Genome-based Protein Function Discovery in the Eukaryotic Alga *Chromochloris zofingiensis*

Fatima Foflonker[1], Sean D. Gallaher[2], Sean McCorkle[1], Sabeeha Merchant[3] and **Crysten E. Blaby-Haas[1]*** (cblaby@bnl.gov)

[1]Biology Department, Brookhaven National Laboratory, Upton, NY; [2]Department of Chemistry and Biochemistry and Institute for Genomics and Proteomics, University of California, Los Angeles; [3]Quantitative Biosciences Institute, Department of Plant and Microbial Biology, University of California, Berkeley

Project Goals: Our overarching research goal is to design and engineer high-level production of biofuel precursors in photoautotrophic cells of the unicellular green *alga Chromochloris zofingiensis*. Our strategy involves large-scale multi-'omics systems analysis to understand the genomic basis for energy metabolism partitioning as a consequence of carbon source. Enabled by cutting-edge synthetic biology and genome-editing tools, we will integrate the systems data in a predictive model that will guide the redesign and engineering of metabolism in *C. zofingiensis.* Toward these objectives, we are implementing a phylogenomics-guided approach that leverages evolutionary relationships between genomes and between proteins encoded on those genomes for contextualized and evidence-based protein function discovery. For more information about the project and our team, please visit: https://sites.google.com/view/czofingiensis/home

The classic, but outdated, view of eukaryotic genomes is of gene islands randomly situated in a sea of non-coding DNA. This picture is derived from the observation that in contrast to prokaryotes, where functionally cooperating proteins are often encoded by operons, such structural organization does not appear to be necessary for co-regulating functional units in eukaryotes, in part because transcription and translation are uncoupled. However, as the number of sequenced eukaryotic genomes and transcriptomes has increased, and the function of those encoded proteins has been revealed, non-random gene organization, such as physical clustering of pathway members and co-regulated genes, has emerged as a characteristic of eukaryotic genomes. Current methods for identifying functionally cooperative gene neighborhoods in eukaryotes rely on the availability of functional annotations, which limits our ability to identify clustered functional gene units in algal genomes. Over half of algal proteins are of unknown function, while functionally annotated genes may be mis-annotated, inaccurate or vague, because of the evolutionary distance between algae and well-characterized model organisms, such as yeast and *E. coli*. A related challenge is the quality of structural annotations that are needed to predict coding regions and serve as the input for downstream comparative genomic analyses. While *C. zofingiensis* has a high quality, chromosome-complete genome assembly, the currently available structural gene annotations for the species are replete with observable mis-annotations and fragmented genes. In an effort to correct this shortcoming, we collaborated with the Joint Genome Institute to generate whole-molecule, long-read sequencing of the transcriptome (IsoSeq) on the PacBio Sequel platform. With a combination of computational analysis and manual curation, this IsoSeq data was used to inform the production of a more highly accurate set of gene annotations. Here, we present a greatly improved *C. zofingiensis* transcriptome and the identification of physical gene clusters in *C. zofingiensis*, using a method independent of functional annotation or coexpression data. We identified over 300 neighborhoods with potential functionally related neighbors including genes involved in carotenoid biosynthesis, photorespiration, nitrogen recycling, and oxidative stress responses. We also illustrate the value of conserved gene neighborhood identification for the discovery of gene function in algae with the discovery and testing of a novel arsenic detoxification pathway that evolved by co-opting glycolysis genes.

This research is supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), award number DE-SC0018301 and a FICUS JGI/EMSL project, award number 503551.