Single-Molecule, Whole-Transcript Sequencing of the Transcriptome of the Green Microalga *Chromochloris zofingiensis* to Accurately Annotate Gene Models and Identify Splice Variants

Sean D. Gallaher[*1] (gallaher@chem.ucla.edu), and **Sabeeha S. Merchant**[2]

[*] Presented By, [1] University of California, Los Angeles, CA, [2] University of California, Berkeley, CA

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.*

Despite having a high-quality genome for *C. zofingiensis*, the current gene annotations are riddled with errors, which is an impediment to our 'omics studies. Here, we have sequenced whole transcripts as single-molecule, long reads from a pool of *C. zofingiensis* mRNA. We are using this data to re-annotate the genome to correct errors, join fragmented genes, and identify splice variants. This should greatly benefit our efforts to conduct multi-'omics analyses in the species.

Chromochloris zofingiensis is a unicellular green alga that is of interest for its ability to produce high quantities of TAGs, as well as other high-value bio-products. In 2017, we published a high quality, chromosome-complete, assembly for *C. zofingiensis*' 58 Mbp genome [1]. As part of this work, we used the AUGUSTUS ab initio gene caller trained on a set of de novo assembled transcripts to annotate 15,274 nuclear genes. However, since that work we have identified a number of short-comings with those annotations. Many genes appear to be erroneously split into two, three, or more gene models. This has the consequence of making it difficult to accurately assign functional annotations to the fragmented genes, and complicates transcriptomic analysis. Many gene models appear to be wrong, specifically with regards to the intron/exon junctions. This leads to inaccurate predictions of the encoded polypeptide, which confounds proteomic analysis. In addition, the current annotations recognize only one splice variant per locus, despite evidence that *C. zofingiensis* utilizes alternative splicing. Lastly, the AUGUSTUS gene annotations lack UTRs. Collectively, these limitations and errors in the currently available AUGUSTUS gene models are obscuring the complete picture of *C. zofingiensis* gene expression that our research demands.

To advance our research in *C. zofingiensis*, we have used a relatively new method, called Iso-Seq, in which PacBio long reads are used to sequence whole transcripts as single molecules. This approach has the advantage over the Illumina short read sequencing that was done previously in that each sequencing read identifies a complete splice variant that is part of the transcriptome. When mapped back to the genome assembly, these reads help to mark the precise locations of intron/exon boundaries, and identify the UTRs. For *C. zofingiensis,* cDNA was made from a pool of RNA collected from cultures grown under a wide range of conditions (phototrophic growth, heterotrophic growth, nutrient-deprived, oxidative stress, etc.), in order to capture the widest possible range of transcripts. The resulting cDNA was ligated to adapters and sequenced on the PacBio Sequel platform. After being subjected to appropriate quality filters, the transcript sequences were re-mapped to the *C. zofingiensis* genome assembly, and compared with the AUGUSTUS gene models.

This analysis identified 26,529 unique transcript isoforms, which mapped to 11,305 genetic loci. The genes identified by Iso-Seq included 9,174 of the 15,274 nuclear genes identified by AUGUSTUS (~60%), as well as 2,131 novel genes. The remaining AUGUSTUS gene models were absent from the Iso-Seq dataset either because those transcripts were under-represented in the Iso-Seq library, or because those genes were originally misannotated by AUGUSTUS. For those genes identified by Iso-Seq, ~50% had two or more splice variants. Importantly, dozens of genes that we had identified as being fragmented in the AUGUSTUS gene models were correctly merged in the Iso-Seq data. While the Iso-Seq data alone is insufficient for complete annotation of all genes, careful merging of this new data with the previously available AUGUSTUS gene predictions and Illumina RNA-Seq data should allow us to produce a highly accurate and detailed picture of the *C. zofingiensis* transcriptome and proteome.

References

1. Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, et al. Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production. Proc Natl Acad Sci U S A. 2017;114. doi:10.1073/pnas.1619928114/-/DCSupplemental

Funding statement.

Funding was generously provided by the DOE / BER grant # DE-SC0018301 and by the Facilities Integrating Collaboration for User Science (FICUS) project of the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL).