The DOE Systems Biology Knowledgebase: Plant Science Domain

Doreen Ware[1, 2]* (ware@cshl.edu), Sergei Maslov[4], Shinjae Yoo[4], Dantong Yu[4], Michael Schatz[1], James Gurtowski[1], Matt Titmus[1], Jer-ming Chia[1], Sunita Kumari[1], Andrew Olson[1], Shiran Pasternak[1], Jim Thomason[1], Ken Youens-Clark[1], Mark Gerstein[5], Gang Fang[5], Darryl Reeves[5], Pam Ronald[6], Chris Henry[7], Sam Seaver[7], and David Weston[3], Priya Ranjan[3], Robert Cottingham[3], Rick Stevens[7], **Adam Arkin**[8]

1 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 2 United States Department of Agriculture–Agriculture Research Service; 3 Oak Ridge National Laboratory, Oak Ridge, TN; 4 Brookhaven National Laboratory, Upton, NY; 5 Yale University, New Haven, CT; 6 University of California, Davis, CA; 7 Argonne National Laboratory, Argonne, IL; 8 Lawrence Berkeley National Laboratory, Berkely, CA. http://kbase.us

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets, and analytical tools and to demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance, and use of predictive models and methods in the study of microbes, microbial communities and plants. The plant component of the KBase will allow users to model genotype-to-phenotype relationships using metabolic and functional networks as well as phenotype measurements and 'omic data. It will also support the reconstruction of new metabolic and functional networks based on expression profiles, protein-DNA, and protein-protein interactions. To accomplish this, we will provide interactive, data-driven analysis and exploration across multiple experiments and diverse data-types. We will provide users access to comprehensive collections of 'omics datasets together with relevant analytical tools and resources.

The major goal for the KBase plants area is to model genotype-to-phenotype relationships through analysis and integration of genomic, transcriptomic, methylomic sequencing data, metabolite and phenotype measurements, and the reconstruction of metabolic and functional networks based on expression profiles, protein-DNA, and protein-protein interactions. To accomplish this goal, KBase will provide interactive, data-driven analysis, and exploration across multiple experiments and diverse data-types. We will provide users access to comprehensive datasets from high-throughput experiments together with relevant analytical tools and resources. Users will be provided with a platform to analyze their own experimental data, integrate publicly available data from other 'omics' platforms, and have these results incorporated into a data exploration framework. We will aid in the translation of basic science through exploration across large 'omics experiments and the initiation of hypothesis-driven genetics studies without the overhead of data flow and management.

The KBase plants effort will consist of two major components, 1) genotyping workflows and 2) data exploration and prediction tools.

Genotyping workflows: Exponential growth in digital demands has motivated extensive research into improved algorithms and parallel systems, especially for genotyping samples, monitoring expression levels, and a host of other important biological applications.

Genotyping workflows will leverage our recent development of Jnomics, as our new Hadoop-based open-source package for rapid development and deployment of cloud-scale sequence analysis tools. Jnomics provides many pre-built tools out-of-the-box that accelerate common tasks, such as mapping, sorting, merging, filtering, and selection, to be performed as distributed tasks spread across a cluster. New tools

can be easily created using an open-source Java API, especially for large-scale genotyping and expression analysis. Because it builds on Hadoop, Jnomics tools inherit Hadoop's efficiency and scalability for very large datasets such as the billions of short reads necessary for genotyping many large plant genomes. Furthermore, Jnomics is "file format agnostic", allowing it to seamlessly read and write most common sequence file formats (SAM, BAM, BED, fastq, fasta), making it easy for Jnomics to interface with other components.

Data exploration and prediction: While the benefits of integrating large independent data sets for value-added research are vast, the influx of raw biological data presents unprecedented challenges in data management and representation. We believe that effective visualization will play a key role in mining and exploration of 'omics data sets. While genome browsers are useful for anecdotal relationships between feature sets and genome maps — especially for bench scientists studying a region of interest — they are not ideal for correlational interrogation of the data. Effective visualization is crucial for generating hypotheses. Displays must be flexible insofar as they allow the end-user to represent highly dimensional data in various forms without compromising pairwise relationships among data points. We intend to develop aesthetic and interactive interfaces that enable researchers to intuitively draw relationships in their complex queries. The visualizations will allow end-users to transform massive data sets in real-time in order to explore latent system-wide relationships among the data. Data exploration tools will focus on integration and visualization of data from multiple 'omics studies, genetic variation, network and pathways models as well as phenotypic association. Users will, for example, map variability onto metabolic models, highlighting pathways predicted to be impacted by genetic variation. By overlaying genetic variation within the context of gene network models, we can predict pathways that are impacted by the pool of genetic diversity in the population.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.