FINAL TECHNICAL REPORT

Project Title:	Transforming our understanding of chloroplast-associated genes through comprehensive characterization of protein localizations and protein-protein interactions
Recipient Organization:	Princeton University
Date of Report:	11/1/2023
Award Number:	DE-SC0020195
Total Project Costs:	$2,974,667
Program Manager:	Pablo Rabinowicz, pablo.rabinowicz@science.doe.gov
Principal Investigator:	Martin Jonikas, Princeton University, mjonikas@princeton.edu
Co-Investigators:	Danny Schnell, Michigan State University, deceased Jianping Hu, Michigan State University, huji@msu.edu

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-SC0020195.

Disclaimer:

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Executive Summary

Bioenergy crops are a renewable source of fuels and are a critical base for building a carbon-neutral economy. Rational engineering of bioenergy crops has the potential to enhance the yields. However, our ability to engineer plants is limited because the functions of most genes remain unknown. Systematic characterization of gene function in plants thus has the potential to greatly accelerate bioenergy research.

Here, we focus on the chloroplast, an underexplored energy-producing organelle that is a hallmark of plants. The chloroplast is one of the promising targets of biofuel crop engineering efforts because of its central role in photosynthesis, metabolism, and intracellular signaling. However, the protein composition of the chloroplast and the functions of most of its proteins remain poorly characterized.

At the core of this project, we sought to comprehensively determine the localization of chloroplast-associated proteins and generate a spatially defined protein-protein interaction network for chloroplast. For this purpose, we used the leading model alga *Chlamydomonas reinhardtii*, which greatly increased experimental speed and throughput. We illustrated the value of our findings to land plants by determining the localization of *Arabidopsis thaliana* land plant homologs of the *Chlamydomonas* proteins.

Altogether, we were successful in determining the localization of 1,034 chloroplast-associated proteins in *Chlamydomonas*. The localizations provide numerous insights into the spatial organization of chloroplasts and how they function to support photosynthesis. The localization patterns of distinct proteins revealed new chloroplast structures and revealed new spatial organization inside the chloroplast. We also identified new components of known chloroplast structures, such as the chloroplast envelope, nucleoid, plastoglobuli, and pyrenoid. We identified these new components by investigating the interacting partners of known proteins. Many proteins localized in both the chloroplast and other cellular structures, thereby hinting at new functions and communication between cellular structures. We also applied machine learning on the atlas to generate predictions for the location of all of the proteins in Chlamydomonas. This enabled us to assign putative functions to many uncharacterized proteins based on their cellular location. Altogether, this research establishes a rich resource that opens new avenues of investigation and guides future work in deciphering and manipulating chloroplast function.

Next, we developed an extensive protein-protein interaction network for the chloroplast by performing affinity purification-mass spectrometry on ~1,150 tagged chloroplast-associated proteins, the first such large-scale study in any photosynthetic organism. This dataset reveals 4,694 high-confidence protein-protein interactions, offering insights into the functions of thousands of conserved poorly-characterized chloroplast proteins. This systematic identification of protein-protein interactions in the

chloroplast also provides multiple exciting new research directions and a detailed blueprint of the chloroplast's operation.

This research lays the groundwork to decipher the inner workings of the chloroplast, the cell structure at the heart of photosynthesis. The spatial atlas and protein-protein interactions reveal chloroplast organizational features that would not have been accessible with traditional approaches. The localization mapping, insights into the function, and research materials generated further provide a rich resource for the research community to advance the understanding of how the chloroplast is organized to enable engineering of enhanced photosynthetic organisms.

Project Accomplishments

1) We determined the localizations of 1,034 chloroplast-associated proteins.

a. We cloned 3,116 genes using our high-throughput cloning pipeline.
b. Using a high-throughput imaging strategy, we successfully localized 1,034 proteins, with 589 proteins localized in the chloroplast, 618 proteins conserved in land plant *Arabidopsis thaliana*, and 702 poorly-characterized proteins.
c. Using two-color microscopy, we identified novel components forming chloroplast punctate structures, such as nucleoid and plastoglobuli.
d. We built a new protein localization predictor using machine learning, mainly based on the training of the proteins localized in this project. By applying the new predictor to the whole proteome of *Chlamydomonas*, we identified 2,245 putative chloroplast proteins.
e. We illustrated the value of the *Chlamydomonas* localizations to land plants by determining the localizations of *Arabidopsis thaliana* land plant protein homologs.
f. We published our findings.
g. We made the localizations available to the research community via a website.
h. We made the strains and plasmids available to the community via the Chlamydomonas Resource Center.

2) We generated a spatially-defined protein-protein interaction network for the chloroplast.

a. We performed immunoprecipitations and mass spectrometry on ~1,150 epitope-tagged lines, more than we aimed for initially. This dataset revealed 4,694 high-confidence protein-protein interactions.
b. The findings are the subject of a manuscript under preparation.

Description of Activities Performed

Target gene selection

We chose 5,874 target proteins for tagging based on the characteristics that suggest they are localized in or associated with chloroplast. In addition to the 4,262 proteins we proposed initially, we added 1,612 more candidate proteins from the Arabidopsis chloroplast proteome and the Chlamydomonas pyrenoid proteome, to cover as many of the chloroplast-associated proteins as we could.

Plasmid construction and cloning

We successfully cloned 3,116 genes from 5,874 target genes using a modified PCR pipeline. We achieved a cloning success rate of 53%, which is higher than the 48% rate in our pilot study. We verified plasmid assembly using Sanger sequencing.

Chlamydomonas transformation

We transformed the 3,116 constructed plasmids into wild-type Chlamydomonas strains using a modified electroporation protocol.

Confocal microscopy imaging

We established a high-throughput imaging pipeline using a Leica SP5 point-scanning confocal microscope. We were able to adjust the focus, locate the target cell, and capture one image within a minute, which enormously accelerated fluorescence imaging. In total, we successfully localized 1,034 proteins out of 3,116 transformed plasmids. The remaining 2,082 proteins failed to be localized, possibly due to insufficient protein accumulation levels.

Indirect immunofluorescence assay

To evaluate the quality of our protein localization datasets, we performed indirect immunofluorescence on the native version of 16 fluorescently-localized proteins using custom-generated antibodies. In these experiments, 14 of the proteins (88%) showed a similar immunofluorescence localization pattern to that of the fluorescently-tagged proteins, indicating the high quality of our protein localization dataset.

Protein localization prediction

To expand our findings of protein localization to the whole proteome, we established a new protein localization predictor, named PB-Chlamy, which showed higher accuracy than the current state-of-the-art protein localization predictor, PredAlgo. PB-Chlamy was trained primarily on the 1,034 proteins we localized in this project. By applying PB-Chlamy to the whole Chlamydomonas proteome, we identified 2,245 putative chloroplast proteins, 725 putative mitochondria proteins, and 2,755 putative secretory proteins.

Affinity purification and mass spectrometry
We established a high-throughput affinity purification and mass spectrometry (AP-MS) pipeline which can process thousands of samples. We performed ~2,300 AP-MS assays on ~1,150 strains, with each strain having two biological repeats. We identified 4,694 high-confidence protein-protein interactions, providing insights into the functions of thousands of chloroplast-associated proteins.

Transient expression of Arabidopsis genes in Tobacco leaves and follow-up studies
We identified hundreds of chloroplast proteins displaying dual-localizations to other organelles. To investigate if the dual-localization phenotype is conserved in land plants, we selected one protein dual-localized to the chloroplast and mitochondria in Chlamydomonas as an example. We then verified the dual localization of its Arabidopsis homolog to the chloroplast and mitochondria, by transiently expressing the Arabidopsis homolog in tobacco leaves. These results suggest that many of the dual localizations of proteins localized in our project could be conserved in land plants. Further, we identified two T-DNA insertional mutants of this Arabidopsis gene, both of which display a slow-growing phenotype, suggesting the involvement of this newly discovered gene in the development of higher plants.

Products

Publications:
Wang, L., Patena, W., Van Baalen, K.A. et al. “A Chloroplast Protein Atlas Reveals Punctate Structures and Spatial Organization of Biosynthetic Pathways.” *Cell* 186,1–20 (2023). *https://doi.org/10.1016/j.cell.2023.06.008. This article is accessible to the public in PAGES (https://www.osti.gov/pages).*

Oral presentations:
2020 The Genomic Sciences Program PI Meeting held by the U.S. Department of Energy, Washington, D.C. ‘Toward an atlas of chloroplast protein localizations and protein-protein interactions’
2021 The 19th International Conference on the Cell and Molecular Biology of Chlamydomonas, France ‘Toward an atlas of the chloroplast proteome’
2022 18th International Congress on Photosynthesis Research, Otepoti Dunedin, New Zealand: ‘A Chloroplast Protein Atlas Reveals Novel Structures and Spatial Organization of Biosynthetic Pathways’

2022 The Genomic Sciences Program PI Meeting held by the U.S. Department of Energy, Washington, D.C. ‘Toward an atlas of chloroplast protein localizations and protein-protein interactions’
2023 The 20th International Conference on the Cell and Molecular Biology of Chlamydomonas, Princeton, U.S.A.: ‘A Chloroplast Protein Atlas Reveals Novel Structures and Spatial Organization of Biosynthetic Pathways’

Poster presentations:
2018 The 18th International Conference on the Cell and Molecular Biology of Chlamydomonas, Washington, D.C.: ‘Toward an atlas of chloroplast protein localizations and protein-protein interactions’
2021 The Genomic Sciences Program Meeting held by the U.S. Department of Energy, Washington, D.C. ‘Transforming our understanding of chloroplast-associated genes through comprehensive characterization of protein localizations and protein-protein interactions’
2022 Gordon Research Conference in Mitochondria and Chloroplasts, Mount Snow. ‘A Chloroplast Protein Atlas Reveals Novel Structures and Spatial Organization of Biosynthetic Pathways’
2023 The 20th International Conference on the Cell and Molecular Biology of Chlamydomonas, Princeton, U.S.A.: ‘A comprehensive map of the chloroplast protein interactome’

Web site:
We generated a website available to the public that allows scientists to explore the protein localizations and protein-protein interaction network: (https://www.chlamylibrary.org/allTaggedStrains).

Protein localization predictor:
We trained a protein language model to predict the protein localization in *Chlamydomonas reinhardtii*. The trained model files are available (https://huggingface.co/wpatena/PB-Chlamy/tree/main).

Predicted protein localization database:
We applied our protein localization predictor to the whole Chlamydomonas proteome. The database is available in the Table S7 of the publication (https://pubmed.ncbi.nlm.nih.gov/37437571/).