Peta- and Exa-scale for Arabidopsis in KBase

Michael R. Garvin[1]* (garvinmr@ornl.gov), David Kainer, Jared Streich, Angelica M. Walker, Daniel A. Jacobson

[1]Oak Ridge National Laboratory, Oak Ridge, TN

Project Goals: To create large interaction networks using SNPs, transcriptomic, metabolomic, microbiome taxa, climatypes and other types of phenotype data as well as tissue-specific epigenetic and expression data. The creation of such network-based models involves the running of ensembles of custom correlation metrics, mixed linear models and explainable-AI methods at extreme scales on the Oak Ridge Leadership Computing Facility supercomputer Summit. The results of these workflows are modeled as networks (and hyper-networks) in order to provide an integrated systems biology view of an organism. This includes using a number of previously developed analysis & modeling methods that, in combination with explainable-AI approaches, predict (high combinatorial order) epistatic architectures for all available traits. This is being developed as a community resource in KBase for the model plant *Arabidopsis thaliana*.

The rapid increase in biological assays, high-throughput phenotyping studies, and computational prediction capabilities has resulted in an enormous wealth of biological data for many model species. These data layers (e.g., genomic, transcriptomic, metabolomic, protein-protein interactions, climatype, phenomic) are developed with the goal of understanding the operation of overarching biological systems and discovering the basis for emergent phenotypes. Each data layer is often interpreted within the context of that specific dataset, which provides useful, but limited, insights. This is because biological elements rarely operate in isolation within and between the cellular environment; data from a single layer reveals only part of the story, and can possibly be misleading. As the primary model plant species, there is a great deal of publicly available data derived from *Arabidopsis thaliana*. We are creating data products composed of *A. thaliana* data layers that will be integrated into KBase. Furthermore, we will develop novel algorithms that can be ported into KBase in order to provide a flexible, open-source product for plant biologists interested in *A. thaliana*. We will also develop a systems-biology resource enabling KBase users to rank candidate genes and to predict the function of unknown genes. While each layer of data intrinsically has a level of ability to predict gene function based on the connectivity and topology of the nodes and edges of its network, some layers are more informative than others. Various methods of analysis, including explainable machine learning, can be applied to evaluate the predictive ability of each layer by cross-validation using known genes from well characterized pathways or interaction networks. This resource will also be integrated into the KBase knowledge engine, providing a powerful new tool to plant biologists.

Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. This program is supported by the U. S. Department of Energy, Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research, under FWP ERKP972.