**Final Scientific Report**
DOE Award: DE-SC0018301
Sponsoring Program Office: Office of Biological & Environmental Research
Recipient Organization: The Regents of University of California, 2150 Shattuck Avenue, Suite 313, Berkeley, CA 94704-5940, USA

**Project Title: Systems Analysis and Engineering of Biofuel Production in *Chromochloris zofingiensis*, an Emerging Model Green Alga**

**Team SWITCh** (Systems biology With Increased TAG in *Chromochloris*)
Krishna K. Niyogi, University of California, Berkeley (Principal Investigator)
Crysten Blaby-Haas, Lawrence Berkeley National Laboratory (Co-Investigator)
Nanette Boyle, Colorado School of Mines (Co-Investigator)
Mary Lipton, Pacific Northwest National Laboratory (Co-Investigator)
Sabeeha S. Merchant, University of California, Berkeley (Co-Investigator)
Trent Northen, Lawrence Berkeley National Laboratory (Co-Investigator)
Melissa S. Roth, University of California, Berkeley (Co-Investigator)

**Project Website**: https://sites.google.com/view/czofingiensis/
**Social Media**: https://twitter.com/zofingiensis
***Chromochloris* genome**: https://phycocosm.jgi.doe.gov/Chrzof1/Chrzof1.home.html

*Acknowledgment: This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-SC0018301.*

## Executive Summary

As a core component of a sustainable bioeconomy, microalgae have the potential to become a major source of biofuels and bioproducts. These photosynthetic microbes utilize solar energy, grow quickly, consume $CO_2$, and can be cultivated on non-arable land. However, there are presently considerable practical limitations in the photosynthetic production of biofuels from microalgae, resulting in low productivity and high costs. Algae are a strikingly diverse group and understudied algae can reveal new opportunities for biofuels. Integrative systems biology and engineering of emerging model systems are needed to expand the possibilities of microbial production of biofuels and bioproducts. Our long-term goal is to design and engineer high-level production of biofuel precursors in microalgae.

The green alga *Chromochloris zofingiensis* (Chlorophyceae) exhibits one of the highest levels of triacylglycerol (TAG) lipid accumulation among microalgae and is under development as a model alga, making it a promising feedstock for economically viable and sustainable production of algal biofuels. A major challenge to algal biofuels is that maximal lipid production typically occurs under nutrient stress, which severely restricts algal growth. However, *C. zofingiensis* turns on lipid production during a reversible switch between photoautotrophic growth (using $CO_2$ as a carbon source) and heterotrophic growth (using glucose). Upon glucose addition, this alga completely shuts off photosynthesis and dramatically increases the production of TAG, which it stores in cytoplasmic lipid bodies, while the cells continue to grow. Thus, the biology of *C. zofingiensis* makes it an ideal organism in which to apply systems analysis and synthetic biology approaches to understand and engineer regulatory switches between algal growth and lipid production.

In this project, we were successful in many areas that we proposed and also made unexpected new discoveries. The key findings from this research include a molecular and functional characterization of a glucose-dependent on/off switch for photosynthesis and redirection of metabolism, uncovering the conserved glycolytic enzyme hexokinase as a critical molecular player in these processes, discovering iron as a key nutrient to rescue photosynthesis and revealing unknown proteins involved in iron prioritization of TAG accumulation at the expense of chloroplast membranes, developing a rapid tool for draft genome-scale metabolic network reconstruction, a refined genome-scale metabolic and systems model, revealing that conserved eukaryotic gene neighborhoods can be used as a new tool for understanding gene function, and discovering widespread polycistronic expression in algae. We conducted cutting edge ‘omics experiments including transcriptomics, proteomics, metabolomics, and lipidomics, and our systems biology studies revealed targets for engineering to increase lipid production. We developed a reproducible transformation system for *C. zofingiensis* and identified another green alga *Auxenochlorella protothecoides* (Trebouxiophyceae) that has robust genetic engineering and a similar lipid switch to *C. zofingiensis.*

## Project Description

### Significance

**Need for sustainable energy sources**: The growing human population and an increase in the average standard of living across the globe continue to escalate the demand for energy. Meeting this demand is exacerbating environmental problems such as global climate change and will further increase our dependence on foreign sources, which threatens national security and undermines economic stability. As a core component of a sustainable bio-economy, microalgae have the potential to become a major source of biofuels and bioproducts. These photosynthetic microbes utilize solar energy, grow quickly, consume $CO_2$, and can be cultivated on non-arable land. However, there are presently considerable practical limitations in the photosynthetic production of biofuels from microalgae, resulting in low productivity and high costs. Integrative systems biology and engineering of emerging model systems will expand the possibilities of microbial production of biofuels and bioproducts. The unicellular green alga *Chromochloris zofingiensis* (formerly known as *Chlorella zofingiensis*) can produce large amounts of lipids and is under development as a model alga, making it a promising source of algal feedstocks for commercial production of biofuels.

**Under-investigated algae may reveal new biofuel opportunities**: Algae are a strikingly diverse group that derived from an endosymbiotic event in which a cyanobacterium was captured by a eukaryotic host >1 billion years ago. Additionally, secondary and tertiary endosymbiotic events involving capture of a primary photosynthetic endosymbiont by a eukaryotic host gave rise to ecologically important organisms like diatoms, haptophytes and dinoflagellates, which are also called algae. Green algae are at the base of green lineage (Viridiplantae), which also includes land plants. The biology of algae has attracted considerable attention in the last decade because the energy sector is looking to photosynthetic organisms as a route for liquid fuel production. Plant biomass production requires land, which presents an opportunity cost with respect to land use for food, housing, and environmental impact from conversion of forests into agricultural lands. Algae have potential as feedstocks for biodiesel production, especially if they can be grown in waste or seawater. In addition, their photosynthetic productivity can be higher than that of land plants. For more realistic exploitation of chlorophyte algae with respect to biofuels and bioproducts, it is critical that we understand the basic biochemistry, and cell and molecular biology of species beyond the laboratory references like *Chlamydomonas reinhardtii* (*Chlamydomonas* hereafter) and, more recently, *Ostreococcus spp*. It is also critical that we develop tools for manipulating the genomes of these organisms. Under-studied algae may reveal novel biology that can be used in generating a sustainable bioeconomy.

### Introduction to *Chromochloris zofingiensis*

*Chromochloris zofingiensis* is a unicellular chlorophyte alga in the Sphaeropleales order. This ~4 μm organism is a haploid coccoid alga, containing a single inter-connected chloroplast (~40% of the cell volume) and a tubular network of mitochondria (Figure 1). Its cell cycle is similar to that of *Chlamydomonas* where multiple fission events result in uniformly sized daughter cells. *C. zofingiensis* has exceptionally high photoprotective capacity compared to other algae and plants (Bonente et al. 2008), and the possibility exists to engineer the organism for higher photosynthetic productivity (Bonente et al. 2008, Kromdijk et al. 2016). Under specific conditions, *C. zofingiensis* can dramatically increase the production of triacylglcyerols (TAGs), the preferred lipid precursor for biofuel products, and it accumulates these TAGs to some of the highest levels analyzed (Breuer et al. 2012). *C. zofingiensis* is fast growing, can be cultured under many different conditions (including with wastewater), and reaches high culture densities ($10^8$ cells/ml). Thus, *C. zofingiensis* is presently considered one of the most promising biofuel feedstocks for commercial production.

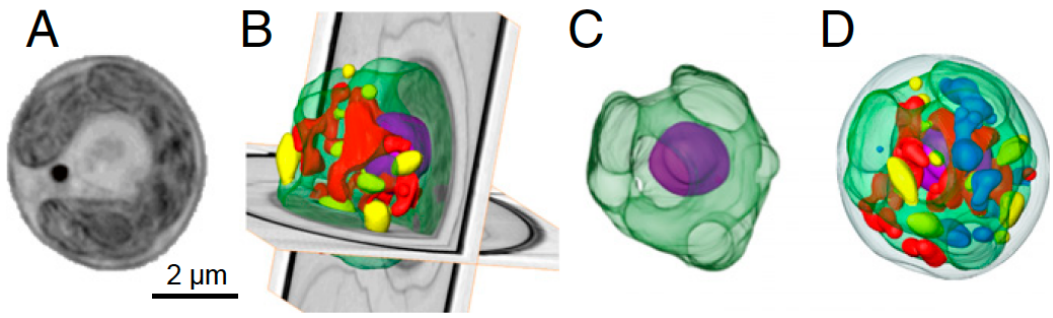


**Figure 1**. *Chromochloris zofingiensis* cell morphology. Cryo-soft X-ray tomography of a reconstructed cell with segmented nucleus (purple), chloroplast (green), mitochondria (red), lipids (yellow), and starch granules within the chloroplast (blue). (A) A representative orthoslice of the reconstructed cell. (B) 3-D segmentation over two orthogonal orthoslices. (C) Segmented chloroplast and nucleus. (D) Fully segmented cell. (Figure from Roth et al., 2017).

*C. zofingiensis* (SAG 211-14) has a high-quality, telomere-to-telomere chromosome nuclear genome assembly with centromeres and organelle genomes (Roth et al. 2017). *C. zofingiensis* has a ~58 Mb haploid nuclear genome distributed over 19 chromosomes with a quality in the tradition of model organism projects, as opposed to the fragmentary "gene space" assemblies typical of many genome sequencing projects. To document this statement, we compared the *C. zofingiensis* genome to those of two model organisms, *Chlamydomonas* and the plant *Arabidopsis*, and three other green chlorophyte genomes, *Coccomyxa subellipsoidea*, *Chlorella* sp. NC64A and *Monoraphidium neglectum* (Roth et al. 2017). *C. zofingiensis* has the most balanced G+C content (~51% nuclear, 53% coding) and low repeat content (~6%), similar to *C. subellipsoidea*. Gene density is quite uniform over chromosomes, not including the pericentromeric heterochromatin. These features are conducive to genome manipulation and engineering. The assembly, called ChrZofV5 (publicly accessible on our paper landing page and JGI Phytozome), successfully extended into telomere-associated repeats (~3.5 kbp on average) for 25 of 38 chromosome tips, and unplaced contigs appear to contain another 11 tips, leaving only 2 unaccounted. For most chromosomes exactly one region was identified as a putative (peri)centromeric locus with a rough estimate of only ~25 kb on average, which might enable the construction of artificial chromosomes.

In addition to the genome, we also completed a *C. zofingiensis* structural annotation based on collected RNA from cells grown under 14 diverse conditions (variation in photon flux density (PFD), nutrient limitation, oxidative stress) to capture a significant fraction of the cell's transcriptional repertoire, and we used paired-end sequencing of transcriptome libraries from rRNA-depleted total RNA to determine splice junctions, resolve paralogous families, and assemble transcripts from all three genomes *de novo* (Roth et al. 2017). The RNA-Seq coverage and *de novo* assembly in conjunction with AUGUSTUS (Stanke et al. 2006) as trained on the de novo transcriptome identified 15,274 nucleus-encoded protein- coding genes, of which 15,194 (99%) are apparently complete. Altogether, the genomes and annotation provide the blueprint for systems level analyses to improve understanding of photosynthetic and metabolic signaling and regulatory networks as well as the foundation for genetic engineering.

## Activities during this grant

### Overview

Our long-term goal is to design and engineer high-level production of TAGs in the understudied alga *C. zofingiensis*. Within the timeframe of this grant, we performed comprehensive, large-scale systems analyses (transcriptome, proteome, metabolome, physiology) of *C. zofingiensis* to understand how the energy metabolism of the cell is redirected based on energy/carbon source and identified potential engineering targets to improve biofuel prospects. We also determined the nutritional status necessary for *C. zofingiensis* to produce high amounts of TAG while continuing photosynthesis. We integrated the

systems data in a predictive genome scale metabolic model of *C. zofingiensis* that can guide us in redesigning and engineering its metabolism. Our experiments and analyses led to challenging long-held paradigms in eukaryotes, including the discoveries of widespread polycistronic gene expression in green algae and conserved eukaryotic gene neighborhoods to provide insight into function. As *C. zofingiensis* is an emerging model organism, we developed an RNA-Seq pipeline, improved the genome assembly and annotation, optimized a defined nutrient medium for high TAG accumulation, generated a proteome using cutting edge proteomics, conducted state of the art lipidomic and metabolomic analyses, established repeatable transformation, and developed broadly applicable computational genomics and synthetic biology tools. We also identified another green alga *Auxenochlorella protothecoides* (Trebouxiophyceae) that is a high producer of lipids and has efficient transformation via homologous recombination.

## Summary of key findings

### Large-scale biology of *C. zofingiensis* during trophic transitions

While many algae have flexible metabolism, *C. zofingiensis* is unusual in that it exhibits a reversible on-off switch of photosynthesis in response changing trophic conditions (Roth et al. 2019a, Figure 2). We conducted a molecular and physiological characterization of this photosynthetic and metabolic switch. We found that with glucose in the light, photosynthesis ceases, thylakoid membranes disappear, and cytoplasmic lipid droplets of TAG and astaxanthin accumulate. In just 3 days with glucose, there is a >20-fold increase in TAG and >3-fold increase in astaxanthin. Within 1 day after glucose removal, the photosynthetic apparatus re-assembles, and photosynthesis turns back on. Time-resolved transcriptomics during glucose addition and removal involve a third of all genes and reveal specific reversible patterns (Roth et al. 2019a). Functional enrichment analyses show downregulation of photosynthetic pathways and upregulation of ketocarotenoid synthesis and heterotrophic carbon metabolism. This study provides a rich large-scale dataset for gene discovery, in particular for those related to lipid biosynthesis.

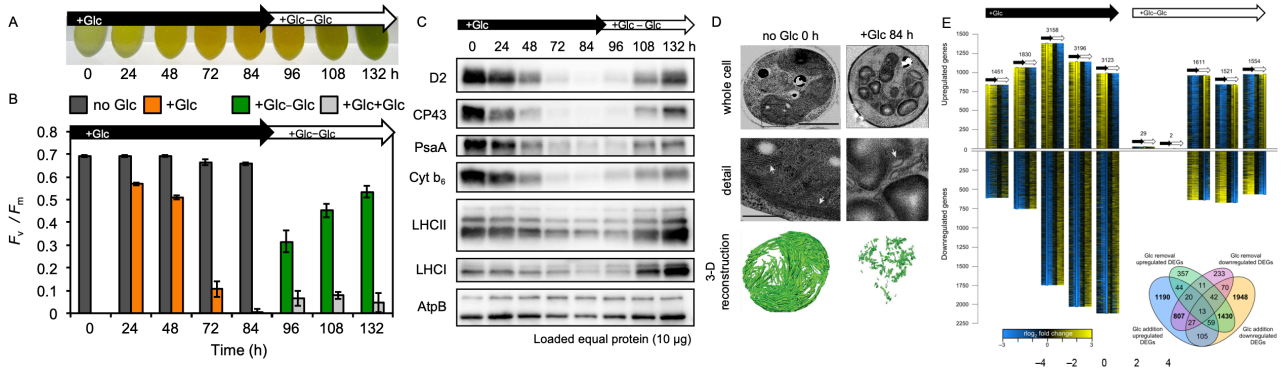


**Figure 2.** Reversible glucose-induced repression of photosynthesis and accumulation of TAG. (A) *C. zofingiensis* cultures. The orange color is due to astaxanthin. (B) Photosynthesis (PSII efficiency). (C) Immunoblots of the photosynthetic apparatus. (D) Transmission electron microscopy. (E) Genome-wide transcriptome responses. Figure adapted from Roth et al., 2019a.

### Hexokinase is necessary for glucose-dependent photosynthetic switch and lipid accumulation

As a simple photosynthetic system, *C. zofingiensis* is poised to bridge the conceptual gap between plants and yeast and could improve general understanding of the critical mechanisms underpinning sugar sensing and signaling in eukaryotes. Using forward genetics and whole-genome sequencing, we identified the widely conserved glycolytic enzyme hexokinase (HXK1) as a glucose sensor in algae and a master regulator of photosynthesis, carbon metabolism, and ketocarotenoid synthesis (Roth et al. 2019b). Glucose-treated *hxk1* mutants do not switch off photosynthesis or accumulate astaxanthin, TAG or

cytoplasmic lipid droplets but do increase respiration and cell size. *C. zofingiensis* has a single-copy gene encoding HXK, which contrasts with plants and yeast which have multiple copies. In plants, HXK1 is thought to have dual roles in metabolism and signaling, but the regulatory network in plants remains unknown. In yeast where mechanistic details are better understood, HXK2 shuttles in/out of the nucleus. While we showed that CzHXK1 is critical to the regulation of genes in those processes, we are currently investigating whether it is due to glucose-6-phophate, the metabolic product of HXK1, or a signaling role of the CzHXK1 protein itself.

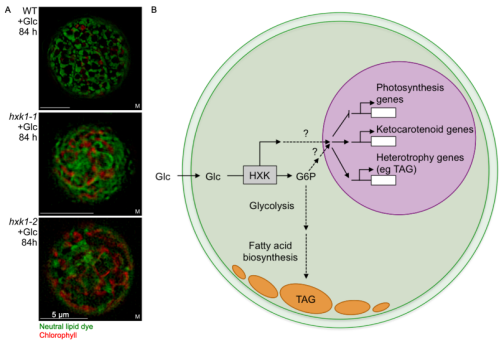


**Figure 3**. Hexokinase is necessary for glucose-mediated photosynthesis repression and lipid accumulation. (A) Live-cell super-resolution structured illumination microscopy. (B) Model of HXK-dependent glucose pathways. Figure adapted from Roth et al., 2019b.

**Iron rescues glucose-mediated repression of photosynthesis**

While testing different nutrients, we discovered that iron supplementation could rescue photosynthesis even in the presence of glucose (Jeffers et al., in revision). We generated a proteome of 12 experimental conditions: +/- Fe x +/- Glc in WT vs. two *hxk1* strains. A WT-Fe+Glc culture that specifically loses all photosynthetic activity was used as a reference point for gene discovery related to photosynthesis. We found that iron partitioning between lipid accumulation and electron transport underlies the loss of photosynthesis. Iron-rich co-factors were tracked in *de novo* fatty acid synthesis, respiration, and photosynthesis. Glucose increases lipid accumulation players despite iron status, while +Glc-Fe uniquely downregulates photosynthesis. Additionally, we identified hundreds of uncharacterized photosynthesis-related proteins conserved across green algae and plants. We developed an ortholog network approach to find proteins co-regulated with photosynthesis across *C. zofingiensis*, *Chlamydomonas*, *Arabidopsis*, rice and maize, which improves the prioritization of potential photosynthetic proteins meriting further molecular characterization. We developed a novel linear model classification pipeline, which revealed more proteins responding to the interaction of iron and glucose interactions than these inputs alone. The dominance of synergistic impacts on regulation highlights the importance of multifactorial environmental responses in the fitness and evolution of regulatory networks. We observed the complete fatty acid biosynthesis pathway was upregulated in WT+Glc independent of iron supply. In addition, players in TAG condensation, lipid droplet storage proteins, and fatty acid desaturase pathways had clear WT+Glc responses. All three players are known to contribute to TAG accumulation in other organisms, but based our proteomics indicates strong regulatory linkage in *C. zofingiensis,* implying regulatory-based bioengineering may improve lipid prospecting.

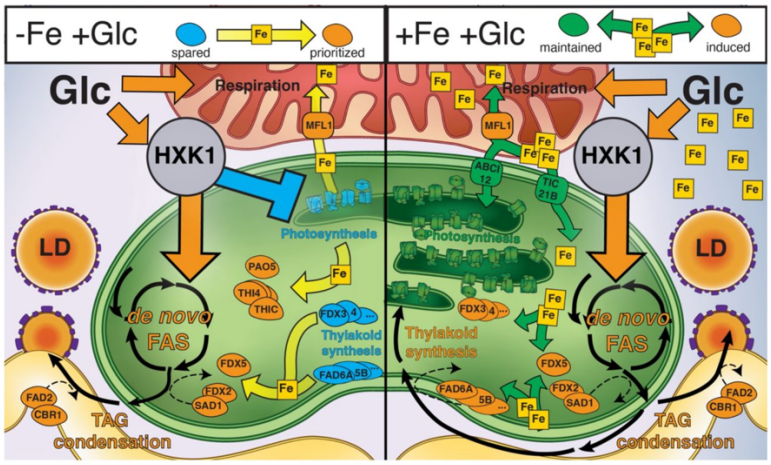


**Figure 4.** Model of intracellular iron partitioning strategies behind iron's role in heterotrophy and TAG accumulation. Figure from Jeffers et al., in revision.

**Automated draft genome-scale metabolic network reconstruction**
To enable more efficient construction of genome-scale metabolic network reconstruction for photosynthetic microorganisms, we developed an algorithm called Rapid Annotation of Photosynthetic Systems (RAPS), which can generate a high-quality draft network in only 20 min (Metcalf et al. 2020). This tool significantly out-performs other automated draft network algorithms for cyanobacteria/algae in terms of completeness of the network. We used RAPS to construct a draft metabolic network for *C. zofingiensis*. To test the completeness of our draft network, we have assumed the biomass composition of *Chlaymdomonas* and used flux balance analysis to predict fluxes for both photoautotrophic and photoheterotrophic growth (on glucose supplemented media) for photosynthetic microorganisms. This tool can be applied for different species, and we used RAPS to generate the first draft metabolic models for seven new species (Metcalf et al. 2020).

**Refined genome-scale metabolic and systems modeling**
A genome-scale metabolic network is a stoichiometric mathematical model of all (or most) enzyme-catalyzed chemical reactions in a cell, tissue, or organism. Once reconstructed, genome-scale networks can be useful in predicting nontrivial metabolic behaviors of organisms by using limited, easily measurable data such as nutrient uptake and product secretion rates. We generated a genome-scale metabolic model of *C. zofingiensis* and used $^{13}$C-MFA to quantify intracellular metabolic fluxes in different growth conditions (unpublished). The RAPS model served as a draft and was manually curated. Simulations conducted using the refined genome-scale metabolic model for *C. zofingiensis* predicted the excretion of lactate in heterotrophic cultures grown on glucose, and we quantified the amount of lactate excretion under different conditions. We have also performed $^{13}$C-MFA on cells grown in a variety of different conditions. Cells grown with glucose have significant overflow metabolism, excreting both lactate and succinate. We do not see this excretion of metabolites in the CORE medium. Interestingly, the additional iron in this medium allows the cell to fix carbon in the light because it enables the cell to maintain the photosynthetic apparatus.

**Eukaryotic polycistronic expression**
During analysis of the structural models, we discovered that polycistronic gene expression is pervasive in the *C. zofingiensis* transcriptome (Gallaher et al. 2021). We identified and confirmed at least 182 loci in which two or more protein-coding ORFs are cotranscribed on a single molecule of mRNA. For the ORFs for which peptides could be identified in the proteomic datasets, we identified N- and C-terminal peptides consistent with independent translation of each ORF, refuting the possibility that a larger polypeptide is post-translationally processed. Finally, we primed in vitro transcription and translation reactions with bicistronic mRNAs derived from *C. zofingiensis*, which resulted in pairs of proteins of the expected size. Some of the polycistronic loci identified in *C. zofingiensis* are conserved in other chlorophyte algae such as *Chlamydomonas* and *Dunaliella salina*, speaking to widespread occurrence of this phenomenon in the green algal lineage (Gallaher et al. 2021).

The discovery of polycistronic expression in *C. zofingiensis* prompted an exploration of the underlying mechanism. One possible mechanism, known as the leaky scanning model, posits that a ribosome may skip upstream AUG start codons at some frequency depending on the adjacent nucleotides, and continue scanning until encountering a more favorable downstream start codon. By aligning the -9 through +6 nucleotides of all genes, we identified the optimal sequence for translation initiation, known as the Kozak-like sequence. To test the leaky scanning hypothesis, we used an *in vitro* coupled transcription and translation system (in collaboration with the Theg laboratory) on constructs adapted from a bicistronic gene pair from *C. zofingiensis*. With the endogenous sequence, the ratio of the upstream to downstream gene product was 1:1 (Gallaher et al. 2021). This ratio increased with a stronger Kozak-like sequence (3.3:1) and decreased with a weak one (0.5:1), thus supporting the leaky scanning model. In addition, this result suggests a powerful mechanism for altering the ratio of gene products

expressed from bicistronic gene expression vectors. This work transforms the scientific beliefs on eukaryotic gene expression.

**Eukaryote gene neighborhoods provide functional information**

Although a popular method in relatively simple prokaryotes, scientists believed function prediction using gene proximity techniques was of little value in eukaryotes. This research reassessed this common assumption (Foflonker and Blaby-Haas 2021). Surprisingly, green algal genomes contain blocks of positionally orthologous genes, whose significance is underscored by their conservation for long periods—700 million years in some cases. Although the selective advantage of maintaining gene neighbors is presently unknown, this study found that these neighborhoods can be used to understand gene function in the same way gene proximity has been used to understand prokaryotic gene function. Integrating the positional analysis with a phylogenomics-based computational genomics approach led to the discovery of a novel arsenic detoxification pathway and a unique way that algae make sunscreen. This work provided both a new broadly applicable computational genomics tool and new biosynthetic routes for bio-based manufacturing of value-added algal products.

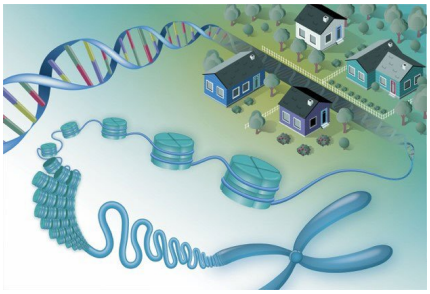

**Figure 5.** Conserved genomic neighborhoods as a new modality in the organization of genetic information.

**Organic nutrient utilization**

Because *C. zofingiensis* can take up many metabolites and carbon sources, it is important to investigate which nutrients yield efficient production of TAGs for biofuel production. In this project, we evaluated organic nutrient (carbon/nitrogen source) utilization order using a defined medium composed of more than 50 metabolites (unpublished). We found that glucose is the preferred carbon substrate and that arginine is the preferred metabolite taken up from the medium. We found additional sugars besides glucose can trigger shifts between photoautotrophic and heterotrophic metabolism. Surprisingly, we have also found that *C. zofingiensis* secreted deoxyadenosine and glutamine in the photoautotrophic state, then rapidly consumed them in the heterotrophic state. We investigated connections between *C. zofingiensis* metabolism and lipid composition. These results provide important insights into the biology and genomics of *C. zofingiensis*.

## Developing the *C. zofingiensis* toolkit

**Optimized growth media**

Because the nutritional status of an organism regulates its biosynthetic pathways and energy production, the media can play a critical role in both growth and accumulation of bioproducts. Furthermore, we found that during the trophic transition to produce high amounts of TAGs, there was also a large increase in biomass. To prevent nutrient stress and support fast growth and TAG accumulation, we defined an optimized medium. The process for optimization is iterative and hence time consuming, so as a starting point, the media chosen was developed from those used for culture of *Chlamydomonas* and *Dunaliella* species. We used inductively coupled mass-spectrometry of the *C. zofingiensis* ionome and spent media to assess the internal ratio of elements and nutrient demands of cells grown on glucose (this work is in preparation). In brief, *C. zofingiensis* was allowed to grow to stationary phase under photo-, hetero- or mixo-trophic conditions, following which the elemental composition of the cells was determined by ICP-MS/MS and related to biomass by measuring total organic carbon and nitrogen (TOC/N). Knowing the elemental composition per cell, we could then calculate what might be the minimal amount of macro- and micro-nutrients to provide in the growth medium to achieve $10^7$ or $10^8$ cells/ml. The media were also buffered to stabilize pH changes during growth. The optimized medium meets the demands of growth with glucose in a normal experimental time frame, while limiting confounding variables by ensuring the cellular ionome was consistent through log phase growth. Sulfur and phosphorus uptake is highly induced

by glucose. These elements likely contribute to *de novo* fatty acid accumulation in with glucose, so their levels in industrial cultures may be manipulated for cost-effective lipid bioprospecting.

The final optimized medium called *Chromochloris zofingiensis* Optimized Ratio of Elements (CORE) was used by Team SWITCh in experiments on metabolic transitions with different iron levels, master regulator TOR kinase, metabolic models, and comparative transcriptomics and proteomics of nutrient replete vs. nutrient deficient cells focusing on the trace elements in both phototrophic and mixotrophic cells with the objective of identifying regulatable promoters that could be used in synthetic biology applications. Genes responding specifically to Fe, Zn and Cu deficiencies were identified. The gene inventory is useful for a) understanding acclimation mechanisms that enable productivity despite nutrient limitations, b) comparative analysis in reference to such mechanisms in diverse green algae, and c) distinguishing metabolic demands in phototrophic vs. heterotrophic cells. Altogether, the transcriptome datasets are also a resource for improving structural annotations of genes that are normally not expressed in the nutrient replete state, and for building metabolic networks.

**RNA-Seq pipeline**

To understand how *C. zofingiensis* acclimates to light and nutrient availability and to distinguish transcription regulatory circuits with the goal of identifying hubs and networks, several time course experiments were designed. We first developed a pipeline to generate high-quality RNA-Seq data (Gallaher and Roth 2018). RNAs were sequenced on the Illumina platform after preparing libraries in our own laboratory or at JGI. The output sequence data were aligned to the *C. zofingiensis* genome for estimating transcript abundances. RNA-Seq experiments during transitions to high TAG accumulation were analyzed to reveal reversible changes in the transcriptome that form the basis of metabolic regulation during trophic transitions (Roth et al. 2019a). Functional enrichment analyses showed that Glc represses photosynthetic pathways while ketocarotenoid biosynthesis and heterotrophic carbon metabolism are upregulated. Additionally, this pipeline has been used in micronutrient and TOR kinase transcriptome experiments (unpublished). These 'omics data helped us identify two phospholipases and one glycolipid lipase that can potentially breakdown both the membrane and plastid lipids respectively. Therefore, upregulation of these lipases can act as a Pull in the Push-Pull-Protect strategy of lipid accumulation. Additionally, we identified a key MYB TF involved in upregulation of *de novo* lipid accumulation genes. These genes will serve as candidates for engineering to improve lipid production.

**Improved genome assembly**

For systems analysis of metabolic acclimation using transcriptomics and proteomics, we need the most accurate gene models for the experimental system. RNA from various experiments was pooled and sequenced at JGI on the PacBio platform to identify full-length processed transcripts and isoforms (Iso-Seq). In parallel, gene prediction tools like Braker were used in conjunction with known and validated gene models to deduce structural models for genes where we did not have transcript data.

Another goal of this project was to annotate transcription start sites using ChIP-Seq. Individual histone modifications are associated with different parts of a gene, such as the promoter vs. the gene body, and these can be identified at a whole-genome level by ChIP-Seq using specific antibodies. We took advantage of our knowledge of *Chlamydomonas* chromatin to develop the methodology for *C. zofingiensis* (Strenkert et al. 2022). We used unmodified histone H3 for methods development and a modified histone H3 to mark gene promoters. Histone H3 is largely conserved between algae and humans, so we tested commercial antibodies, which worked well in immunoblotting. After confirming the efficacy of the protocol by PCR, we completed the ChIP-Seq analysis to show that H3K4me3 predominantly maps to promoter regions of genes in *C. zofingiensis*, but enrichment does not necessarily correlate with the magnitude of expression of the underlying genes. The data were subsequently used to further improve gene structural annotation based on the promoter marks.

The resulting annotations are a major improvement both with respect to inclusion of previously missed genes and with respect to accuracy of included exons. These data are a major resource for capturing proteomic data or building metabolic models with this species. For example, in when we re-

aligned the peptides from proteomics experiments to the newer gene models, a larger fraction of the proteome could be quantified (increase from 65 to 80%) and more peptides were aligned per gene model. The high-quality assembly with good structural annotation is also a resource for comparative genomics. We also quantified the various mRNA species (isoforms) from each gene and ranked them based on abundance. The ChIP-Seq experiments allowed us to distinguish and validate the transcript models. A final set of high-quality gene annotations was produced and is now available on Phytozome and Phycocosm through JGI.

**Identification of transcription factors**

Transcription factors (TFs) are important proteins that control the rate of transcription of genetic information by binding to specific DNA sequences to regulate gene expression. Additionally, TFs can regulate multiple genes such as those in the same pathway. We developed a tool to identify TFs within any novel genome (unpublished). Using PFAM and InterProScan IDs, we made a database of different TF families from PlantTFDB and from domains literature that have DNA-binding domains then filtered all the known and unknown TFs using overlapping DNA binding domains and families from within the genome. Using this tool, we identified 340 different TFs in the *C. zofingiensis* genome. This tool can be applied to different organisms.

**Genetic transformation**

Genetic engineering is a powerful approach to modify an existing organism or adding in a synthetic pathway to another organism for sustainable biofuel production. When we began this project, we hoped to optimize an existing method for nuclear transformation (Liu et al. 2014). However, we have been unable to reproduce their method, and the authors have not published any follow-up studies. Despite *C. zofingiensis* proving to be a recalcitrant organism to genetically transform, we have now established reliable yet low-efficiency transformation. We tested many parameters including cell wall digestion, selectable markers, promoters, and different methods of DNA delivery. Currently, we have two selectable markers working and will continue to optimize efficiency and expand its genetic repertoire (unpublished).

**State of the art proteomics**

Proteomics involves the extraction and digestion of proteins from any sample to tryptic peptides. The peptides are then analyzed by liquid chromatography mass spectrometry (LC-MS/MS) where the peptide mass is measured and then the peptide is fragmented in the second MS step. The peptide identification results from the comparison of the measured fragmentation pattern to a sequenced genome using an algorithm such as MSGFplus (Kim and Pevzner 2014).

The first important question in our proteomic characterization is how detailed a protein profile can we obtain from a single analysis (using liquid chromatography mass spectrometry (LC-MS/MS) on a sample as opposed to separating the sample into multiple fractions and analyzing each by LC-MS/MS separately. We determined that a single run yielded 29,394 peptides corresponding to 4,386 protein identifications (determined by at least 2 peptides per protein), while fractionating the same sets of samples using high pH separations (Yang et al. 2012) into 12 fractions each yielded 222,431 total peptides corresponding to 10,462 proteins. This increase in both number of proteins and an average number of peptides per protein illustrated the need for fractionation in our experimental design.

There are a number of limitations in using fractionation in proteomics experiments. First, the number analyses increase by at least an order of magnitude for every experiment. Second, because peptides can be included in multiple fractions, peptide abundances as measured by the ion intensity in the MS is the result of an average of many fractions and lowers accuracy. Third, this inaccuracy in the peptide concentration hampers the comparison of relative changes in peptide and protein abundance. For these reasons, as well as the need for fractionation to increase protein coverage, we decided to use the Tandem Mass Tag (TMT) approach to quantification of peptide abundances (O'Connell et al. 2018).

The TMT approach allows for a multiplexed analysis of many different samples to efficiently use instrument time and increase protein coverage through fractionation. The approach uses isobaric tags as

markers for the peptides in each sample which are then measured in the MS step. Briefly, cells are lysed and proteins digested to peptides. The TMT labels are added to individual samples, where at least one label is reserved for a pooled sample to serve as a reference.

Once the peptides have been labeled, they are mixed, and the peptide mixture is fractionated using high pH fractionation and analyzed by LC-MS/MS. In the MS process, the isobaric label is cleaved off and analyzed at a low molecular weight range of the analysis. The peptide is fragmented in the mass spectrometer and the resulting fragmentation pattern is used for peptide identification. Since the same peptide from all the samples elutes in the same fraction and at the same time from the LC part of the LC-MS/MS, the isobaric tags are used for quantitation purposes. The resulting output looks like figure Y where the colored lines are the peaks from the isobaric tags. The peptide abundances can be compared directly in the experiment through the abundances of each of these tags. The TMT method effectively allows for the direct comparison of the peptide abundance as well as the use of fractionation while still reducing instrumentation demand.

Protein post-translational modifications are an important regulation step in biomolecular and metabolic pathways. Since the time frame for modifying a protein is on the order of seconds and protein synthesis is on the order of minutes to hours, biological systems have used protein post-translational modifications as a way to fine tune metabolic activities. Proteomics is one of the only methods for determining the state of protein post-translational modification since it is the only omics method to directly examine the proteins. However, searching for modifications of peptides in the global tryptic digests is difficult since the abundance of modified peptides are usually much lower than those of unmodified peptides and their signals become suppressed. For that reason, we have employed a number of enrichment methods for the modifications that we are targeting.

Phosphorylation is the most common protein post-translational modification since it plays a crucial role in both metabolism and regulation. For that reason, there are well established strategies for the enrichment of phospho-peptides. In these experiments, the peptides are applied to an Immobilized Metal Affinity Column (IMAC) where the metal chelates with the phospho group, and the phosphorylated peptides are retained on the column. They are eluted with a salt buffer and analyzed by mass spectrometry in the same manner as described above for the global peptides.

While phospho-peptides are the most common target for the investigation of protein post-translational modifications, there are other types of modifications that we have also investigated with the *C. zofingiensis* samples. Using samples that have been grown up on different substrates, we applied different enrichment methods for each of the samples to determine which proteins are labeled.

Ubiquitination is the biochemical process in which proteins are marked by ubiquitin, a 76 amino acid protein. Ubiquitin is a conserved polypeptide unit that plays an important role in the ubiquitin-proteasome pathway. Ubiquitin can be covalently linked to many cellular proteins by the ubiquitination process, which targets proteins for degradation by the 26S proteasome. Three components are involved in the target protein-ubiquitin conjugation process. Ubiquitin is first activated by forming a thio-ester complex with the activation component E1; the activated ubiquitin is subsequently transferred to the ubiquitin-carrier protein E2, then from E2 to the E3 ubiquitin ligase for final delivery to the epsilon-$NH_2$ of the target protein lysine residue. Ubiquitination occurs intracellularly in eukaryotes and regulates a wide variety of biological processes. This type of modification targets proteins to the proteasome and is key in many cellular signaling pathways. For example, protein processing requires ubiquitination, this process plays an important role in regulating apoptosis, and membrane trafficking also requires mono-ubiquitination. The enrichment method for this protein post-translational modification uses a commercially available kit from Cell Signaling Technology which employs a proprietary methodology for peptide enrichment based on immunoprecipitation using a specific bead-conjugated antibody. Simply, the peptide digest is applied to a column containing the antibody where the ubiquitin binds selectively to the antibody. A buffer is applied to the column rinsing all non-modified peptides out leaving the modified peptides bound. A salt buffer is then applied to disrupt the antibody binding thus eluting the modified peptides for analysis by LC-MS as described above.

Protein acetylation is another very common protein modification. Here, the acetyl group is co- or post-translationally attached either to the lysine residues at ε-amino group or to the N-terminus of protein at α-amino group. This modification can change the function of a protein through alteration of its properties, including hydrophobicity, solubility, and surface properties, all of which may influence protein conformation and interactions with substrates, cofactors and other macromolecules. Most central metabolic enzymes possess lysines that are acetylated in a regulated fashion, and many of these regulated sites are conserved across the spectrum of bacterial phylogeny. The interconnectedness of acetylation and central metabolism suggests that acetylation may be a response to nutrient availability or the energy status of the cell. Similar to the ubiquitin modification enrichment, the commercially available kit from Cell Signaling Technology targets the acetyl group using based on immunoprecipitation using a specific bead-conjugated antibody. The acetylated peptides are enriched using the selective binding of the solid phase resin and eluted with a salt buffer for analysis by LC-MS.

Proteomics data were used in Gallaher et al. (2021) and Jeffers et al. (in revision), and both proteomics and phosphoproteomics data are used in understanding metabolic transitions with different iron levels and regulation of the master regulator TOR kinase (unpublished). Additionally, a PTM-focused study on phosphorylation, acetylation and ubiquitination has been conducted on TAG accumulation during glucose addition (unpublished).

## Outlook

This grant enabled a significant amount of progress toward our long-term goal to design and engineer high-level production of TAGs in green algae (see Grant Products). We completed systems analysis of *C. zofingiensis* and identified potential targets that will enhance lipid production. Additionally, we advanced the *C. zofingiensis* toolkit, which enables greater understanding of its lipid biosynthesis and can be applied for biofuel production. Although the project is closed, data analysis is continuing and additional manuscripts on *C. zofingiensis* will be published. Moreover, this grant has led us to the genetically tractable green alga *Auxenochlorella protothecoides* that has a similar but different switch to TAG accumulation and can also serve as a genetic testbed for targets identified by the systems biology work in *C. zofingiensis*. *A. protothecoides* is the focus of our current DOE BER grant and encompasses our entire Team SWITCh with the addition of two colleagues, Jeffrey Moseley (UC Berkeley) and Setsuko Wakao (Lawrence Berkeley National Laboratory).

## **Grant Products** (September 2017-January 2024)

### Journal Articles


Iwai, M, Roth, MS, Niyogi, KK. 2018. Subdiffraction-resolution live-cell imaging for visualizing thylakoid membranes. *The Plant Journal 96*, 233-243. DOI: 10.1111/tpj.14021

Gallaher, SD, Roth, MS. 2018. RNA purification from the unicellular green alga, *Chromochloris zofingiensis*. *Bio-protocol 8*, e2793. DOI: 10.21769/BioProtoc.2793

Roth, MS, Gallaher, SD, Westcott, DJ, Iwai, M, Louie, KB, Mueller, M, Walter, A, Foflonker, F, Bowen, BP, Ataii, NN, Song, J, Chen, J-H, Blaby-Haas, C, Larabell, C, Auer, M, Northen, T, Merchant, SS, Niyogi, KK. 2019a. Regulation of oxygenic photosynthesis during trophic transitions in the green alga *Chromochloris zofingiensis*. *The Plant Cell* 31: 579-601. DOI: 10.1105/tpc.18.00742

Blaby-Haas, CE, Merchant, SS, 2019. Comparative and Functional Algal Genomics. *Annual Review of Plant Biology*, *70*, pp.605-638. DOI: 10.1146/annurev-arplant-050718-095841

Roth, MS, Westcott, DJ, Iwai, M, Niyogi, KK. 2019b. Hexokinase is necessary for glucose-mediated photosynthesis repression and lipid accumulation in a green alga. *Communications Biology* 2: 347. DOI: 10.1038/s42003-019-0577-1

Sake, CL, Metcalf, AJ, Boyle, NR. 2019. The challenge and potential of photosynthesis: unique considerations for metabolic flux measurements in photosynthetic microorganisms. *Biotechnology Letters* **41**, 35–45. DOI: 10.1007/s10529-018-2622-4

Metcalf, AJ, Nagygyor, A, Boyle, NR. 2020. Rapid Annotation of Photosynthetic Systems (RAPS): Automated algorithm to generate genome-scale metabolic networks from algal genomes. *Algal Research*, 50, 101967. DOI: 10.1016/j.algal.2020.101967

Foflonker, F, Blaby-Haas, CE, 2020. Colocality to cofunctionality: Eukaryotic gene neighborhoods as a resource for function discovery. *Molecular Biology and Evolution*, 38 (2), pp.650-662. DOI: 10.1093/molbev/msaa221

Gallaher, SD, Craig, RJ, Ganesan, I, Purvine, SO, McCorkle, SR, Grimwood, J, Strenkert, D, Davidi, L, Roth, MS, Jeffers, TL, Lipton, MS, Niyogi, KK, Schmutz, J, Theg, SM, Blaby-Haas, CE, Merchant, SS. 2021. Widespread polycistronic gene expression in green algae. *Proceedings of the National Academy of Sciences USA* 118 (7): e2017714118. DOI: 10.1073/pnas.2017714118

Strenkert, D, Mingay, M, Schmollinger, S, Chen, C, O'Malley, RC, Merchant, SS. (2022). An optimized ChIP-Seq framework for profiling histone modifications in *Chromochloris zofingiensis*. *Plant Direct*, 6(3), e392. DOI: 10.1002/pld3.392

Schwechheimer, C, Schröder, PM, Blaby-Haas, CE. Plant GATA factors: their biology, phylogeny, and phylogenomics. *Annual Review of Plant Biology* 73 (2022): 123-148. DOI: 10.1146/annurev-arplant-072221-092913

Öztürk Y, Andrei A, Blaby-Haas CE, Daum N, Daldal F, Koch HG. Metabolic sensing of extracytoplasmic copper availability via translational control by a nascent exported protein. *mBio*. 2023 Feb 28;14(1):e0304022. DOI: 10.1128/mbio.03040-22.

Meagher, M, Tamburro, J, Boyle, NR. Accepted. Development of a 3D printed paddlewheel to improve algal growth in a flat panel photobioreactor. *Biotechnology Progress*


### Preprints


Jeffers, TL, Purvine, SO, Nicora, CD, McCombs, RM, Upadhyaya, S, Stroumza, A, Whang, K, Gallaher, SD, Dohnalkova, A, Merchant, SS, Lipton, MS, Niyogi, KK, Roth, MS. Iron rescues glucose-

mediated photosynthesis repression during lipid accumulation in the green alga *Chromochloris zofingiensis*. *bioRxiv* DOI:10.1101/2023.07.31.551119

Meagher, M, Metcalf, AJ, Ramsey, SA, Prentice, W, Boyle, NR. Genome-scale metabolic model accurately predicts fermentation of glucose by *Chromochloris zofingiensis. bioRxiv* DOI: 10.1101/2021.06.22.449518

**Book Chapters**

Jeffers, TL, Roth, MS. 2021. Revealing mechanisms of algal astaxanthin production and bioengineering potential using multi-omics. *In* Ravishankar, GA and Ambati, RR (Eds.). *Global perspective of astaxanthin: From industrial production to food, health and pharmaceutical applications*. Academic Press, USA. Pp 181-208. DOI: 10.1016/B978-0-12-823304-7.00010-6

Blaby-Haas, Crysten E. The Mosaic Landscape of Algal Metal Transport and Usage. *Microbial Metabolism of Metals and Metalloids*. Cham: Springer International Publishing, 2022. 55-75. DOI: 10.1007/978-3-030-97185-4

**Patents**

Roth, MS, Niyogi, KK, Gallaher, SD, Merchant, SS. 2021. Genes involved in astaxanthin biosynthesis (US Patent Number 11,124,819). *United States Patent and Trademark Office* URL

Merchant, SS, Gallaher, SD, Moseley, JL, 2023. Expressing multiple genes from a single transcript in algae and plants (US Patent Application Number US-20230110869-A1). *United States Patent and Trademark Office* URL

## Cited References


Bonente, G, Passarini, F, Cazzaniga, S, Mancone, C, Buia, MC, Tripodi, M, Bassi, R, Caffarri, S. 2008. The occurrence of the *psbS* gene product in *Chlamydomonas reinhardtii* and in other photosynthetic organisms and its correlation with energy quenching. *Photochemistry and Photobiology* 84(6):1359-1370.

Breuer, G, Lamers, PP, Martens, DE, Draaisma, RB, Wijffels, RH. 2012. The impact of nitrogen starvation on the dynamics of triacylglycerol accumulation in nine microalgae strains. *Bioresource Technology* 124:217-226. 861.

Foflonker, F, Blaby-Haas, CE, 2020. Colocality to cofunctionality: Eukaryotic gene neighborhoods as a resource for function discovery. *Molecular Biology and Evolution*, 38 (2), pp.650-662. DOI: 10.1093/molbev/msaa221

Gallaher, SD, Craig, RJ, Ganesan, I, Purvine, SO, McCorkle, SR, Grimwood, J, Strenkert, D, Davidi, L, Roth, MS, Jeffers, TL, Lipton, MS, Niyogi, KK, Schmutz, J, Theg, SM, Blaby-Haas, CE, Merchant, SS. 2021. Widespread polycistronic gene expression in green algae. *Proceedings of the National Academy of Sciences USA* 118 (7): e2017714118. DOI: 10.1073/pnas.2017714118

Gallaher, SD, Roth, MS. 2018. RNA purification from the unicellular green alga, *Chromochloris zofingiensis*. *Bio-protocol 8*, e2793. DOI: 10.21769/BioProtoc.2793

Jeffers, TL, Purvine, SO, Nicora, CD, McCombs, RM, Upadhyaya, S, Stroumza, A, Whang, K, Gallaher, SD, Dohnalkova, A, Merchant, SS, Lipton, MS, Niyogi, KK, Roth, MS. In revision. Iron rescues glucose-mediated photosynthesis repression during lipid accumulation in the green alga *Chromochloris zofingiensis*. *Nature Communications* (preprint DOI:10.1101/2023.07.31.551119)

Kim, S, Pevzner, PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* 5:5277.

Kromdijk, J, Glowacka, K, Leonelli, L, Gabilly, ST, Iwai, M, Niyogi, KK, Long, SP. 2016. Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science* 354(6314):857-861.

Liu, J, Sun, Z, Gerken, H, Huang, J, Jiang, Y, Chen, F. 2014. Genetic engineering of the green alga *Chlorella zofingiensis*: a modified norflurazon-resistant phytoene desaturase gene as a dominant selectable marker. *Applied Microbiology and Biotechnology* 98(11):5069-5079.

Metcalf, AJ, Nagygyor, A, Boyle, NR. 2020. Rapid Annotation of Photosynthetic Systems (RAPS): Automated algorithm to generate genome-scale metabolic networks from algal genomes. *Algal Research*, 50, 101967. DOI: 10.1016/j.algal.2020.101967

O'Connell JD, Paulo JA, O'Brien JJ, Gygi SP. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J Proteome Res*. 2018;17(5):1934-1942.

Roth, MS, Cokus, SC, Gallaher, SD, Lopez, D, Walter, A, Erickson, E, Endelman, B, Westcott, D, Larabell, C, Merchant, SS, Pellegrini, M, Niyogi, KK. 2017. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. *Proceedings of the National Academy of Sciences USA* 114 (21): E4296–E4305. DOI: 10.1073/pnas.1619928114

Roth, MS, Gallaher, SD, Westcott, DJ, Iwai, M, Louie, KB, Mueller, M, Walter, A, Foflonker, F, Bowen, BP, Ataii, NN, Song, J, Chen, J-H, Blaby-Haas, C, Larabell, C, Auer, M, Northen, T, Merchant, SS, Niyogi, KK. 2019a. Regulation of oxygenic photosynthesis during trophic transitions in the green alga *Chromochloris zofingiensis*. *The Plant Cell* 31: 579-601. DOI: 10.1105/tpc.18.00742

Roth, MS, Westcott, DJ, Iwai, M, Niyogi, KK. 2019b. Hexokinase is necessary for glucose-mediated photosynthesis repression and lipid accumulation in a green alga. *Communications Biology* 2: 347. DOI: 10.1038/s42003-019-0577-1

Stanke, M, Schoffmann, O, Morgenstern, B, Waack, S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62.

Strenkert, D, Mingay, M, Schmollinger, S, Chen, C, O'Malley, RC, Merchant, SS. (2022). An optimized ChIP-Seq framework for profiling histone modifications in *Chromochloris zofingiensis*. *Plant Direct*, 6(3), e392. DOI: 10.1002/pld3.392

Yang, F, Shen, Y, Camp, DG 2nd, Smith, RD. 2012. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. Expert Review of Proteomics 9(2):129-34.