New Computational Pipelines to Prioritize Candidate Genes for Optimal Biomass Production under Drought in C4 Plants *Setaria viridis* and *Sorghum bicolor*

Cheng Zhao[1], Fan Lin[1], Elena Lazarus[1], Pascal Schläpfer[1], Hye-In Nam[1], Scott Lee[3], Phil Ozersky[3], Edward J. Wolfrum[2], Jennifer Barrett[3], Allen Hubbard[3], Hui Jiang[3], Xiaoping Li[3], Erica Agnew[3], Todd Mockler[3], **Ivan Baxter[3]**, Seung Y. Rhee[1,*] (srhee@carnegiescience.edu)

[1]Carnegie Institution for Science, Stanford, CA; [2]National Renewable Energy Lab, Golden, CO; [3]Donald Danforth Plant Science Center, St. Louis, MO

Project Goals: This project aims to leverage *Setaria viridis* as a model system to develop novel technologies and methodologies to redesign the bioenergy feedstock *Sorghum bicolor* to enhance water use and photosynthetic efficiencies.

URL: www.foxmillet.org

C4 plants, such as *Sorghum bicolor* and *Setaria viridis*, have CO_2 concentrating mechanisms in specialized cell types (bundle sheath and mesophyll cells) to enhance water use and photosynthetic efficiencies. Current mathematical modeling of C4 photosynthesis does not sufficiently capture leaf biochemical and anatomical phenotypes under dynamic environments. Linkage mapping has been widely used to identify quantitative trait loci (QTL) in many plant species but usually requires a time-consuming and labor-intensive fine-mapping process. Here, we developed two computational pipelines to identify candidate genes to improve important agricultural traits, such as height and biomass production. First, we developed QTG-Finder2, a machine learning-based algorithm to prioritize the causal genes in QTLs, and used orthologs of known causal genes as a training set. The model trained with orthologs could recall about 64% of Arabidopsis and 83% of rice causal genes when the top 20% ranked genes were considered, which is similar to the performance of models trained with known causal genes. Using QTG-Finder2, we trained and cross-validated *Sorghum bicolor* and *Setaria viridis* models. The sorghum model was validated by causal genes curated from the literature and could recall 70% of causal genes when the top 20% ranked genes were considered. We also applied the *Setaria* model and public transcriptome data to prioritize a plant height QTL and identified thirteen candidate genes. Second, we will present a computational framework of multiscale modeling to investigate how plants allocate metabolic resources for biomass production in response to drought. The framework is centered on a cell type-specific genome-scale metabolic network model of *S. bicolor* constrained by cell type-specific RNA-seq data. A C4 photosynthesis biochemical model was then integrated with the cell type-specific model to simulate dynamic environments by controlling carbon and energy sources of the metabolic network model. We collected a variety of data to inform the metabolic network model, such as photosynthesis data, biomass composition data, and RNA-seq data for Sorghum under well-watered and water-limiting conditions at multiple time points. Using the computational framework, we simulated plant growth and predicted that knocking out 23 genes and overexpressing 28 genes can improve biomass production. Finally, we have developed a pipeline to mine pan-genome information to identify potential causative polymorphisms in linkage with GWAS identified loci. We will cross-validate these independent pipelines computationally and test the prioritized candidate genes experimentally by generating and phenotyping knock-out lines.

Funding Statement: *This research was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program grant nos. DE- SC0018277 and DE-SC0008769.*