Phylogenomics-guided Approaches for Predicting and Characterizing Functions Across a Multi-functional Protein Family

Miriam Pasquini* (mpasquini@bnl.gov) and **Crysten E. Blaby-Haas**

Brookhaven National Laboratory, Upton, NY

Project Goals: QPSI (Quantitative Plant Science Initiative) is a DOE-mission driven, interdisciplinary, team-based capability that aims to accelerate the acquisition of a core knowledgebase of experimentally validated protein family function. Knowing what and how to engineer and predicting the consequence of that redesign on complex systems, such as a bioenergy crop, requires a systems biology approach grounded in molecular-level understanding. Because of the lack of sequence-based function understanding that affects all BER-relevant plant genomes, our goals are to: (1) reduce uncertainty in plant protein function annotation through integrated computational and experimental approaches, (2) identify genome-based principles underlying highly conserved processes over plant evolution, and (3) define fundamental rules of sequence-function relationships that translate across bioenergy crops. The resulting knowledge will underpin genome-based functional genomics experiments, while accelerating a sequence-based understanding of genotype-to-phenotype in specific plants.

Very few plant-specific proteins are characterized to the level required for inclusion in a redesign strategy. Indeed, most proteins in the plant lineage remain completely uncharacterized. This knowledge gap impedes accomplishing Biological and Environmental Research (BER) mission in enabling design and reengineering of plants for energy independence. However, it is impossible to experimentally characterize every protein: presently 99.8% of functional annotations across plant genomes are predictions. These predictions at the genome-wide scale are typically based on sequence similarity to a database "best hit", for example, by searching against NCBI's RefSeq database using BLAST or using profile hidden Markov models to search domains in Pfam. The annotation from the best-scoring hit is then transferred to the uncharacterized gene product. Unfortunately, in the case of BLAST-based annotations, the "best hit" was also likely annotated by a BLAST search, and the relationship to the original experimentally annotated protein is often lost, and distantly taxonomically related. In the case of model-based annotation, the domain annotation is typically derived from characterization of a single bacterial, yeast, or animal protein and/or mutant phenotype. The result is that over the last 20 years there has been significant error-prone propagation of functional annotations from one genome to another.

The functional annotation challenge is particularly acute for large proteins families that contain multiple paralogs and subfamilies that contain members from taxonomically divergent lineages. With recent advancements in genomics, post-genomic experimentation and high-throughput (HTP) experimental tools, we are able to address protein function in a critical way: integrating genomics, functional genomics, genetics, biochemistry and biophysical characterization. Using

this multi-disciplinary approach, we are addressing the challenge of accurately annotating large, functionally diverse protein families. We are employing large-scale phylogenomic analyses combined with conserved gene neighborhood detection, co-expression networks, co-occurrence profiles, and protein fusion discovery for function prediction. We then take our predictions to the bench for experimental characterization, which further guides evidence-based propagation of annotations across sequenced space. Here, we describe the phylogenomics-guided discovery and biochemical characterization of a protein family that we predict to be required for the maturation of metal-dependent proteins involved in processes that range from bacteriochlorophyll and cobalamin biosynthesis to post-translational modification of chloroplast-localized proteins. In addition, we will present our results toward the experimental verification of these functions, including meta-functional and sub-functional activity assays, and planned experimentation toward understanding the phenotypic manifestation of protein function in *Sorghum bicolor*.

This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER)