Deep Green: Structural and Functional Genomic Characterization of Conserved Unannotated Green Lineage Proteins

Jianlin Cheng[1], Eric Knoshaug[2], Vladimir Lunin[2], Ambarish Nag[2], Ru Zhang[3], and **James Umen**[3]*
(jumen@danforthcenter.org)
*presenting author.

[1]University of Missouri, Columbia, MO; [2]National Renewable Energy Laboratory, Golden, CO; [3]Donald Danforth Plant Science Center, St. Louis, MO.

Project Goals and Overview

Sequence-homology and experimental approaches have enabled functional annotation of many plant and algal genes, but around half of the predicted proteins in most sequenced green-lineage genomes remain as unknowns, with no information on structure or function. While some of these unknown proteins are lineage-specific or even species-specific, a sizable number are conserved within the Viridiplantae (green algae + land plants) or within large sub-groups of plants (e.g. monocots, dicots). These unknown conserved proteins are likely to play important roles in core plant biological processes. Through this project, we will establish a pipeline and associated databases for structural prediction and functional characterization of plant proteins of unknown function (Deep Green proteins), including around 500 unknown proteins from the model dicot *Arabidopsis thaliana* (Arabidopsis) and/or the model C4 bioenergy monocot *Setaria viridis* (Setaria) with homologs in the model green alga *Chlamydomonas reinhardtii* (Chlamydomonas), where we can perform high throughput functional genomics screening. This project leverages expertise in structural genomics and high-performance bioinformatics computing from team members at the National Renewable Energy Laboratory (NREL), omics-based computational predictions from team members at University of Missouri (MU), and algal and plant functional genomics expertise from team members at Donald Danforth Plant Science Center.

The Deep Green project is divided into five major objectives to characterize and annotate unknown plant proteins, which include **1)** Assembly and ongoing curation of the Deep Green candidate protein set; **2)** *in silico* structural predictions and network analyses to assign structures and predict function; **3)** Assembly and curation of reverse genetic resources in Chlamydomonas; **4)** functional genomics characterization and prioritization in Chlamydomonas; and **5)** structural validation of selected candidates and functional validation in two important reference plants, Arabidopsis and Setaria.

The rich new data resources produced under the Deep Green project will be curated in one or more public databases, including DOE-supported KBase. These data will help guide researchers in investigating the contribution of conserved unknown proteins to diverse aspects of photosynthetic biology that impact photosynthesis, biomass accumulation, and stress responses. This work will also help fill a major gap in the annotation for large sets of plant proteins whose structures and functions have not yet been characterized, and which represent a relatively untapped resource for bioenergy and synthetic biology applications that underlie the DOE mission.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-SC0020400.