraged towards the goal of understanding the complexity of plant-microbe interactions. Components of biological interactions and the tools and architecture necessary to elucidate them.

many of the fungal endophyte cultures represent new species.[5] Within the fungal collection, the majority of the isolates correspond to the Ascomycota; >50% belong to the Nectriaceae and Helotiaceae. The Helotiaceae endophytic fungi are transcriptionally active on *Populus* roots in bioassays, suggesting key active roles in overall plant health. Complementing this unique PMI microbial collection, >400 *Populus* bacterial and 65 fungal isolate (endo- and ectomycorrhizal) genomes, including members of the diverse ECM family Russulaceae have been sequenced in collaboration with the DOE JGI.[6-11] The fungal genome sequences are an expansion of our original model fungal symbiont *Laccaria bicolor*, which helped to define plant-microbe symbiosis.[12-14] These isolates and genome sequences have facilitated laboratory constructed community experiments[15,16] and comparative genomic analyses.[9,17-21] Furthermore, they are beginning to uncover characteristics of host-microbe selectivity and identify molecular mechanisms underlying community assembly and function. This microbial resource is complemented by extensive metagenomics analyses of samples from a *Populus* field site and has resulted in defining microbial community structure and functional potential. All these sequencing data are publicly available from DOE-JGI (https://img.jgi.doe.gov). These resources allow us to test specific hypotheses of how host genotypes and phenotypes shape symbiotic communities, define functional composition of the microbiome, and create environmentally and spatially defined niches.

Poplar data resources. With a collection of 28 million single nucleotide polymorphisms (SNP) derived from genomic sequence of 1084 variant genotypes of *P. trichocarpa*, there is a genetic marker every 17 bp in the population, allowing precise location of causal alleles linked to measured phenotypes including host-microbe interactions.[22-24] Additionally, we have xylem-,

leaf- and root-based expression (*i.e.*, RNAseq) quantitative trait nucleotide (eQTN) resources that enables us to identify cis- and trans-regulatory elements and loci controlling host-microbe interactions.[25] This is the most extensive eQTN resource attempted for any plant species. Furthermore, we have characterized the metabolome of >850 genotypes consisting of >440 metabolites (with ~half being fully identified) that have been integrated into genetic networks for targeted gene discovery.[26] Finally, a *P. trichocarpa* × *P. deltoides* pseudo-backcross pedigree has been established for uncovering the genetic factors controlling species-specific *Populus*-microbe interactions through quantitative trait locus (QTL) mapping. This population has been extensively genotyped using >3,500 genome-anchored SNP markers allowing for high resolution mapping.[27] As a complement to the GWAS eQTN resource, RNAseq analyses were performed on xylem transcriptomes of 310 progeny from the pedigree to establish species-dependent transcriptional regulatory networks of traits underlying microbial colonization.

Computational resources. The ORNL Leadership Computing Facility (OLCF) and the National Center for Computational Sciences (NCCS) provide resources for high-performance computational science and computing science research and offer storage resources as well as computing, networking, and visualization resources. The NCCS is a world leader in high-performance computing (HPC) and home to Summit, a system capable of processing more than 200,000 trillion calculations each second (200 petaflops [PF]). At 200 PF and >2.3 petabytes of memory, Summit is the world's most powerful computing platforms for open science and the central component of the OLCF. The Compute and Data Environment for Science (CADES): CADES is a fully integrated infrastructure offering compute and data services for researchers at ORNL. The CADES infrastructure supports multiple data security levels with separate open research and secures network enclaves. The open research enclave includes small-scale high-performance computing (HPC) and cloud compute systems and interconnects with DOE's Energy Sciences Network at 100 Gigabits per second. The Department of Energy Systems Biology Knowledgebase (www.KBase.us)[28] is an online software and data platform for researchers to address fundamental questions in systems biology. KBase has an easy graphical user interface that integrates both data and tools bypassing complex bioinformatic software and high-performance computing systems command line structure and hierarchy. KBase allows users to perform large-scale analyses and combine multiple lines of evidence to model plant and microbial physiology and community dynamics.

Collectively, these resources, coupled with our development of experimental system and analytical measurement approaches, allow us to make significant progress in advancing computational approaches for interpreting the functioning of plant-microbe interactions. Collectively, we are gaining a molecular-level understanding of plant colonization and material exchange and signaling between the plant and its microbial components.

Comparative genomic analysis

Comparative genomics provides a powerful tool to understand variation among organisms and helps identify the genes that are both conserved and divergent among species. We have leveraged our cultured *Populus* microbiome collection and carried out comparative genomics studies of key genera.[8,18,29] Furthermore, we used comparative genomics to identify genes which are enriched in the bacterial genomes relative to non-plant associated microbes,[9,18,21] thereby confirming the hypothesis that bacterial mutualists of plants have acquired genetic adaptations

that allow them to competitively colonize plant host tissues. These studies identified numerous functions/domains enriched in plant associated bacterial genomes, relative to non-plant associated and rhizosphere bacteria, including proteins that function in carbon metabolism, motility, chemotaxis, nitrogen fixation, quorum sensing, pili, type III and VI secretion systems, and siderophore production.

Comparative genomic analyses have also been carried out on *Pseudomonas* species isolated from *Populus*. Nineteen strains of 21 *Pseudomonas* isolated from *Populus* were classified as originating from either the rhizosphere (15) or endosphere (4). Metabolic models for each isolate were generated using publicly-available KBase tools. Metabolic processes for each isolate ranged from 1235 to 1324 reactions with 1151 reactions common to all models. Of the 281 reactions distributed differentially throughout the models, 42 were predicted transporters and 61 were not classified in KEGG maps. Over one-third of the differentially distributed reactions (105/281) were only found in endosphere isolates, while only one reaction was unique to rhizosphere isolates, suggesting significant diversity and bias in these environmental isolates.[21]

Metagenomic analysis and microbial community structure

Metagenomics and computational analyses are key to identifying the genes involved in specific *Populus* associations. Extensive analyses of bacterial community structure and functional potential were carried out using metagenomics data collected in collaboration with JGI, including metagenome assembly (and binning for individual genome assembly), gene calling and annotation. The resulting annotation was used for functional enrichment and differential analyses across compartments (soils, rhizosphere and root endosphere) and across host species (*P. deltoides vs. P. trichocarpa x deltoides* hybrids). Taxa were assigned using ParaKraken (described below), and Proportional Similarity Indices were used to define community structures and perform comparisons across compartments and host species. The results were modeled as networks for visualization and interpretation. These analyses have identified taxa and functions significantly enriched in endosphere samples relative to soil and rhizosphere compartments (e.g., regulatory, transporters, signaling, and adhesins). Mapping of related genes to our reference isolate genomes has allowed us to prioritize strains for hypothesis testing.

We also analyzed sequenced genomes of isolates from the *Populus* root microbiome to characterize organisms for the development of natural products (NP) potential and uncovered a wealth of novel chemical signatures in the *Populus* rhizosphere.[30] The assembled *Populus* metagenomes were analyzed by antiSMASH[31] for the presence of genes associated with secondary metabolite biosynthesis. Samples were compiled and separated based on location of sample extraction: endosphere, rhizosphere, or bulk soil. We found > 3400 individual gene clusters identified in 339 bacterial isolates, including 173 newly sequenced organisms, that were diverse across NP type and distinct from known NP clusters. The sequenced collection captures a broad phylogenetic diversity of the *Populus* microbiome that could lead to the discovery of compounds involved in communication and control of key plant-microbial interactions.

Traditional identification of microbiome abundance relies on marker-based operational taxonomic units (OTUs) such as 16S rRNA, which can limit the scope of organisms identified and requires a separate sequencing run. However, RNA-Seq offers the opportunity to identify both the host transcriptome and the high abundance organisms using whole genome sequences of

all associated organisms. To optimize taxonomic identification, we have developed a pipeline that combines traditional RNA-Seq with a modified and parallelized version of Kraken[32]: ParaKraken. As it is impractical to store every individual genome in a single database, ParaKraken, allows us to utilize whole genome sequences without binning the genomes into subsets and losing accuracy. Kraken uses a kmer approach, which breaks up genomes into short nucleotide sequences (31 mers) that allows high precision with fast matching speeds. For example, in an analysis of *Populus* undergoing drought stress, we have created databases that contain whole genomes from > 33,000 bacteria, 734 archaea, 571 fungi, 25 nematodes, 2 aphids, >7,000 viruses, and *P. deltoides.* We have since expanded the databases to every known sequenced organism, allowing for unprecedented classification ability.

In this drought stress experiment, *Populus deltoides* WV94 plants were subjected to two different drought conditions: an acute drought experiment where plants were left unwatered for eight days; and a cyclic drought experiment where plants were subjected to four drought and three re-watering cycles. Samples for RNA-seq were collected throughout each experiment, and transcriptomics showed a strong water deficit response in the host with some differences between cyclic and acute drought. However, the phytobiome was highly dependent on the cyclic or acute nature of the drought. To assess the relationship between the host and the organisms living on the leaves we looked at both co-differential expression/abundance and associations created by DUO (Figure 2), an algorithm that identifies positive and negative relationships between taxa and genes. Effects of drought on both the plant and phytobiome are dependent on severity and prior drought exposure, with more severe drought causing severe metabolic impairment. Despite this, cyclic severe drought plants have relatively higher photosynthetic and ROS metabolism than severe acute drought; conversely, severe acute drought plants have a phytobiome that is more associated with plant death and pathogenicity. Changes that occur during cyclic drought suggest that these plants are better acclimated to drought compared to those subjected to a progressive acute drought.

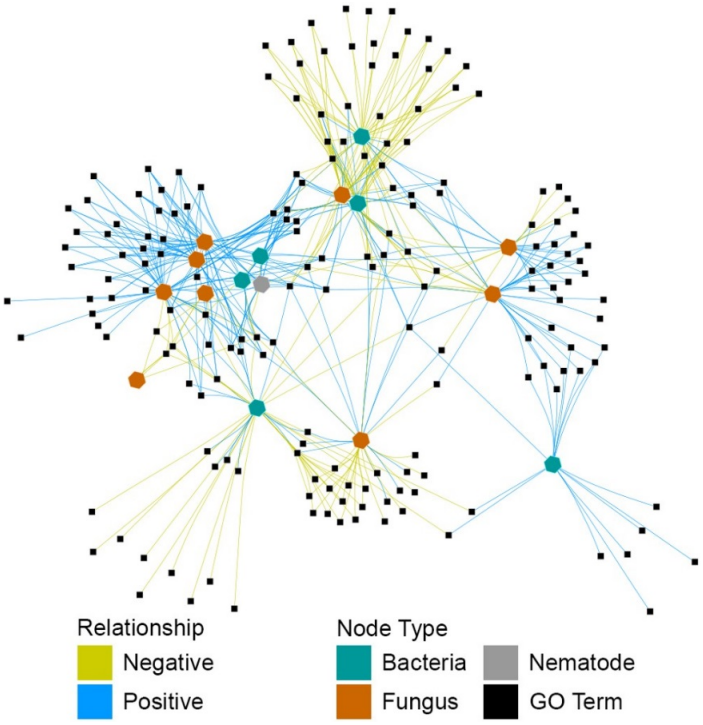

Figure 2. DUO associations between taxa and GO terms. *Rhizophagus* and *Trichinella* are associated with poplar host ROS metabolism, further supporting the hypothesis that *Rhizophagus* may be mitigating drought stress and *Trichinella* may be modulating the host immune system. Additionally, *Streptomyces* and *Pseudomonas* are positively associated with disease resistance genes, which may help protect plants from disease during drought stress. Also, acute drought and cyclic drought have non-overlapping associations of disease resistant genes and taxa suggesting that both conditions are experiencing different types of disease stresses.

High-Performance Computing Applications

Artificial Intelligence – Genome Wide Association Phytobiome Analysis (AI-GWAPA) machine learning, deep learning and general artificial intelligence (AI) techniques elucidate the interactions between microbial and viral constituents of the *Populus trichocarpa* GWAS population arrayed in common gardens in the Pacific Northwest.[33] Taxa were identified from the leaf and xylem transcriptome using ParaKraken; to improve our confidence in taxonomic assignment we processed the sparse data for putative outlier taxa using factorization machines (FM) outlier approach in Pytorch.[34] The phytobiome taxa that remained after the FM-outlier analysis were treated as phenotypes in a genome wide association analysis (Figure 3). Approximately 10 million single-nucleotide polymorphisms (SNPs) were then filtered and used as the genotype information. Only SNPs with a minor allele frequency greater than 0.01 were analyzed. DUO was used to compare taxa abundance across the population. Taxa abundance vectors are compared pair-wise after categorizing individual measurements into High, Medium or Low, based on the quantiles of the entire dataset. The metric then evaluates how correlated the high (up) /low (down) components of the vector are. This results in 4 correlation values: UU, UD, DU and DD (U=up,D=down). Mutualism is suggested by a UU or DD correlation, while antagonism is suggested by a UD or DU correlation. SSN values are generated by removing a genotype and recalculating the DUO metric. By doing this for all genotypes and then observing the resultant change from the original metric, we can estimate the genotype's contribution to the DUO correlations.[33]

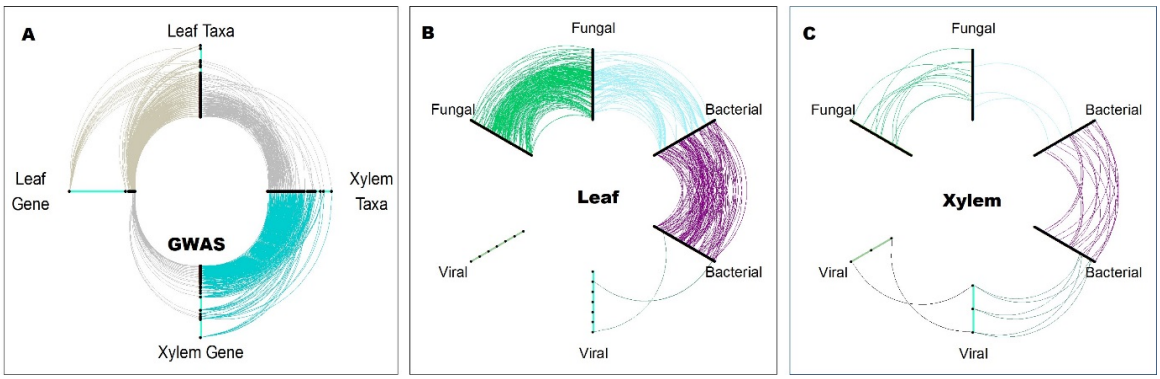

Figure 3. Hive plots of the GWAS results. An arch drawn between nodes on each axis indicates a relationship. (A) The four axes represent either genes or taxa nodes, from the leaf/xylem GWAS analysis. Nodes are arranged based on the number of taxa associations (the outer nodes therefore have higher connectivity), similarly taxa nodes are arranged on the number of GWAS results. Arcs between taxa and gene nodes indicate a significant GWAS result. Taxa that are both in leaf and xylem are connected, similarly genes both in leaf and xylem are connected. (B) Hive plot of the neighborhood of the gene associated with a particular taxon from the leaf. We see that there is one virus that is associated with a gene which in turn associates with a bacterium. Similarly, there are large fungal to fungal, bacterial to bacterial and bacterial to fungal neighborhoods. (C) Similar to B, but for the xylem GWAS instead. Here we see a viral to viral neighborhood that is absent from the leaf, a few more viral to bacterial neighborhoods.

This provides a comprehensive framework for a systems biology approach to the analysis and interpretation of the complex interactions between a host and its phytobiome. From the meta-transcriptome samples, we obtain a snapshot of the putative constituents of this phytobiome. The FM learning allows us to model up to 3rd order interactions between the respective taxa, thereby retaining signal that would otherwise be missed using standard metric-based analysis. GWAS analysis helps to uncover the nature of this complex interaction. We find a few highly connected genes involved in functions such as phospholipid transportation, protein degradation, transcriptional regulation, etc. With more advanced metrics, such as DUO, we see differences in the types mutualistic/antagonistic relationships when comparing leaf and xylem. With sample specific networks we can start to understand the genotypic effect on these relationships. By further using a deep learning-based protein interaction model we can work towards a protein level understanding of this dynamical system.

Annotation-based Lines of Evidence (LOE) Network Mining. Biological organisms are complex systems that are composed of functional networks of interacting molecules and macro-molecules. Complex phenotypes are the result of orchestrated, hierarchical, heterogeneous collections of expressed genomic variants. However, the effects of these variants are the result of historic selective pressure and current environmental and epigenetic signals, and, as such, their co-occurrence can be seen as genome-wide correlations in a number of different manners. With the computational tools we have used to help deconstruct elements of biomass recalcitrance,[26] we can look at other complex polygenic phenotypes of importance to plant-microbe interactions: microbial recognition, signaling; disease susceptibility/resistance and defense response, apoptosis, cell wall regulation, signaling, biotic stress, etc.

This type of analysis makes use of data derived from many different data layers (Figure 4). A Lines Of Evidence (LOE) scoring system is developed to integrate the information in the different layers and quantify the number of lines of evidence linking genes to target functions. This new scoring system was applied to quantify the lines of evidence linking genes to lignin-related genes and phenotypes across the network layers and allowed for the generation of new hypotheses surrounding potential new candidate genes involved in lignin biosynthesis.[26] The resulting Genome Wide Association Study networks, integrated with Single Nucleotide Polymorphism (SNP) correlation, co-methylation, and co-expression networks through the LOE scores are proving to be a powerful approach to determine the pleiotropic and epistatic relationships underlying cellular functions and, as such, the molecular basis for complex phenotypes described above.

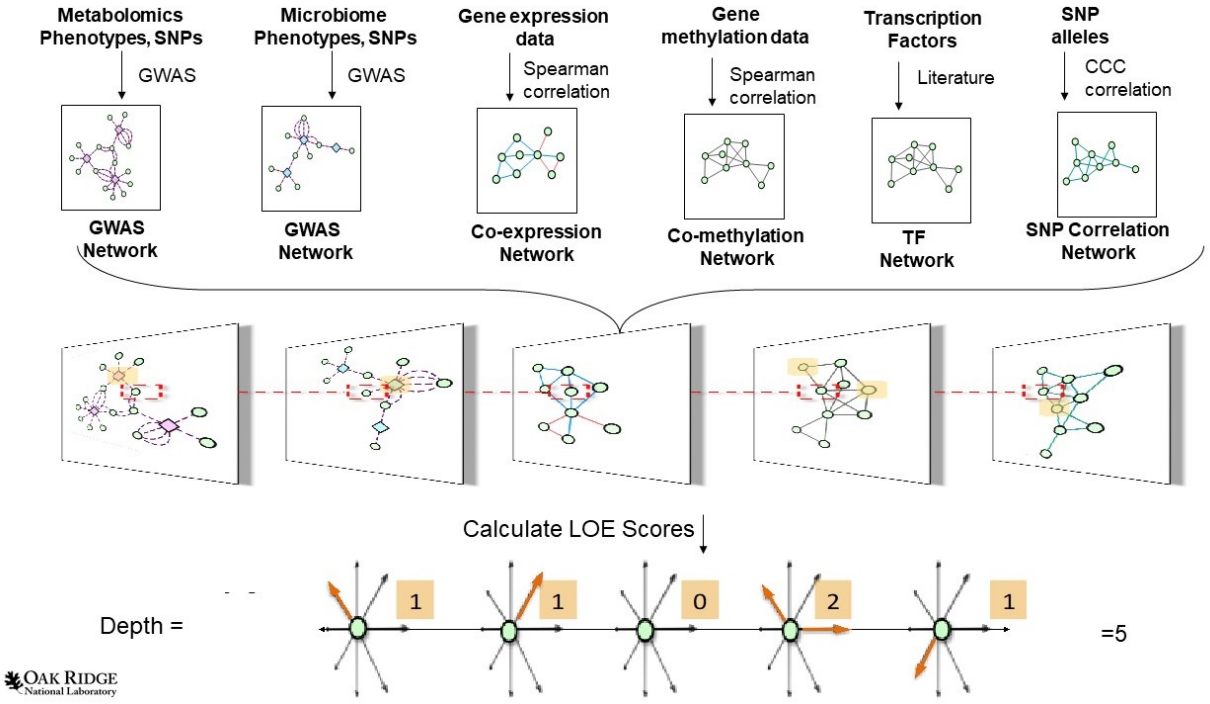

Figure 4. Deeper Discoveries in Systems Biology: The Balance Between Type 1 and Type 2 Error. Our ability to reconstruct the entirety of a complex biological system improves as the number of population-scale endo-, meso- and exo-phenotypes are measured and combined with deep layers of experimental data collected on individual genotypes.

Co-evolutionary networks for understanding host and symbiotic systems. SNP Correlation/Co-evolution – Leading the Way to Exascale. We have achieved 2.36 ExaOps (mixed precision ExaFlops) at 4,560 nodes (99% of Summit) using the Tensor Cores Equivalent to 86.4 TF per GPU for the whole computation (including communications and transfers) at 4,560 nodes. This superior scaling is made possible by Summit fat tree network with adaptive routing, which is >15,000X faster than the closest competing code.[35]

Computation of the mathematical relationships between pairs of vectors is required in many science domains. In the field of genomics, the Custom Correlation Coefficient (CCC)[36] was developed to calculate the correlation between mutations (SNPs) across a population of

individuals. This can be used to identify groups of SNPs which tend to co-occur in a population, and consequently can be used to find combinations of SNPs which associate with certain phenotypes, such as a disease phenotype.[37] CCC also considers genetic heterogeneity and finds correlations. The resulting sets extracted from CCC network topologies represent higher-order combinatorial sets that allow for much more complex interactions to be tested than simple SNP pairs tested by extant methods.

We are actively developing new computational methods to account for any influences of population structure on the global co-evolutionary signatures (SNP correlations) that we see in the *P. trichocarpa* GWAS population. We will use a genome relationship matrix to develop bootstrap sampling strategies for iterative SNP correlation network building to test and adjust for population structure in a scalable manner. The goal is to suppress the effect of population structure. RLK activation and salicylate structure will be the examples for this method. We have developed a test framework for clustering individuals into sub-populations based on pairwise kinship, followed by performing Leave-One-Subpopulation-Out bootstrapping of pairwise SNP correlation. We have also established a metric to determine if a SNP correlation is or is not affected by presence of population substructure. We have successfully completed a test run of the new GPU Correlation Coefficient Calculations (CCC) software on Titan[38] and are now filtering the set of 10M SNPs for minor allele frequency and call rate and will run the resulting set of SNPs through the pipeline. The GPU optimized CCC algorithm we have developed for this work can meet the needs of these types of ambitious projects. Thus, pipelines we have constructed will enable the high-throughput discovery of epistatic networks directly from population genomics and phenomics data, providing a new tool for GWAS and QTL studies. The potential impact is substantial. Once created, these epistatic networks and SNP sets can be used to test for associations with any phenotype that has been measured across a population. As such, this will allow for a tremendous amount of new information to be derived from publicly available phenotypes and plant genomes as they become available.

Proportional Similarity metric for pleiotropy discovery. Our overall research project in each of the species that we are studying involves the creation of extensive systems biology models of each species that capture the molecular interactions in the cell that lead to emergent properties and complex, organismal-scale phenotypes. This goal includes the determination of genome-variant to phenotype relationships for all available phenotypes. For some species we already have collected thousands of phenotypes (160,000 phenotypes in the case of our bioenergy project with *Populus trichocarpa*). As such, pleiotropy is an important factor that needs to be captured in these extensive models. Pleiotropy is the phenomenon in which a gene is involved in multiple phenotypes.[39]

Explainable AI with iterative Random Forests (iRF) – Recent developments include random intersection trees (RITs)[40] and iterative RFs (iRFs)[41] that we have re-written and scaled up to run on ORNL supercomputers such as Summit. These approaches use density estimates (or response surfaces) that can be mined not only for feature importance but also for interactions between features. It is now possible to identify interactions of any form or order at the same computational cost as main effects[40,41] enabled by importance sampling on the space of all subsets (order 2P). However, there is a major bottleneck on the utility of these methods – they specialize in one- or low-dimensional Y. Hence, they are of limited utility in data fusion

exercises where we wish to understand a potentially high dimensional process (*e.g.,* a tensor) as a function of another high-dimensional process. Classically, one would simply regress each Y vector one at a time, but it is costly, and also lossy. We are often interested in the joint distribution of Y and X in detail; *e.g.,* we may care about the variance of Y given X for process control or stability analysis. Tensor-ready versions of ensembled density estimators have the potential to open fundamentally near areas of research in statistical machine learning: such algorithms will inherit the astounding interpretability of RF "for free," and stand to change the way we think about coupled high-dimensional processes.

From Matrices to Cubes to Polytopes. Furthermore, we are not just interested in pairwise matrix comparisons. Plant-microbe systems contain many layers of data/molecule types (genome, epigenome, chromatin structure, transcriptome, proteome, metabolome, lipidome, microbiome, *etc*.) and various layers of emergent phenotypes that range from localized mechanisms up to whole organism scales. We are interested in finding associations within and across each of these layers which can be represented as matrices. If one considers spatial relationships for the attributes of one of those matrices, then each matrix becomes a data cube. However, each of those layers may have other attributes including temporal (longitudinal), quantum chemical, and interaction tensors to name a few and will thus become polytopes. As such there is a need for algorithms that can find higher order associations within and across an arbitrary number of polytopes of differing scales. Algorithms that we are currently developing, such as Tensor iterative Random Forests (TiRFs) are aimed at this problem.[32,40,41]

Summary

With novel computational tool development and advanced analytical infrastructure, we are poised to determine the commonalities and genetic underpinnings of bacterial and fungal recognition, discover new regulators, and define molecular mechanisms underlying selectivity and reveal signaling cascades leading to root colonization. Further details on PMI project efforts can be found in our research publications. A complete listing of PMI project publications is available at https://pmiweb.ornl.gov/portfolio/

References

1 Bonito, G. *et al.* Atractiella rhizophila, sp. nov., an endorrhizal fungus isolated from the Populus root microbiome. *Mycologia* **109**, 18-26, doi:10.1080/00275514.2016.1271689 (2017).

2 Cregger, M. A. *et al.* The Populus holobiont: dissecting the effects of plant niches and genotype on the microbiome. *Microbiome* **6**, 31, doi:10.1186/s40168-018-0413-8 (2018).

3 Gottel, N. R. *et al.* Distinct microbial communities within the endosphere and rhizosphere of *Populus deltoides* roots across contrasting soil types. *Appl Environ Microbiol* **77**, 5934-5944, doi:10.1128/AEM.05255-11 (2011).

4 Shakya, M. *et al.* A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature Populus deltoides trees. *PLoS One* **8**, e76382, doi:10.1371/journal.pone.0076382 (2013).

5 Aime, M. C., Urbina, H., Liber, J. A., Bonito, G. & Oono, R. Two new endophytic Atractiellomycetes, Atractidochium hillariae and Proceropycnis hameedii. *Mycologia* **110**, 136-146, doi:10.1080/00275514.2018.1446650 (2018).

6 Bonito, G. *et al.* Isolating a functionally relevant guild of fungi from the root microbiome of Populus. *Fungal Ecology* **22**, 35-42, doi:10.1016/j.funeco.2016.04.007 (2016).

7 Brown, S. D. *et al.* Draft genome sequence of Rhizobium sp. strain PDO1-076, a bacterium isolated from Populus deltoides. *J Bacteriol* **194**, 2383-2384, doi:10.1128/JB.00198-12 (2012).

8 Klingeman, D. M. *et al.* Draft Genome Sequences of Four Streptomyces Isolates from the Populus trichocarpa Root Endosphere and Rhizosphere. *Genome announcements* **3**, doi:10.1128/genomeA.01344-15 (2015).

9 Levy, A. *et al.* Genomic features of bacterial adaptation to plants. *Nature genetics* **50**, 138-150, doi:10.1038/s41588-017-0012-9 (2018).

10 Looney, B. P. *et al.* Russulaceae: a new genomic dataset to study ecosystem function and evolutionary diversification of ectomycorrhizal fungi with their tree associates. *New Phytol* **218**, 54-65, doi:10.1111/nph.15001 (2018).

11 Uehling, J. *et al.* Comparative genomics of Mortierella elongata and its bacterial endosymbiont Mycoavidus cysteinexigens. *Environ Microbiol* **19**, 2964-2983, doi:10.1111/1462-2920.13669 (2017).

12 Martin, F. *et al.* Symbiotic sequencing for the *Populus* mesocosm. *New Phytologist* **161**, 330-335, doi:DOI 10.1111/j.1469-8137.2004.00982.x (2004).

13 Pellegrin, C. *et al.* Laccaria bicolor MiSSP8 is a small-secreted protein decisive for the establishment of the ectomycorrhizal symbiosis. *bioRxiv*, doi:10.1101/218131 (2017).

14 Plett, J. M. *et al.* A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Curr Biol* **21**, 1197-1203, doi:10.1016/j.cub.2011.05.033 (2011).

15 Bible, A. N. *et al.* A Carotenoid-Deficient Mutant in Pantoea sp. YR343, a Bacteria Isolated from the Rhizosphere of Populus deltoides, Is Defective in Root Colonization. *Frontiers in microbiology* **7**, 491, doi:10.3389/fmicb.2016.00491 (2016).

16 Timm, C. M. *et al.* Two Poplar-Associated Bacterial Isolates Induce Additive Favorable Responses in a Constructed Plant-Microbiome System. *Frontiers in plant science* **7**, 497, doi:10.3389/fpls.2016.00497 (2016).

17 de Freitas Pereira, M. *et al.* Secretome Analysis from the Ectomycorrhizal Ascomycete Cenococcum geophilum. *Frontiers in microbiology* **9**, 141, doi:10.3389/fmicb.2018.00141 (2018).

18 Jun, S. R. *et al.* Diversity of Pseudomonas Genomes, Including Populus-Associated Isolates, as Revealed by Comparative Genome Analysis. *Appl Environ Microbiol* **82**, 375-383, doi:10.1128/AEM.02612-15 (2016).
19 Martino, E. *et al.* Comparative genomics and transcriptomics depict ericoid mycorrhizal fungi as versatile saprotrophs and plant mutualists. *New Phytol* **217**, 1213-1229, doi:10.1111/nph.14974 (2018).
20 Schaefer, A. L. *et al.* LuxR- and luxI-type quorum-sensing circuits are prevalent in members of the *Populus deltoides* microbiome. *Appl Environ Microbiol* **79**, 5745-5752, doi:10.1128/AEM.01417-13 (2013).
21 Timm, C. M. *et al.* Metabolic functions of Pseudomonas fluorescens strains from Populus deltoides depend on rhizosphere or endosphere isolation compartment. *Frontiers in microbiology* **6**, 1118, doi:10.3389/fmicb.2015.01118 (2015).
22 McKown, A. D. *et al.* Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol* **203**, 535-553, doi:10.1111/nph.12815 (2014).
23 Porth, I. *et al.* Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. *New Phytol* **200**, 710-726, doi:10.1111/nph.12422 (2013).
24 Slavov, G. T. *et al.* Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol* **196**, 713-725, doi:10.1111/j.1469-8137.2012.04258.x (2012).
25 Zhang, J. *et al.* GWAS and eQTL analyses reveal roles of hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase 2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors. . *New Phytol* **in press** (2018).
26 Weighill, D. *et al.* Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery. *Frontiers in Energy Research* **6**, doi:UNSP 30 10.3389/fenrg.2018.00030 (2018).
27 Muchero, W. *et al.* High-resolution genetic mapping of allelic variants associated with cell wall chemistry in *Populus*. *BMC Genomics* **16**, 24, doi:10.1186/s12864-015-1215-z (2015).
28 Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* **36**, 566-569, doi:10.1038/nbt.4163 (2018).
29 Brown, S. D. *et al.* Twenty-one genome sequences from Pseudomonas species and 19 genome sequences from diverse bacteria isolated from the rhizosphere and endosphere of Populus deltoides. *J Bacteriol* **194**, 5991-5993, doi:10.1128/JB.01243-12 (2012).
30 Blair, P. M. *et al.* Exploration of the biosynthetic potential of the Populus microbiome. *mSystems* **in press** (2018).
31 Weber, T. *et al.* antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237-243, doi:10.1093/nar/gkv437 (2015).
32 Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46, doi:10.1186/gb-2014-15-3-r46 (2014).
33 Jones, P. C. *et al.* in *2018 Genomic Sciences Program Annual Principal Investigator Meeting* (Tysons Corning, VA., 2018).
34 Paszke, A. *et al. Automatic differentiation in PyTorch*, <https://openreview.net/pdf?id=BJJsrmfCZ> (2017).

35 Hines, J. *Genomics code exceeds exaops on SUMMIT supercomputer*, <https://www.olcf.ornl.gov/2018/06/08/genomics-code-exceeds-exaops-on-summit-supercomputer/> (2018).

36 Climer, S., Yang, W., de las Fuentes, L., Davila-Roman, V. G. & Gu, C. C. A custom correlation coefficient (CCC) approach for fast identification of multi-SNP association patterns in genome-wide SNPs data. *Genet Epidemiol* **38**, 610-621, doi:10.1002/gepi.21833 (2014).

37 Climer, S., Templeton, A. R. & Zhang, W. Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis GWAS. *PLoS Comput Biol* **10**, e1003766, doi:10.1371/journal.pcbi.1003766 (2014).

38 Joubert, W., Nance, J., Climer, S., Weighill, D., and Jacobson, D. . Parallel Accelerated Custom Correlation Coefficient Calculations for Genomics Applications. *CoRR* **abs/1705.08213** (2017). <http://arxiv.org/abs/1705.08213>.

39 Joubert, W. J. *et al.* in *SC18*.

40 Shah, R. D. & Meinshausen, N. Random Intersection Trees. *J Mach Learn Res* **15**, 629-654 (2014).

41 Basu, S., Kumbier, K., Brown, J. B. & Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proc Natl Acad Sci U S A* **115**, 1943-1948, doi:10.1073/pnas.1711236115 (2018).