Title: Curation and Characterization of Conserved Green Lineage Proteins

Authors: James Umen[1]* (jumen@danforthcenter.org), Chen Chen[2], Jianlin Cheng[2], Eric Knoshaug[3], Jian Liu[2], Vladimir Lunin[3], Ambarish Nag[3], Huong Nguyen[1], Peter St. John[3], Peipei Sun[1], Ru Zhang[1].

*PI, senior author, presenter

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

Website: http://tulip.rnet.missouri.edu/deepgreen/deepgreen/index.html

Project Goals:
Around half of the predicted proteins in most sequenced green-lineage genomes remain as unknowns, with no information on structure or function. Through this project, we will characterize 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 will perform high-throughput functional genomics screening. Our objectives are: 1. Assembly and 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 Arabidopsis and Setaria.

Abstract:
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). 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 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. Ongoing work on Deep Green proteins has produced three curated lists of unknown protein families from the three focal species Arabidopsis, Setaria and Chlamydomonas as well as a overlaps between these sets established based on sequence

homology criteria. A manuscript describing the curation process and preliminary characterization of Deep Green proteins is in preparation, including 460 members from Chlamydomonas, and . Under Objective 3 (assembly of reverse genetic resources for Chlamydomonas Deep Green Proteins) we have identified pre-existing CLiP library(1) insertions for 345 mutants, and for the remaining 115 we have adapted an efficient genome editing procedure (2) that uses CRISPR-Cas9 and a barcoded selectable marker cassette to generate tagged mutants. Under Objective 2 we applied our MULTICOM tool ranked among top predictors in the 14 Critical Assessment of Protein Structure Prediction (CASP14) to predict the tertiary structures and structural features (i.e., secondary structure, solvent accessibility, disorder, domain boundaries, inter-residue contacts) for 825 out of 1658 shared Setaria and Arabidopsis Deep Green proteins. The prediction results are available at a user-friendly, browsable website (http://tulip.rnet.missouri.edu/deepgreen/deepgreen/index.html). These results are being compared with *de novo* structure predictions obtained using the I-TASSER software and the Rosetta Abinitio Relax module. Together, 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.

References:

1. X. Li, *et al.*, An Indexed, Mapped Mutant Library Enables Reverse Genetics Studies of Biological Processes in Chlamydomonas reinhardtii. *The Plant Cell* **28**, 367–387 (2016).

2. T. Picariello, *et al.*, TIM, a targeted insertional mutagenesis method utilizing CRISPR/Cas9 in Chlamydomonas reinhardtii. *Plos One* **15**, e0232594 (2020).

Funding statement:

This research is supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program grant no. DE-SC0020400 and by JGI Community Sequencing Proposal 506032