**Comparative Genomic and Transcriptomic Analyses for Pathway Discovery in *Chromochloris zofingiensis***

Fatima Foflonker* (foflonker@bnl.gov), **Crysten E. Blaby-Haas**

Brookhaven National Laboratory, Upton, NY

**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.**

The development of 'omics technologies has provided an unprecedented opportunity for the study of organisms as complex systems. With the genome-wide data these technologies provide, nearly any species can be fast-tracked to a level of understanding that was previously attainable for only a few "model organisms". The present challenge is shifting from acquiring genomic data, such as whole-genome sequences and transcriptomes, to using that knowledge to enable predictive biology and the rational redesign of biosystems. At the core of this challenge is the bottleneck in protein function understanding. Every genome-wide analysis, such as an RNA-Seq experiment, relies on functional annotations for translating data into information. As a result, the ability of an 'omics dataset to inform on a biological system is dependent on knowledge of that system's parts. However, for emerging organisms, nearly every functional annotation is a prediction that is typically based solely on sequence similarity to a database hit. Approaches are needed that place genomic resources in the hands of experimentalists allowing them to verify and test those function predictions in silico before taking a hypothesis to the bench.

With the successful implementation of an integrated systems biology approach to model, design and engineer high levels of biofuel precursors and value-added products, microalgae have the potential to become major sources of sustainable bioenergy and bioproducts. Toward this goal, over 100 algal whole-genome sequences are either presently available or are soon to be published. As fundamental resources, these data combined with thousands of published transcriptomes are precipitating a paradigm shift in the way we understand one of the most diverse, complex and understudied groups of photosynthetic eukaryotes. Remarkably, over half of the proteins encoded by algal genomes are of unknown function, highlighting both the volume of unique functional capabilities yet to be discovered and a fundamental knowledge gap that impedes successful biosystem design.

We are using a phylogenomic-based approach supported by a comprehensive, large-scale systems analyses for the discovery of novel, economically valuable, functional capabilities in the unicellular green alga *Chromochloris zofingiensis*. We are utilizing comparative genomics approaches to infer protein function from evidence-based associations. Ten chlorophyte algal genomes were used to find conserved gene neighborhoods, defined as: proximal orthologous genes within a 5 gene window, in a minimum of 4 species, and from at least two taxonomic classes to minimize effects of background synteny. This resulted in 183 conserved gene neighborhoods with predicted functionality in carotenoid biosynthesis, photorespiration, thiamine metabolism, nitrogen recycling, oxidative stress responses, and detoxification. Furthermore, relaxed constraints were used to capture proximal orthologous genes that co-occur in *C. zofingiensis* and at least one other algal genome. Gene fusions were identified by searching for domains that are encoded by two separate genes in at least two genomes and encoded by a single gene in at least two other genomes. These analyses were combined with a phylogenetic profile of orthologous proteins and co-expression analysis of condition-specific RNA-Seq data for the discovery and support of co-functional proteins and potential bioengineering targets for the production of value-added bioproducts and redesign of carbon and nutrient handling in *C. zofingiensis*.

*This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), award number DE-SC0018301.*