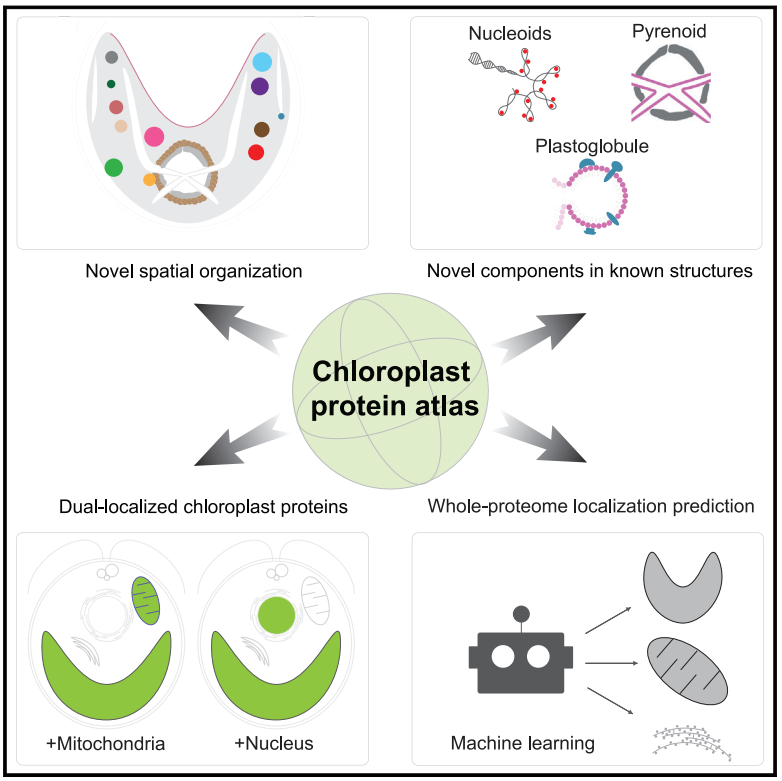
Cell

Resource

A chloroplast protein atlas reveals punctate structures and spatial organization of biosynthetic pathways

Graphical abstract



Highlights

- 1,034 candidate chloroplast proteins localized by fluorescent tagging
- This protein atlas reveals chloroplast structures, functional regions, and components
- Dual-organelle localizations suggest extensive cross-compartment coordination
- Atlas-trained machine learning predicts localizations of all *C. reinhardtii* proteins

Authors

Lianyong Wang, Weronika Patena, Kelly A. Van Baalen, ..., Danny J. Schnell, Claire D. McWhite, Martin C. Jonikas

Correspondence

mjonikas@princeton.edu

In brief

Localization analyses of 1,034 candidate chloroplast proteins reveal insights into chloroplast architecture and functions in *Chlamydomonas reinhardtii*.

Wang et al., 2023, Cell *186*, 3499–3518
August 3, 2023 © 2023 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2023.06.008

CellPress

Cell

CellPress
OPEN ACCESS

Resource

A chloroplast protein atlas reveals punctate structures and spatial organization of biosynthetic pathways

Lianyong Wang,[1] Weronika Patena,[1] Kelly A. Van Baalen,[1] Yihua Xie,[1] Emily R. Singer,[1] Sophia Gavrilenko,[1] Michelle Warren-Williams,[1] Linqu Han,[2,3] Henry R. Harrigan,[1] Linnea D. Hartz,[1] Vivian Chen,[1] Vinh T.N.P. Ton,[1] Saw Kyin,[1] Henry H. Shwe,[1] Matthew H. Cahn,[1] Alexandra T. Wilson,[1] Masayuki Onishi,[4] Jianping Hu,[2,3] Danny J. Schnell,[2] Claire D. McWhite,[1] and Martin C. Jonikas[1,5,6,*]

[1]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[2]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA
[3]MSU-DOE Plant Research Lab, Michigan State University, East Lansing, MI 48824, USA
[4]Department of Biology, Duke University, Durham, NC 27708, USA
[5]Howard Hughes Medical Institute, Princeton University, Princeton, NJ 08544, USA
[6]Lead contact
*Correspondence: mjonikas@princeton.edu
https://doi.org/10.1016/j.cell.2023.06.008

SUMMARY

Chloroplasts are eukaryotic photosynthetic organelles that drive the global carbon cycle. Despite their importance, our understanding of their protein composition, function, and spatial organization remains limited. Here, we determined the localizations of 1,034 candidate chloroplast proteins using fluorescent protein tagging in the model alga *Chlamydomonas reinhardtii*. The localizations provide insights into the functions of poorly characterized proteins; identify novel components of nucleoids, plastoglobules, and the pyrenoid; and reveal widespread protein targeting to multiple compartments. We discovered and further characterized cellular organizational features, including eleven chloroplast punctate structures, cytosolic crescent structures, and unexpected spatial distributions of enzymes within the chloroplast. We also used machine learning to predict the localizations of other nuclear-encoded *Chlamydomonas* proteins. The strains and localization atlas developed here will serve as a resource to accelerate studies of chloroplast architecture and functions.

INTRODUCTION

The chloroplast is a hallmark organelle of eukaryotic photosynthetic organisms. Over 85% of global biological light energy capture, CO_2 fixation, and O_2 production happens in chloroplasts, driving the Earth's biochemistry.[1,2] In addition to photosynthesis, the chloroplast has essential roles in key cellular processes including amino acid synthesis,[3] starch synthesis,[4] lipid metabolism,[5] isoprenoid synthesis,[6] purine/pyrimidine synthesis,[7] and the immune response of land plants.[8] Despite its importance, the mechanisms of chloroplast function and regulation are still not well understood.

Chloroplasts are thought to originate from a single primary endosymbiosis of a free-living photosynthetic cyanobacterium by a host eukaryotic cell.[9] This endosymbiosis event is thought to have given rise to the *Archaeplastidia* eukaryotic supergroup, which includes land plants, red algae, and green algae. Secondary endosymbiosis of members of *Archaeplastidia* then produced the chloroplasts found in other eukaryotic supergroups including the coccolithophores and diatoms. Hereafter, we focus on the chloroplast of *Archaeplastidia*, which remain dominant on a global scale, with land plants performing the vast majority of photosynthesis on land and green algae performing a significant portion of photosynthesis in the oceans.[1,2]

To understand the function and regulation of the chloroplast, we need to study its proteins and its sub-organellar organization. While the chloroplast has a minimal genome, the vast majority of its proteins are nuclear encoded and imported.[10] Although hundreds of nuclear-encoded proteins have recently been associated with the chloroplast through proteomics,[11–13] phylogenetics,[14] and bioinformatics studies,[15,16] the protein composition of the chloroplast remains poorly defined. Moreover, most chloroplast-associated proteins remain functionally uncharacterized.[10,17,18]

One promising starting point for understanding the functions of chloroplast-associated proteins is the systematic determination of their cellular and sub-chloroplast localizations.[19,20] Chloroplast functions are highly spatially organized into distinct regions within the organelle: sub-chloroplast regions called nucleoids contain the chloroplast's DNA,[21] chloroplast-traversing thylakoid membranes specialize in the photosynthetic capture of light energy,[22] and thylakoid-membrane-associated lipid

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

droplets called plastoglobules play roles in lipid metabolism.[23] Localizing a protein of unknown function to a functionally specialized region immediately suggests a corresponding function for the protein.

Much of our current knowledge of sub-chloroplast protein localizations comes from proteomic analyses of chloroplast fractions or sub-compartments, providing transformative advances and accelerating our understanding of the chloroplast.[23–27] Fluorescent protein tagging provides an opportunity to go beyond proteomic analyses of chloroplast fractions. It is more accurate and offers higher spatial resolution, and it can reveal new localizations and sub-organellar organization.[28] Furthermore, tagged strains can be affinity-purified and subjected to mass spectrometry-based proteomics to identify associating proteins, which can provide functional insights and identify other components of cellular structures.

To date, only a small subset of chloroplast proteins has been localized using fluorescent tagging or immunofluorescence (IF). A recent comprehensive survey[17] found that altogether only 582 of the ~3,000 bioinformatically predicted chloroplast proteins[15,16] (~19%) have been experimentally localized in the leading model land plant, *Arabidopsis thaliana*. These numbers suggest that many opportunities lie ahead for discovering novel chloroplast structures and protein functions through systematic localization of fluorescently tagged proteins.

The green alga *Chlamydomonas reinhardtii* (*Chlamydomonas* hereafter, Figure 1A), an evolutionary relative of land plants,[29,30] is a powerful model system for studying the cell biology of photosynthetic eukaryotes. Its unicellular nature allows higher throughput than land plant model systems, enabling systematic large-scale analysis of gene and protein function.[31] Work in *Chlamydomonas* has revealed conserved pathways and key principles of chloroplast biology including electron transport,[32] photosynthetic regulation,[33] assembly of photosynthetic complexes,[34] and chloroplast genome segregation.[21] Further study of the *Chlamydomonas* chloroplast will continue to shed light on the chloroplast biology of land plants, including agriculturally important crop species.

In this study, we establish a comprehensive atlas of the subcellular localizations of 1,034 chloroplast candidate proteins in *Chlamydomonas* (Figure 1). Our results reveal novel chloroplast structures and spatial organization, new components of known cellular structures, and widespread dual-localized proteins. We also use this dataset to train a more accurate *Chlamydomonas* protein localization predictor through machine learning. These insights and the associated plasmid, strain, and protein localization prediction resources open doors to the characterization of chloroplast spatial organization and poorly characterized proteins in green algae and land plants.

RESULTS

Systematic localization of 1,034 tagged proteins

To determine protein localizations, we used a previously established system[28] for expressing fluorescently tagged proteins (STAR Methods). Specifically, we cloned the open reading frame of each gene into a vector containing a constitutive promoter, C-terminal fluorescent Venus tag[35] for localization, and 3X FLAG epitope[36] for affinity purification. Electroporation of each construct into wild-type (WT) *Chlamydomonas* cells produced stable insertions at random sites within the genome.[37] We imaged protein localizations in photoautotrophically grown live cells using confocal microscopy.

To maximize the number of proteins localized to or associated with the chloroplast, we selected target proteins for fluorescent tagging from seven sources (Figures 1C and S1A; Table S1; STAR Methods). To facilitate classification of localizations, we included 29 proteins with known localization to the chloroplast or other organelles in *Chlamydomonas* (Figure S1B). Altogether, we successfully mapped the localization of 1,034 tagged proteins to 141 distinct patterns across 17 major organelles/cellular sites (Figure 1D; Tables S2 and S3).

Localization dataset validation

We first investigated the reproducibility and agreement of our results with protein localizations from previous studies to rule out inaccurate protein localization due to either protein complex disruption or alteration of native regulation (limitations of the study). Of the proteins examined, 62% were represented by at least two independent strains (Figure 1E). The localizations observed in the independent strains for a given protein agreed in >99% of the cases (Figures S1C and S1D).

As expected, all 32 known photosynthetic complex proteins and all 23 plastid ribosome proteins represented in our dataset were enriched in the chloroplast (Figure S1E). Furthermore, our localizations matched previously published localizations by fluorescent protein tagging for 27 of 28 proteins (96%) (Figure 1F; Table S4). The only exception was EZY1 (Cre06.g255750), which is normally expressed exclusively in early diploid zygotes[38] and whose mis-localization in our data (Table S2) is likely due to expression under non-native conditions.

To orthogonally validate our dataset, we assayed the localizations of 17 proteins from our dataset by indirect IF using antibodies to the native proteins in WT cells (Figures 1H–1J and S2A–S2N). Of the 16 proteins for which the cellular IF signal was observed, 14 (88%) showed a similar localization pattern to the Venus-tagged protein and two showed different patterns in the same compartment as the Venus-tagged protein. Taken together, the excellent agreement of our localization data with previous studies and IF validation suggest that our dataset provides reliable localizations for uncharacterized proteins.

Our fluorescence images are particularly effective in validating reported organelle proteomics data and identifying potential contaminant proteins in those datasets (Figure 1G; Table S2). Our localization data suggest that 26 out of the 233 proteins from the published *Chlamydomonas* chloroplast proteome[13] are actually not in the chloroplast under our experimental conditions. Similarly, 17 out of the 22 proteins previously detected in the pyrenoid proteome,[39] 56 out of the 81 proteins previously detected in the mitochondrial proteome,[40] and 21 out of the 25 reported high-confidence flagellar proteome proteins[41] do not match our localization data. We note that these numbers should not be interpreted as reflecting the overall accuracy of the mitochondrial or flagellar proteomes: we only tagged the subsets of these proteomes for which other omics evidence suggested a

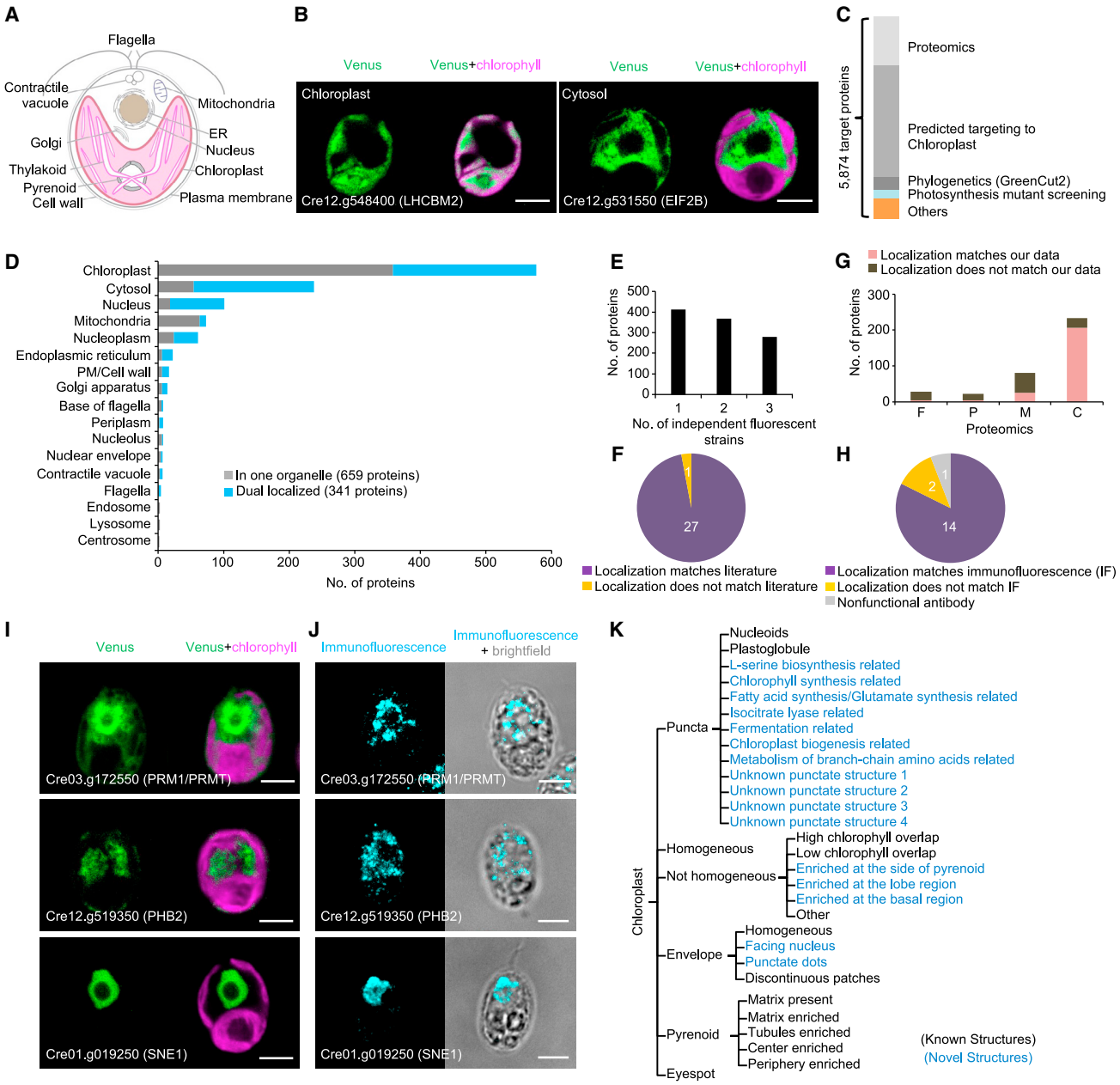


Figure 1. 1,034 tagged proteins localized to diverse patterns in 17 major compartments

(A) Cell structure of *Chlamydomonas reinhardtii*.

(B) Representative images of Venus-tagged chloroplast protein Cre12.g548400 (LHCBM2) and cytosol protein Cre12.g531550 (EIF2B).

(C) Summary of target protein sources.

(D) Number of proteins per subcellular location: proteins observed in one organelle (gray) and proteins observed in multiple organelles (blue).

(E) Number of independent strains imaged for determining localization patterns.

(F) Comparison of our localization data with those of previous literature.

(G) Comparison of our localizations with several proteomes in *Chlamydomonas*, including flagella (F), pyrenoid (P), mitochondria (M), and chloroplast (C).

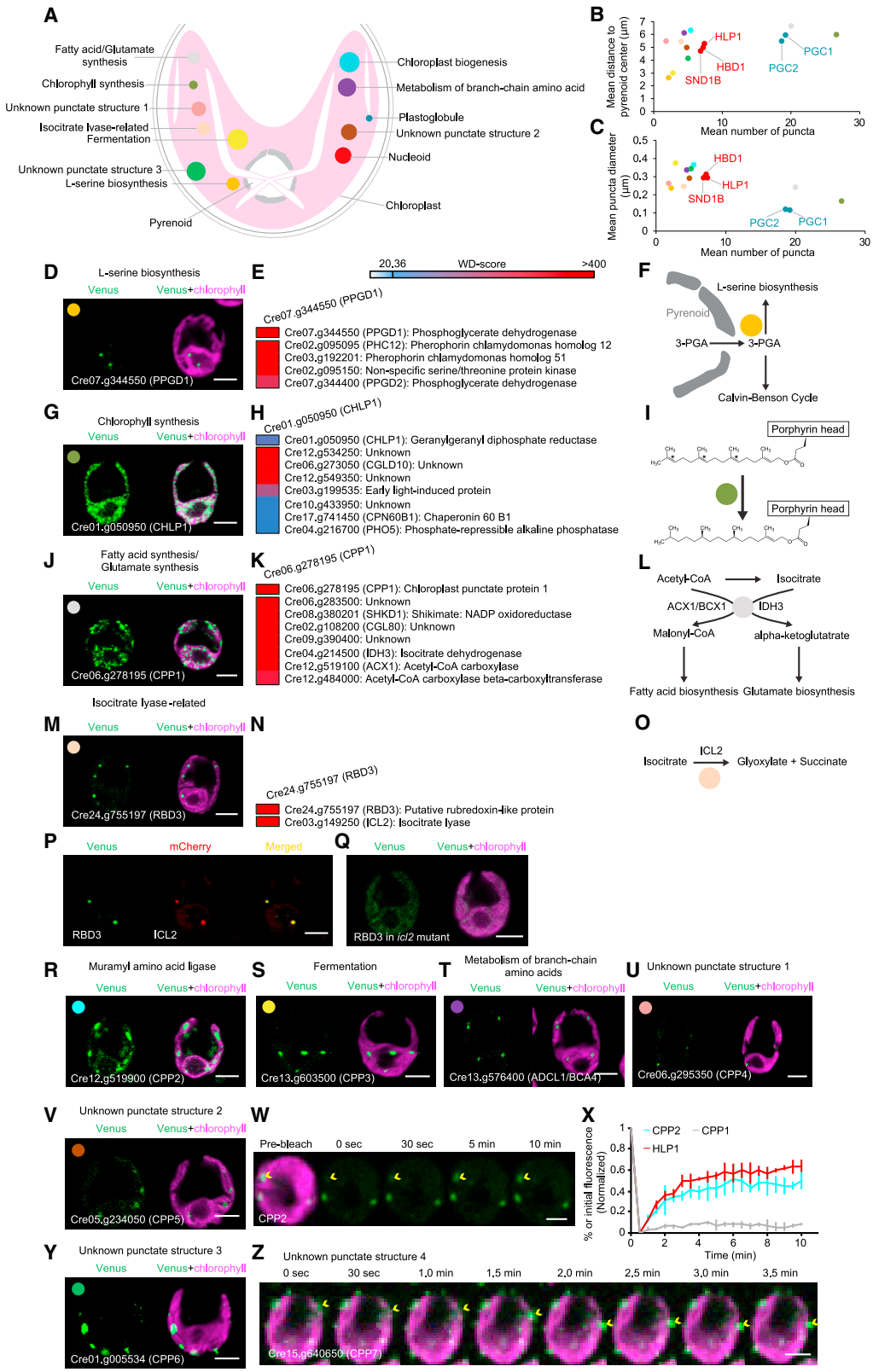
(H) Agreement of our data with localizations identified using indirect immunofluorescence (IF).

(I) Localizations of proteins that disagree with proteomics-based localizations in *Chlamydomonas*: Cre03.g172550 (PRM1/PRMT), previously found in chloroplast; Cre12.g519350 (PHB2), previously found in mitochondria; Cre01.g019250 (SNE1), previously found in flagella.

(J) Immunofluorescence localizations of PRM1/PRMT, PHB2, and SNE1 in WT.

(K) Decision tree for assigning chloroplast proteins to specific subcellular locations. Known structures in black.

All scale bars represent 5 μm.



(legend on next page)

possible chloroplast localization, which enriches our dataset for mitochondrial or flagellar proteome false positives.

The IF data validated our Venus-tagged dataset's localizations of three proteins that had been assigned to other compartments by previous proteomics studies. The conserved histone-arginine N-methyltransferase (PRM1/PRMT: Cre03.172550),[42] which had been detected in the chloroplast proteome,[13] was localized to the ER/nucleus in our data (Figure 1I). Prohibitin 2 (PHB2: Cre12.g519350),[43] previously found in the mitochondrial proteome,[40] was localized to the cytosol in our data. A GreenCut2 protein (SNE1: Cre01.g019250),[44] which had been detected in the flagellar proteome,[41] was localized to nucleoplasm in our data. Consistent with our localization dataset, we detected the native PRM1, PHB2, and SNE1 mainly in the ER/nucleus, cytosol, and nucleoplasm, respectively, by IF (Figure 1J).

Novel chloroplast punctate structures suggest compartmentalized biosynthetic reactions

We assigned the 581 chloroplast proteins to one or more of 30 sub-chloroplast locations (Figure 1K). Among the most striking were 11 unique punctate localization patterns that we could not associate with previously described structures within the chloroplast (Figures 2 and S1I; Table S2). The localization patterns differed in the number, diameter, and position of puncta within the chloroplast (Figures 2A–2C), suggesting that they correspond to distinct structures. We named the seven unnamed punctate-localized proteins chloroplast punctate proteins (CPP1–7).

To characterize these 11 structures, we performed immunoprecipitation-mass spectrometry (IP-MS) on the tagged proteins. We identified on average five high-confidence protein interactors per structure, for a total of 59 proteins associated with these punctate structures (Figures 2E, 2H, 2K, and 2N; Table S5). Many constituent proteins are conserved in land plants, suggesting that at least some of these structures are broadly conserved.

Many of the punctate-localized proteins or their interactors correspond to metabolic enzymes, suggesting that these punctate structures play functional roles in spatial organization of biosynthetic reactions. Two themes emerge. (1) Typically, only some of the enzymes of a pathway are localized to puncta, suggesting that the puncta enhance or regulate a subset of the reactions. (2) In some cases, punctate localization of an enzyme may allow it to perform its reaction at a location where its substrate is most available. These observations are consistent with previous observations of metabolism associated with cellular condensates.[45–48] Below, we discuss our data on composition and potential functions of some of these structures.

L-serine biosynthesis

The conserved predicted 3-phosphoglycerate dehydrogenase (PGDH) Cre07.g344550 (Table S2), which catalyzes the commitment step of L-serine biosynthesis, localized to puncta, most of which were directly adjacent to the pyrenoid (Figures 2D and S2A). Because the pyrenoid is the site of production of 3-phosphoglycerate (3-PGA) by ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco), the localization of PGDH to puncta next to the pyrenoid may enhance its activity through metabolic channeling. The pyrenoid is surrounded by presumably impermeable starch plates that are only punctured in a few places by thylakoid membranes[49]; we speculate that the PGDH puncta localize to these openings to capture exiting 3-PGA (Figure 2F). Cre07.g344550 co-precipitated with another predicted PGDH encoded adjacent to it in the genome, Cre07.g344400 (Figure 2E), suggesting that both enzymes may function in these puncta. From these observations, we propose naming these enzymes pyrenoid-associated

Figure 2. Novel punctate structures in the chloroplast

(A) Diagram illustrating the 12 chloroplast puncta structures observed. Relative size and approximate distance from pyrenoid center are represented; for simplicity, only one of each structure is shown. Each structure is assigned a unique color that is used throughout this figure and Figure S1I.

(B and C) Punctate structures showed differences in the average position, number, and size of puncta. For each punctate structure, its mean distance to the pyrenoid center, the average number of puncta, and the mean punctum size are shown. Additional information is shown in Figure S1I.

(D) Representative images of Cre07.g344550 (PPGD1).

(E) High-confidence interacting proteins of PPGD1. WD scores (an associated *P*-value and *Z*-score to identify high-confidence interacting proteins in the Comparative Proteomics Analysis Software Suite) represent our confidence in the interactions; scores greater than 20.36 correspond to the 3.7% highest-confidence interactions (STAR Methods).

(F) Diagram illustrating how adjacent localization of 3-phosphoglycerate dehydrogenase PPGD1 to the pyrenoid could enhance its activity.

(G) Representative images of Cre01.g050950 (CHLP1).

(H) High-confidence interacting proteins of CHLP1.

(I) Diagram illustrating how geranylgeranyl diphosphate reductase CHLP1 could catalyze the last step of chlorophyll biogenesis.

(J) Representative images of Cre06.g278195 (CPP1).

(K) High-confidence interacting proteins of CPP1.

(L) Diagram illustrating how CPP1 could regulate the branching of metabolism between fatty acid synthesis and glutamate production.

(M) Representative images of Cre24.g755197 (RBD3) in WT.

(N) High-confidence interacting proteins of RBD3.

(O) Diagram illustrating how RBD3 puncta could support isocitrate lyase 2 (ICL2).

(P) Co-localization of RBD3-Venus and ICL2-mCherry in the same cell.

(Q) RBD3 localization in *icl2* insertional mutant.

(R–Y) Representative images of Cre12.g519900 (CPP2) (R), Cre13.g603500 (CPP3) (S), Cre13.g576400 (ADCL1/BCA4) (T), Cre06.g295350 (CPP4) (U), Cre05.g234050 (CPP5) (V), and Cre01.g005534 (CPP6) (Y).

(W) Fluorescence recovery of CPP2-Venus punctum during 10 min after photobleaching of the whole punctum. Yellow arrow, bleached punctum.

(X) Fluorescence recovery profile of puncta of CPP2, Cre06.g285401 (HLP1), and CPP1. Shown are mean ± SD of three different puncta for each protein.

(Z) Representative time course of Venus-tagged Cre15.g640650 (CPP7). Yellow arrows track one punctum.

All scale bars represent 5 μm.

PGDH (PPGD1 and PPGD2, respectively) and the puncta glydehydrosomes.

Chlorophyll biosynthesis

CHLP1 (Cre01.g050950), a conserved predicted geranylgeranyl diphosphate reductase (Table S2) that catalyzes a series of reductions during the last step of the biogenesis of the photosynthetic pigment chlorophyll,[50] formed the most-numerous puncta of all the punctate localization patterns we observed (Figure 2G). As none of the eight other enzymes in the chlorophyll biosynthesis pathway we tagged and examined showed punctate localization (Table S2), this is likely the only step of chlorophyll biosynthesis that benefits from being performed in puncta. Because CHLP1 performs three separate reductions on its substrate (Figure 2I), we speculate that its localization to puncta increases the enzyme's local concentration, allowing released product to more efficiently re-bind a CHLP1 active site during these sequential reductions. Consistent with a role of these CHLP1 structures in photosynthesis, two *chlp1* insertional mutants showed defective growth in normal light but not in dim light conditions (Figures S3A–S3C).

CHLP1 physically interacted with Cre03.g199535, a conserved early light-induced protein with a chlorophyll a/*b* binding protein domain,[51] and with the conserved but poorly characterized protein CGLD10 (Cre06.g273050) (Figure 2H), suggesting that Cre03.g199535 and CGLD10 also play roles in chlorophyll biosynthesis, possibly by enhancing CHLP1's function.

Metabolic regulation

The punctate-localized conserved protein Cre06.g278195 (Figures 2J and S2B; Table S2), which we named CPP1, co-precipitated with two subunits of acetyl-CoA carboxylase, ACX1 (Cre12.g519100) and BCX1 (Cre12.g484000), and with the chloroplastic isocitrate dehydrogenase 3 (IDH3) (Cre04.g214500) (Figure 2K). Both of these enzymes perform essentially irreversible reactions downstream of citrate/acetyl-CoA. Thus, we hypothesize that the puncta formed by CPP1 regulate the branching of metabolism between fatty acid synthesis, which is downstream of acetyl-CoA carboxylase,[52] and the production of glutamate, which is downstream of IDH[53] (Figure 2L).

Glyoxylate cycle

The conserved punctate-localized protein RBD3 (Cre24.g755197) (Figures 2M and S2C) contains a predicted Rubredoxin-like domain (Figure 2N), suggesting a role in electron transfer. RBD3 physically interacted and co-localized with predicted isocitrate lyase 2 (ICL2) (Cre03.g149250), a key enzyme in the glyoxylate cycle, which also localized to chloroplast puncta (Figures 2N, 2P, and S2D) and allows cells to metabolize two-carbon compounds such as acetate when simple sugars are not available (Figure 2O). The RBD3 puncta disappeared in *icl2* insertional mutants (Figures 2Q and S3D–S3I), indicating that ICL2 is essential for the formation of RBD3 puncta.

The glyoxylate cycle is thought to occur in peroxisomes,[54] but evidence from our data and the literature suggest that the enzymes necessary for the cycle are also present in the chloroplast. Specifically, we observed chloroplast localization of a malate dehydrogenase 1 (MDH1) (Cre03.194850) and a citrate synthase (Cre13.g579050) (Table S2). Succinate dehydrogenase activity has been observed in spinach chloroplasts,[55] and bioinformatics[16] (Table S7) predicts the chloroplast targeting of aconitases (Cre06.g252650 and Cre01.g004500), succinate dehydrogenase (Cre12.g528450), and fumarase (Cre06.g272500). Our observations therefore support the possibility that chloroplasts are able to operate a glyoxylate cycle, which could increase the cell's metabolic flexibility, and suggest that a portion of this cycle occurs in punctate structures.

Punctate structures differ in their exchange and movement dynamics

The punctate structures exhibit different dynamics in exchange of components with the chloroplast stroma and the movement of the structures within the chloroplast. Puncta of the predicted muramyl amino acid ligase CPP2 (Cre12.g519900) demonstrated rapid exchange of components with the stroma, similar to the behavior of puncta of the chloroplast-DNA-binding nucleoid component HLP1 (Figures 2W and 2X). In contrast, the punctate structure formed by CPP1, which we associated with branching of metabolism between fatty acid synthesis and the production of glutamate (Figure 2X), did not exhibit such rapid exchange. Moreover, whereas most structures did not move significantly in 10 min, puncta that contained CPP7 (Cre15.g640650) showed rapid movement on the timescale of minutes (Figure 2Z; Video S1). We speculate that the rapid exchange of CPP2 and nucleoid components with stroma and the rapid movement of CPP7 are important to the function of these compartments.

Localization data reveal components of chloroplast substructures

In addition to discovering the structures above, we identified novel components across different known substructures within the chloroplast.

Nucleoid

One of two previously known punctate structures in our dataset were nucleoids[21] (Figure 3A). Our dataset revealed one novel nucleoid protein, SND1B (Cre06.g256850) (Figure 3B), which co-precipitated and co-localized with the previously characterized nucleoid proteins HLP1 (Cre06.g285401)[56] and HBD1 (Cre16.g672300)[57] (Figures 3C and 3D; Table S2). SND1B contains a predicted histone-lysine N-methyltransferase and a SAND DNA-binding domain,[58] further supporting a nucleoid function.

Plastoglobule

The other previously known structures were plastoglobules, which are thylakoid-membrane-associated lipid droplets containing triacylglycerols, plastoquinone, phylloquinone, carotenoids, and proteins related to their biosynthesis[23] (Figure 3E). Our data revealed two novel plastoglobule-localized proteins, Cre03.g197650 and Cre03.g145507, which we named plastoglobule component 1 and 2 (PGC1 and PGC2). PGC1 contains a PAP fibrilin domain found in structural proteins of plastoglobules,[59] leading us to hypothesize that the puncta it formed (Figure 3F) corresponded to plastoglobules. PGC2 showed a similar localization pattern to PGC1, co-localized, and co-precipitated with it (Figures 3F–3H), suggesting that they are part of the same structure. Immunoprecipitation of these two proteins pulled down six proteins whose homologs were

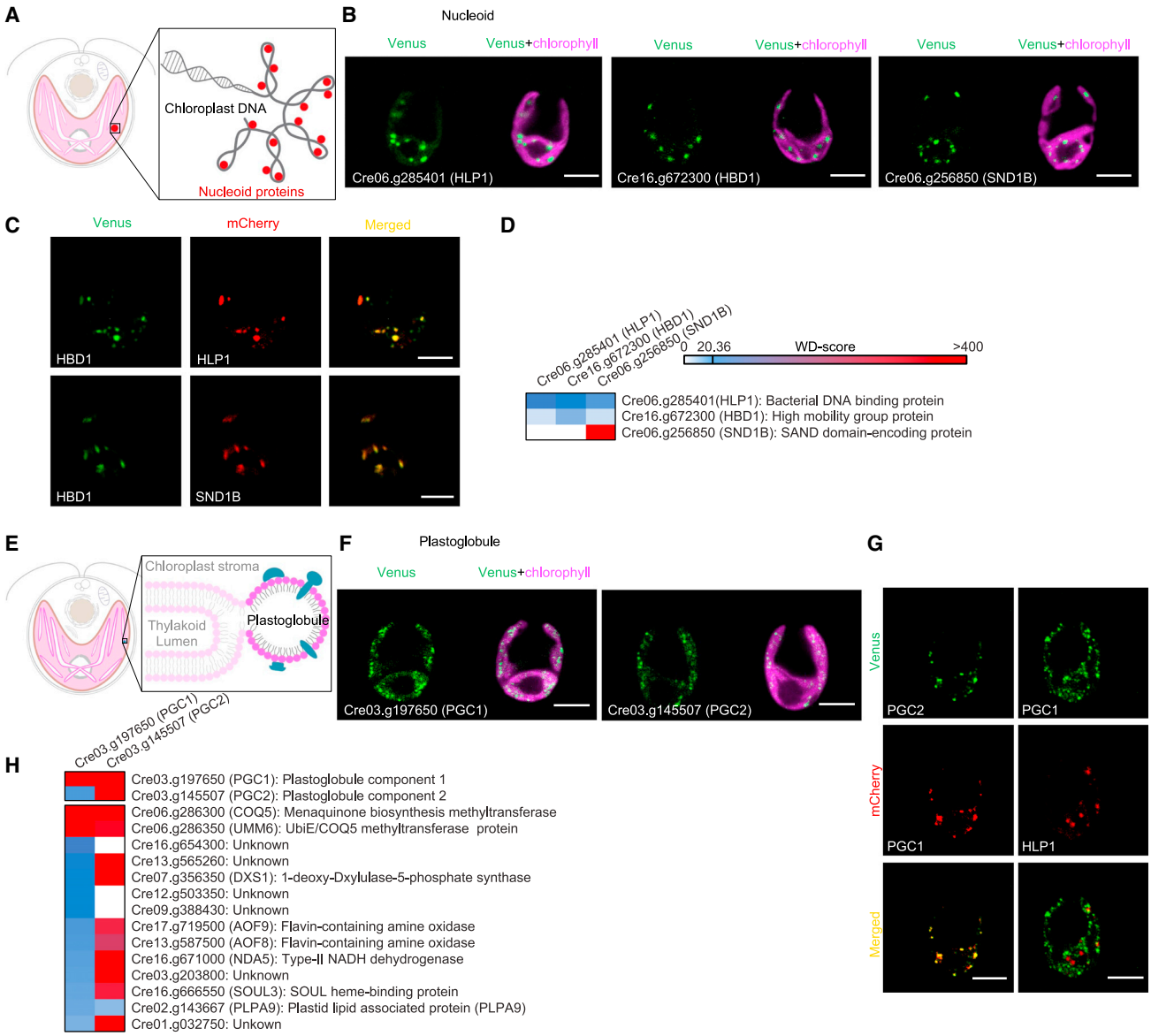
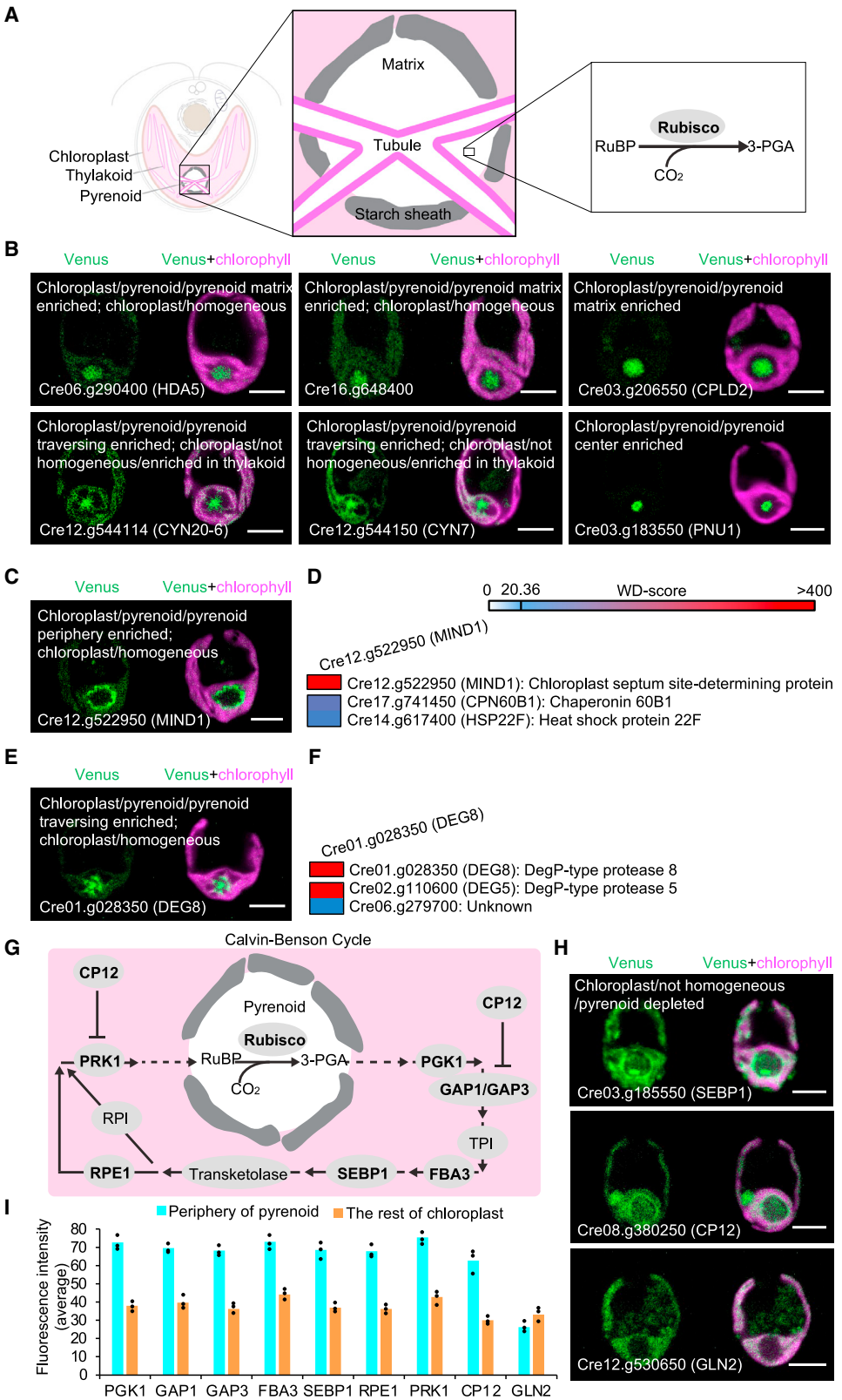


Figure 3. Novel components of chloroplast nucleoids and plastoglobules

(A) Diagram of the chloroplast nucleoid; chloroplast DNA is organized into DNA-protein conglomerates.
(B) Representative images of fluorescently tagged Cre06.g285401 (HLP1), Cre16.g672300 (HBD1), and Cre06.g256850 (SND1B).
(C) Co-localization of HLP1-mCherry, HBD1-Venus, and SND1B-mCherry.
(D) Protein-protein interactions among HLP1, HBD1, and SND1B.
(E) Diagram of a plastoglobule.
(F) Representative images of fluorescently tagged Cre03.g197650 (PGC1) and Cre03.g145507 (PGC2).
(G) Co-localization of PGC1-mCherry and PGC2-Venus.
(H) High-confidence interacting proteins of PGC1 and PGC2.
All scale bars represent 5 μm.

previously found in the *Arabidopsis* plastoglobule proteome[23,27] (Figure 3H), including the electron transport protein NAD5 (Cre16.g671000), the SOUL heme-binding protein SOUL3 (Cre16.g666550), two phylloquinone biosynthesis-related proteins UMM6 (Cre06.g286350) and COQ5 (Cre06.g286300), and the plastid lipid-associated protein PLPA9 (Cre02.g143667). We conclude that PGC1 and PGC2 are plastoglobule proteins.

Interestingly, our immunoprecipitation experiments also identified predicted protein functions not previously thought to be present at plastoglobules. Specifically, we found DXS1 (Cre07.g356350), a conserved predicted 1-deoxy-D-xylulose 5-phosphate synthase, which would generate a precursor for isoprenoid and vitamin B_1 and B_6 synthesis.[60] The immunoprecipitation also identified AOF8 (Cre13.g587500) and AOF9



(legend on next page)

(Cre17.g719500), two conserved predicted flavin-containing amine oxidases that catalyze the oxidative cleavage of alkylamines into aldehydes and ammonia. These findings suggest that plastoglobules perform previously unappreciated functions in 1-deoxy-D-xylulose 5-phosphate and alkylamine metabolism.

Pyrenoid

The pyrenoid is a non-membrane-bound proteinaceous suborganelle of the chloroplast in which the rate of CO_2 fixation into organic carbon is enhanced by supplying the CO_2-fixing enzyme Rubisco with a high concentration of CO_2[61,62] (Figure 4A). Within our dataset, we observed the localization of 18 novel proteins to the pyrenoid periphery, matrix, tubules, or pyrenoid center (Table S2). Two of the pyrenoid matrix-localized proteins, the predicted histone deacetylase HDA5 (Cre06.g290400) and uncharacterized protein Cre16.g648400 (Figure 4B), harbor predicted Rubisco-binding motifs,[63] suggesting that they bind directly to Rubisco.

MIND1 (Cre12.g522950), the *Chlamydomonas* homolog of the *Arabidopsis* chloroplast division site regulator MinD1,[64] was enriched at the pyrenoid periphery (Figure 4C). MIND1 co-precipitated with plastid chaperonin 60 beta 1 subunit (CPN60B1) (Cre17.g741450) (Figure 4D), whose *Arabidopsis* homolog has also been implicated in plastid division,[65] suggesting the conservation of this interaction in plastid division in algae. Considering that the pyrenoid typically divides by fission during chloroplast division,[66] we hypothesize that MIND1's localization to the pyrenoid periphery plays a role in coordinating pyrenoid fission with chloroplast division.

CPLD2 (Cre03.g206550), the *Chlamydomonas* homolog of the *Arabidopsis* xylulose-1,5-biphosphate (XuBP) phosphatase CbbY (AT3G48420),[67] was enriched in the pyrenoid matrix (Figure 4B). XuBP phosphatase consumes XuBP, a misfire product of Rubisco that potently inhibits the enzyme.[68] The localization of CPLD2 to the pyrenoid likely allows the cell to consume XuBP at its source.

The pyrenoid tubules are modified thylakoid membranes that traverse the pyrenoid and are thought to supply it with concentrated CO_2. We observed nine proteins localizing to the pyrenoid tubules, including two predicted peptidyl-prolyl *cis-trans* isomerases (CYN20: Cre12.g544114 and CYN7: Cre12.g544150) (Figure 4B) and the predicted DegP-type protease DEG8 (Cre01.g028350) (Figure 4E), which co-precipitated with another DegP-type protease, DEG5 (Figure 4F). These observations suggest that tubules may have a role in protein folding, degradation, and/or protein import into the pyrenoid.

Finally, our data support a role for the pyrenoid in nucleic acid degradation. The bifunctional nuclease-domain-containing protein Cre03.g183550, which we named pyrenoid nuclease 1 (PNU1), localized to the pyrenoid center (Figure 4B). In plants, bifunctional nucleases are responsible for the degradation of RNA and single-stranded DNA in several biological processes.[69] Considering that oxidized RNA localizes to the pyrenoid in *Chlamydomonas*,[70] we speculate that localizing RNA-degrading enzymes to the pyrenoid allows for increased specificity of degradation for damaged RNA.

Calvin cycle enzymes are enriched in the stroma surrounding the pyrenoid

The Calvin cycle is the metabolic cycle that enables the assimilation of CO_2. It includes the CO_2-fixing enzyme Rubisco and 11 other enzymes that convert Rubisco's product, phosphoglycerate, into its substrate, ribulose-1,5-bisphosphate, allowing the cycle to continue. In *Chlamydomonas*, Rubisco is the only Calvin cycle enzyme present in the pyrenoid, while the other enzymes are all in the stroma[46] (Figure 4G).

From our dataset, we observed that the Calvin cycle enzyme sedoheptulose-1,7-bisphosphatase (SEBP1) and the Calvin cycle regulatory protein CP12 were both enriched in a region of the stroma immediately surrounding the pyrenoid (Figure 4H). This enrichment was not noticed in our previous study[46] because of the lack of other stromal-localized proteins for comparison. Re-examining the localization of the proteins, it is now apparent that the Calvin cycle enzymes phosphoglycerate kinase 1 (PGK1), glyceraldehyde 3-phosphate dehydrogenase (GAP1 and GAP3), fructose-1,6-bisphosphate aldolase (FBA3), SEBP1, ribulose phosphate-3-epimerase (RPE1), and phosphoribulokinase (PRK1) are all enriched in the region of the stroma immediately surrounding the pyrenoid (Figures 4I and S4A–S4F). This enrichment in the periphery of the pyrenoid may enhance the activity of the Calvin cycle, considering that Rubisco resides inside the pyrenoid and its substrates and products must therefore diffuse in and out of the pyrenoid. These observations motivate questions for future research, including how these enzymes localize to the pyrenoid periphery and how their localization changes under conditions, such as high CO_2, where Rubisco dissolves into the stroma.

Unexpected thylakoid associations and protein distributions

Several proteins exhibited unexpected thylakoid association, with one showing an intriguing gradient distribution. Thylakoid

Figure 4. Identification of pyrenoid components and Calvin-Benson cycle enzyme enrichment in the stroma around the pyrenoid

(A) Diagram of the pyrenoid, highlighting the starch sheath, pyrenoid tubules, and pyrenoid matrix, where most of the carbon-fixing enzyme Rubisco is located.

(B) Representative images of Cre06.g290400, Cre16.g648400, Cre03.g206550 (CPLD2), Cre12.g544114 (CYN20-6), Cre12.g544150 (CYN7), and Cre03.g183550 (PNU1).

(C) Representative images of Cre12.g522950 (MIND1).

(D) High-confidence interacting proteins of MIND1.

(E) Representative images of Cre01.g028350 (DEG8).

(F) High-confidence interacting proteins of DEG8.

(G) Simplified model of the Calvin-Benson cycle in *Chlamydomonas*.

(H) Representative images of Cre03.g185550 (SEBP1) and Cre08.g380250 (CP12) in comparison to Cre12.g530650 (GLN2, glutamine synthetase).

(I) Comparison of Calvin-Benson cycle enzyme fluorescence in the periphery of the pyrenoid and in the rest of the chloroplast. Fluorescence intensity was measured with Fiji (ImageJ). The measured value and mean from three independent cells are shown.

All scale bars represent 5 μm.

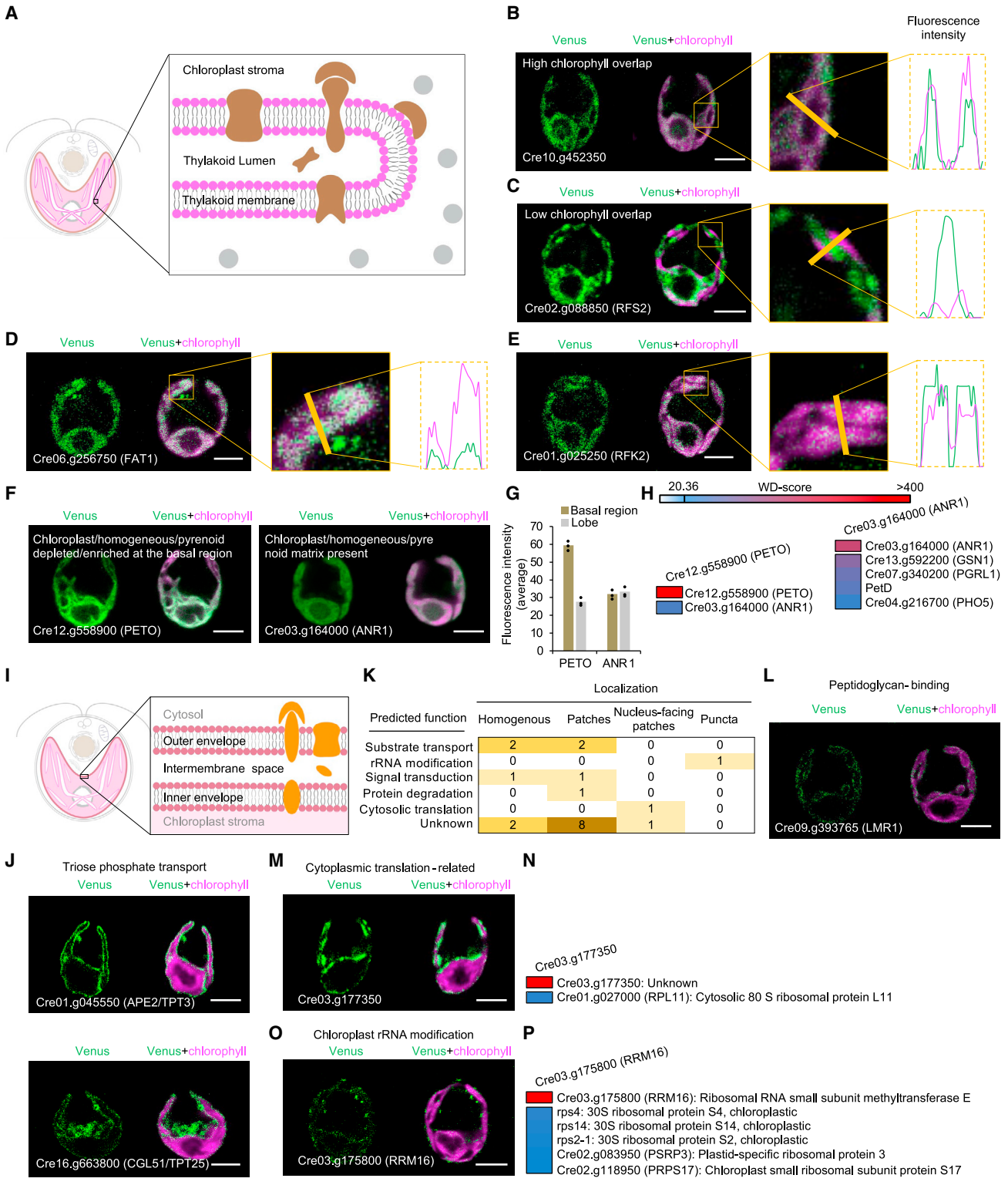


Figure 5. Localizations of membrane-associated proteins within the chloroplast

(A) Diagram showing the thylakoid membrane, thylakoid-associated enzymes (brown), and non-thylakoid-associated enzymes (gray).

(B) Representative images and line intensity profile of fluorescently tagged Cre10.g452350 and chlorophyll. Fluorescence intensity was measured with Fiji (ImageJ).

(legend continued on next page)

membranes host chlorophyll-containing protein complexes that capture light and generate ATP and NADPH for the cell. Of the proteins with non-homogeneous chloroplast localization in our dataset, 40 exhibited high localization overlap with chlorophyll (Figures 5A and 5B), while 31 exhibited low overlap (Figure 5C). We interpret high localization overlap with chlorophyll as indicative of thylakoid membrane association: of the 71 proteins with non-homogeneous localization patterns, all 11 proteins with transmembrane domains showed high chlorophyll overlap ($p = 0.001$, Fisher's exact test). Below, we illustrate how our observation of thylakoid membrane association advances understanding of protein functions.

In photosynthetic eukaryotes, fatty acids are made by fatty acid synthase in the chloroplast stroma,[71] but the localization of the enzymes that process nascent fatty acids has not been completely defined. Our data show that the only predicted chloroplastic acyl-ACP thioesterase FAT1 (Cre06.g256750), which releases fatty acids from fatty acid synthase,[5] is associated with thylakoid membranes (Figure 5D), suggesting that nascent fatty acids are released in the proximity of thylakoid membranes, into which they may initially partition.

Our data also suggest that riboflavin kinase (RFK2) (Cre01.g025250) is associated with the thylakoid membrane (Figure 5E). Riboflavin kinase phosphorylates riboflavin to produce flavin mononucleotide, an essential cofactor for the thylakoid-localized NADH dehydrogenase.[72] However, the localization of riboflavin kinase within the chloroplast was previously unknown. The localization of RFK2 to the thylakoid membrane suggests that flavin mononucleotide is produced in proximity to where it is needed for assembly into NADH dehydrogenase.

We also uncovered an intriguing distribution of a known thylakoid-associated protein, PETO (Cre12.g558900). PETO has been proposed to be important for photosynthetic cyclic electron flow, a poorly understood pathway of photosynthesis that pumps additional protons across the thylakoid membrane without producing net reducing equivalents.[73] In our dataset, PETO stood out as the only protein that showed a gradient localization pattern across the chloroplast, with a 2-fold enrichment at the base of the chloroplast (Figures 5F, 5G, and S4G). While the specific function of PETO in cyclic electron flow remains unknown, our observation of a gradient localization suggests that cyclic electron flow may be more active at the base of the chloroplast. This activity could result in pumping additional protons into the thylakoid lumen in the proximity of the pyrenoid, where they are needed to drive the conversion of HCO_3^- to CO_2 by carbonic anhydrase.[74,75]

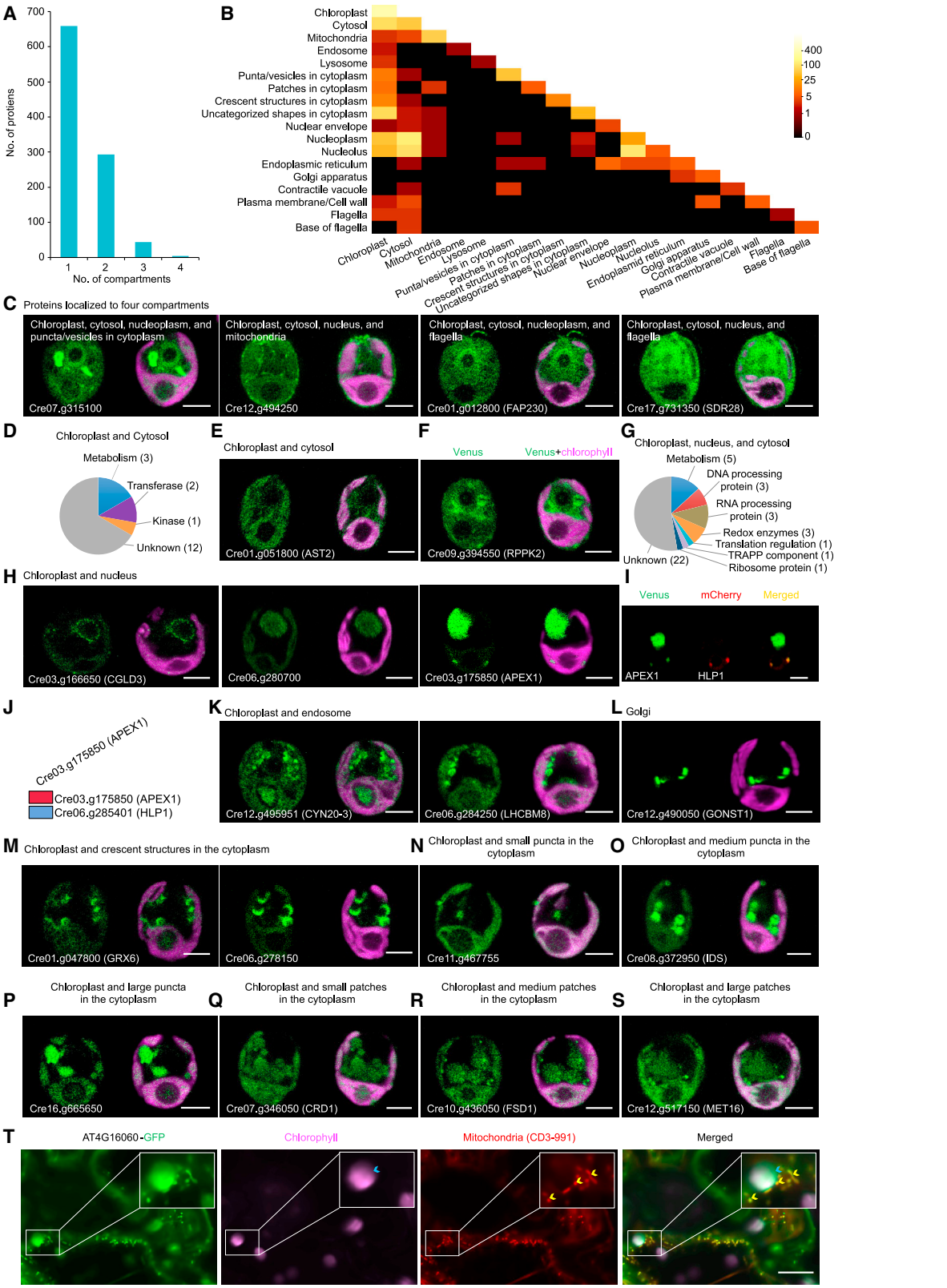
Our immunoprecipitation data confirm the previously observed physical interaction of PETO with the cyclic electron flow regulator ANR1 (Cre03.g164000)[73] (Figure 5H). Unlike PETO, however, ANR1 did not show a gradient localization (Figure 5F), and affinity purification of ANR1 did not yield detectable amounts of PETO (Figure 5H; Table S5), suggesting that only a fraction of ANR1 is associated with PETO. In addition, ANR1 co-precipitated with cytochrome b_6f subunit IV (PetD) and with the cyclic electron flow regulator proton-gradient related-like 1 (PGRL1) (Cre07.g340200)[76] (Figure 5H), supporting a possible direct role of ANR1 in the regulation of cyclic electron flow. The highest-confidence interactor of ANR1 was the predicted NADH-dependent glutamate synthase (GSN1) (Cre13.g592200), suggesting the possibility that ANR1 could downregulate cyclic electron flow in response to increased need for NADPH by GSN1.

Chloroplast envelope localization patterns suggest functionally specialized regions

The chloroplast envelope, as the interface between the chloroplast and surrounding cytosol, controls the exchange of ions, metabolites, proteins, and signals (Figure 5I). Of the 20 chloroplast-envelope-localized proteins, only five showed a homogeneous localization throughout the envelope (Figures 5J, 5K, and S4H), whereas the other 15 showed one of three distinct heterogeneous localization patterns: patches (12 proteins), nucleus-facing patches (2 proteins), and puncta (1 protein) (Figures 5K–5M, 5O, and S4I–S4K; Table S2). These observations suggest that most proteins operate in specialized regions at the chloroplast envelope. Below, we discuss protein functions associated with each localization pattern.

Proteins localized to patches along the chloroplast envelope included LMR1 (Cre09.g393765), which contains two predicted peptidoglycan-binding LysM domains[77] (Figure 5L). While some chloroplasts are surrounded by peptidoglycan, as in the moss *Physcomitrella patens*,[78] the apparent absence of most of the peptidoglycan biosynthesis genes in the *Chlamydomonas*

(C) Representative images and line intensity profile of Cre02.g088850 (RFS2).
(D) Representative images and line intensity profile of Cre06.g256750 (FAT1).
(E) Representative images and line intensity profile of Cre01.g025250 (RFK2).
(F) Representative images of Cre12.g558900 (PETO) and ANR1 (Cre03.g164000).
(G) The average fluorescence intensity of PETO-Venus and ANR1-Venus in the basal or lobe region are shown for three independent cells; bars represent mean values.
(H) High-confidence interacting proteins of PETO and ANR1. GSN1, glutamate synthase; PetD, cytochrome b_6f subunit IV; PGRL1, proton-gradient related-like 1; PHO5, alkaline phosphatases.
(I) A diagram of the chloroplast envelope highlights the outer membrane, intermembrane space, inner membrane, and chloroplast-envelope-associated proteins.
(J) Representative images of Cre01.g045550 (APE2/TPT3) and Cre16.g663800 (CGL51/TPT25).
(K) Heatmap of observed localization and predicted function of 20 chloroplast envelope proteins.
(L) Representative images of Cre09.g393765 (LMR1).
(M) Representative images of Cre03.g177350.
(N) High-confidence interacting proteins of Cre03.g177350.
(O) Representative images of Cre03.g175800 (RRM16).
(P) High-confidence interacting proteins of Cre03.g175800 (RRM16).
All scale bars, 5 μm.



(legend on next page)

genome suggests that LMR1 instead binds other glycans at the chloroplast envelope.

Proteins localized to nucleus-facing patches included the conserved protein Cre03.g177350 (Figure 5M; Table S2). This protein physically interacted with the cytosolic 80S ribosomal protein L11 (Figure 5N), suggesting that Cre03.g177350 could be involved in the cytosolic translation of chloroplast proteins before their import into the chloroplast.

The protein that localizes to puncta along the chloroplast envelope (Figure 5O) is the conserved protein RRM16 (Cre03.g175800) (Table S2), which bears two ribosomal RNA (rRNA) methyltransferase domains. This localization suggests that the chloroplast envelope could be a site where rRNA modification takes place. Consistent with this hypothesis, we detected high-confidence physical interactions between RRM16 and several chloroplast ribosome small subunit components, including rps4, rps14, rps2-1, PSRP3, and PRPS17 (Figure 5P). The presence of a predicted chloroplast-targeting sequence in RRM16[16] and its physical interactions with chloroplast-encoded ribosome subunits suggest that RRM16 is acting on chloroplast rRNA rather than cytosolic rRNA. The localization of chloroplast rRNA modification to the chloroplast envelope could provide an opportunity for cytosolic signals to regulate the chloroplast ribosome.

Many proteins have unexpected localizations to multiple compartments

The localization of a specific protein to multiple cellular compartments is a widespread phenomenon[20,79,80] that can enable signaling between organelles[81,82] or increase the number of coding products within a restricted genome size.[83]

We identified 341 proteins with multiple compartment localizations (Figure 6A), more than the approximately 250 previously identified across all studies in plants to date.[79] We observed multiple targeting in 87 distinct localization patterns (Figure 6B; Table S3), six times more distinct patterns than seen previously in plants. Four proteins were multiply localized to four compartments (Figures 6A and 6C).

Because of the selection of proteins in this study, our dataset is particularly enriched in proteins where one of the sites of localization is the chloroplast. Of the 341 multiple-localized proteins, 214 proteins were dual targeted to the chloroplast and one of 13 other regions (Figures 6B and S5A–S5U; Table S2).

Chloroplast and cytosol

We observed 16 proteins with clear dual localizations to the chloroplast and cytosol (Figures 6D, 6E, and S5A). Many of these proteins contained predicted enzymatic domains (Figure 6D), suggesting that they are enzymes that function in both compartments. In some cases, our observed dual localizations identify candidate enzymes for activities that have been observed biochemically in those compartments. For example, the activity of ribose-phosphate pyrophosphokinase, which catalyzes a key step in purine nucleotide synthesis, has been detected in both the chloroplast and cytosol in spinach,[84] but the protein responsible for the activity in the chloroplast has not previously been identified. Our observation that the conserved ribose-phosphate pyrophosphokinase (RPPK2) (Cre09.g394550) (Table S2) shows dual localization to the cytosol and chloroplast suggests that this enzyme mediates the synthesis of phosphoribosyl diphosphate in both compartments (Figures 6F and S2F).

We also observed 31 proteins with a primary fluorescence signal in the chloroplast and relatively weak signal in the cytosol (Figure S5B; Table S2). Some of these proteins are likely to be functional only in the chloroplast, as they are components of the photosynthetic apparatus or of the plastid ribosome. The observation of these proteins in the cytosol may reflect a longer cytosolic residence time before chloroplast import,[85] or could be an overexpression artifact of our system.

Chloroplast and nucleus

Our dual localization data suggest that the chloroplast and nucleus share nucleic acid processing and repair factors. We identified five proteins showing dual chloroplast and nucleus

Figure 6. Many proteins localize to multiple compartments

(A) The number of proteins localized to one, two, three, or four compartments is shown.
(B) A heatmap shows observed dual localizations.
(C) Representative images of proteins localized to four compartments.
(D) Functional classification of 18 proteins dual-localized to the chloroplast and cytosol.
(E) Representative images of Cre01.g051800 (AST2).
(F) Representative images of Cre09.g394550 (RPPK2).
(G) Functional classification of 39 proteins dual-localized to the chloroplast, nucleus, and cytosol.
(H) Representative images of proteins dual-localized to the chloroplast and nucleus.
(I) Co-localization of Cre03.g175850 (APEX1-Venus) and HLP1-mCherry in the chloroplast.
(J) The immunoprecipitation-mass spectrometry (IP-MS) data revealed a high-confidence interaction between APEX1 and HLP1.
(K) Representative images of Cre12.g495951 (CYN20-3) and Cre06.g284250 (LHCBM8).
(L) Representative images of Cre12.g490050 (GONST1) whose homolog in *Arabidopsis* (AT2G13650) is localized to the Golgi apparatus.
(M) Representative images of Cre01.g047800 (GRX6) and Cre06.g278150.
(N) Representative images of Cre11.g467755.
(O) Representative images of Cre08.g372950 (IDS1).
(P) Representative images of Cre16.g665650.
(Q) Representative images of Cre07.g346050 (CRD1).
(R) Representative images of Cre10.g436050 (FSD1).
(S) Representative images of Cre12.g517150 (MET16).
(T) Representative images of AT4G16060, the *Arabidopsis* homolog of Cre12.g494250 in (C), in tobacco leaf cells. AT4G16060 was observed in the chloroplast, cytosol, and mitochondria. The yellow arrows indicate mitochondria labeled by mCherry marker CD3-991. The blue arrows indicate chloroplasts.
All scale bars represent 5 μm.

localizations (Figures 6G, 6H, and S5E). These proteins all had predicted functions related to nucleic acids. Of these five, the conserved predicted RNA helicase CGLD3 (Cre03.g166650) and putative RNA splicing factor Cre06.g280700 localized to the nucleus and throughout the chloroplast (Figure 6H), suggesting that they act on RNA in both compartments. The conserved DNA repair exonuclease APEX1 (Cre03.g175850) (Table S2) localized to the nucleus and chloroplast nucleoids (Figure 6H) and co-localized and co-precipitated with the nucleoid component HLP1 (Figures 6I and 6J), suggesting that it contributes to the repair of both genomes.

Chloroplast and endosome or lysosome

We observed 6 proteins localized to the chloroplast and either the endosome or the lysosome (Figures 6K and S5K–S5M; Table S2). For some of these, dual localization likely reflects a functional role in both compartments. For example, the conserved peptidyl-prolyl *cis-trans* isomerase CYN20-3 (Cre12.g495951) could isomerize prolines in both the chloroplast and in the endosome (Figure 6K). However, other proteins showing dual chloroplast and endosome localization, such as the light-harvesting protein LHCBM8 (Cre06.g284250) (Figure 6K), are likely proteins that function in the chloroplast and are degraded in the lysosome by chlorophagy[86,87] or are sequestered by quality control machinery before chloroplast import.[88]

Chloroplast and crescent structures in the cytoplasm

Among the most striking dual localizations were proteins present in both the chloroplast and cytoplasmic crescent structures, a localization pattern not described previously to our knowledge (Figures 6M, S2G, S2H, and S5N). Depending on the localized protein, the crescent structures were either small (~1 μm) or medium-sized (~2 μm), representing either distinct structures or different stages of development of the same structure. The structures did not appear to be Golgi (Figure 6L), endosomes (Figures S1B and S5L), or lysosomes (Figure S5M).

Predicted domains of 9 out of 20 proteins that localize to these crescent structures suggest that the crescent structures play roles in nucleotide and phosphate metabolism (Table S2). These proteins included predicted polynucleotide phosphatase/kinase Cre11.g467709 (Figure S5V), predicted purine biosynthesis enzyme Cre17.g734100 (Figure S5N), and predicted phosphate transporter Cre07.g325740 (Figure S5V). Cellular phosphate is primarily used for nucleotide biosynthesis, so it is logical that the two functions are spatially co-localized.

The crescent structure size and the predicted protein functions lead us to speculate that the crescents correspond to the matrix of acidocalcisomes, poorly characterized vesicular structures that store phosphate as a single, large spherical granule of polyphosphate.[89–92] Fluorescently tagged proteins localizing to the matrix of acidocalcisomes would show a crescent structure due to their exclusion from the spherical polyphosphate granule (Figures 6M, S5N, and S5V). Indeed, the acidocalcisome matrix observed by electron microscopy[89] appeared as crescents of similar size to the structures we observed by microscopy.

Acidocalcisomes are possibly the only organelle conserved from bacteria to plants and humans.[90] They are essential for cellular survival under nutrient deprivation, but we are only beginning to understand their protein composition in any organism.[93] Our identification of 20 candidate acidocalcisome proteins advances the molecular characterization of these fascinating structures. Moreover, the relatively large number (12) of proteins dual-localized to the chloroplast and these structures suggests that there could be extensive interactions between chloroplasts and acidocalcisomes, with potential for cycling of phosphate between the two compartments.

Chloroplast and other structures

We observed 27 proteins dual-localized to the chloroplast and cytoplasmic puncta of one of three different diameters: small (~1 μm), medium (~2 μm), or large (~3 μm) (Figures 6N–6P, S2I–S2K, and S5O–S5Q; Table S2). All three classes of puncta contained proteins with predicted enzymatic domains (Table S2) but contained no homologs of well-characterized proteins, precluding us from conclusively assigning these localizations to known structures. We also observed 12 proteins dual-localized to the chloroplast and small (~2 μm diameter), medium (~3.5 μm), or large (~5 μm) cytoplasmic patches (Figures 6Q–6S and S5R–S5T; Table S2) or to one of many uncategorized shapes in the cytoplasm (Figure S5U). It is possible that some of these cytosolic structures consist of misfolded proteins sequestered in regions analogous to the juxtanuclear quality control compartment (JUNQ) or insoluble protein deposit (IPOD).[88] These structures are intriguing targets for future characterization.

Machine learning enables proteome-wide protein localization predictions

Machine learning allowed us to expand the scope of our experimental protein localization findings to the genome-wide scale. The current state-of-the-art predictor for *Chlamydomonas* protein localization, PredAlgo,[16] has been a tremendously useful resource for the scientific community. However, it was trained on a relatively small dataset of 152 proteins. The much larger number of protein localizations in this present work, combined with advances in machine learning classification of protein sequences, allowed us to train a more accurate protein localization predictor.

We built our predictor, PB-Chlamy, based on ProtBertBFD,[94] a natural language processing model of protein features pre-trained on the BFD database,[95] which contains 2.5 billion protein sequences from diverse organisms. We trained three separate protein sequence classifiers on *Chlamydomonas* protein localization data (Figure 7A): one each to recognize chloroplast, mitochondrial, and secretory proteins. For each localization category, we generated a combined dataset using this work, our previous protein localization study,[28] and the training dataset assembled for PredAlgo.[16] Each dataset is composed of a set of positives (proteins known to localize to a particular subcellular location) and negatives (proteins found to not localize to the location). We split each dataset into training, validation, and testing subsets, with a 3:1:1 ratio (Table S7). We used the training set to train a BertForSequenceClassification model to distinguish proteins that do or do not localize to a compartment, evaluating against the validation set of proteins during training.

We evaluated the performance of PB-Chlamy in comparison to PredAlgo. To ensure that neither of the predictors being compared had been trained on any of the proteins in the test sets, we used testing datasets with proteins used to train

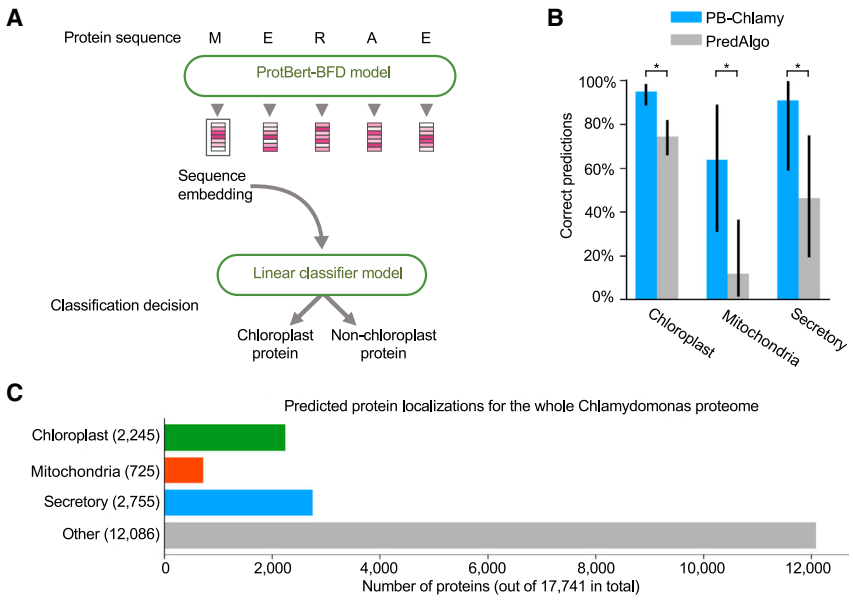


Figure 7. We predicted localizations of all proteins in *Chlamydomonas reinhardtii*

(A) The PB-Chlamy set of predictors was created based on a pre-trained ProtBertBFD model with a second-stage of supervised fine-tuning using our protein localization dataset.

(B) Comparison of PB-Chlamy performance to PredAlgo on a test set randomly chosen from our localized proteins and not used for training. The asterisk * indicates $p < 0.05$ (STAR Methods).

(C) PB-Chlamy was used to predict localizations for the whole *Chlamydomonas* proteome (v5.6, primary transcripts only).

PredAlgo excluded. PB-Chlamy reliably performs better than PredAlgo on our test sets for proteins localized to the chloroplast, mitochondrial, and secretory pathway (Figures 7B and S6).

We proceeded to use PB-Chlamy to predict protein localizations for the entire *Chlamydomonas* proteome (Figure 7C; Table S7), finding 2,245 putative chloroplast proteins, 725 putative mitochondrial proteins, and 2,755 putative secretory proteins. These numbers include 70 proteins with predicted dual localizations, mostly chloroplast + mitochondria (Table S7). Notably, we predict only two-thirds as many chloroplast proteins and one-quarter as many mitochondrial proteins as PredAlgo (which predicts 3,375 chloroplast and 2,843 mitochondrial proteins), providing a sharper view of the predicted proteome of these organelles.

DISCUSSION

Our systematic protein localization resource is useful for advancing the understanding of the molecular functions of poorly characterized proteins: the molecular functions of 702 (68%) of our localized proteins are unknown and 459 (44%) of the localized proteins were previously unnamed (Figure S1F). Our protein localizations revealed extensive spatial organization of the chloroplast, including 11 punctate structures, which appear to be metabolic hubs that enhance or regulate specific reactions, such as the commitment step of L-serine biosynthesis or the final step of chlorophyll biosynthesis.

Our study is also the largest-scale survey to date of proteins with multiple compartment localization in any photosynthetic eukaryote, making it a resource for studying the biological functions of this phenomenon, the mechanisms that underlie it, and its evolution. Indeed, 933 of the *Chlamydomonas* proteins we localized are conserved in the green alga *Volvox carteri*, 696 are conserved in the green alga *Coccomyxa subellipsoidea*, and 618 are conserved in the land plant *Arabidopsis thaliana* (Figure S1G; Table S2). We observed similarities and differences in protein localizations between *Chlamydomonas* and land plants for proteins that show single and multiple localizations (Table S2). For example, the *Chlamydomonas* protein Cre12.g494250 localized to the chloroplast, cytosol, nucleus, and mitochondrion; its *Arabidopsis* homolog AT4G16060 showed a similar multiple localization to the chloroplast, cytosol, and mitochondrion when expressed in tobacco leaves (Figures 6T, S5W, and S5X).

For proteins where we could not obtain experimental localization data, our PB-Chlamy classifier accurately predicts their localization. Together, our localization data, protein-protein interactions, and computational predictions greatly narrow down the possible functions of poorly characterized proteins and facilitate generation of specific hypotheses for their further characterization, accelerating the elucidation of chloroplast organization and function.

The images and protein-protein interactions from this study are available at https://www.chlamylibrary.org/. This site also provides links for ordering the corresponding strains and plasmids from the Chlamydomonas Resource Center.

Climate change and the rising global population drive a pressing need to understand the basic biology of photosynthetic organisms and to advance our ability to engineer them. Our study is a rich resource that lays the groundwork and paves the way for understanding the remaining mysteries of the chloroplast, the organelle at the heart of photosynthetic organisms.

Limitations of the study

Some protein localizations and protein-protein interactions reported here could be inaccurate due to technical limitations. Potential sources of artifacts include the presence of a C-terminal tag, which can disrupt protein complexes, and overexpression driven by a constitutive promoter, which can lead to abnormally high protein levels. Examples of proteins that are likely mis-localized in our study due to such artifacts are chaperonin 10 (Cpn10) (Cre03.g178450) and Rubisco accumulation factor Raf1 (Cre06.g308450) (Table S2). The cytosolic localization that we observed for the tagged versions of these proteins is inconsistent with the chloroplast localizations expected of the native proteins, based on functional characterization *in vitro*.[96,97] Finally, some proteins

may localize differently and/or have different interaction partners under growth conditions different from the ones used here.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Strains and culture conditions
- METHOD DETAILS
  - Target genes selection
  - Plasmid Construction and Cloning
  - Chlamydomonas transformation
  - Confocal Microscopy
  - Indirect Immunofluorescence Assay
  - Immunoblotting Analysis
  - Mutant generation by CRISPR-Cas9
  - Protein localization prediction
  - Affinity Purification and Mass Spectrometry
  - Transient expression of Arabidopsis gene in Tobacco leaf
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Peptide identification
  - Calculating WD-scores
  - Data visualization
  - Transmembrane prediction and Protein homology prediction
  - Statistical tests

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.06.008.

ACKNOWLEDGMENTS

We thank Princeton University Confocal Microscopy manager Gary Laevsky for instrumentation support; Princeton Institute for Computational Science and Engineering for research computing resources; Laurent Cournac, Ben Engel, Ursula Goodenough, Olivier Vallon, Christoph Benning, Ned Wingreen, Yonghua Li-Beisson, and present and former laboratory members for manuscript discussions; and Marie Bao, as part of Life Science Editors, for manuscript editing help. 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; HHMI/Simons Foundation grant 55108535; and the Lewis-Sigler Scholars Fund. M.C.J. is a Howard Hughes Medical Institute Investigator.

AUTHOR CONTRIBUTIONS

Conceptualization, L.W. and M.C.J.; gene cloning and confocal microscopy, L.W., K.A.V.B., Y.X., S.G., and H.R.H.; medium preparation, M.W.-W.; mass spectrometry analysis, S.K., H.H.S., L.W., E.R.S., and L.D.H.; prediction of protein localizations, C.D.M. and W.P.; prediction of protein structure, C.D.M., V.C., and V.T.N.P.T.; IF and immunoblotting assay, L.W.; characterization of CRISPR-Cas9 mutants, L.W. and M.O.; protein localization in tobacco, L.H., D.J.S., and J.H.; building the data/material-sharing website, M.H.C., W.P., L.W., and M.C.J. All authors analyzed the data. L.W., A.T.W., and M.C.J. wrote the manuscript, with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: May 15, 2022
Revised: May 6, 2023
Accepted: June 11, 2023
Published: July 11, 2023

REFERENCES

1. Behrenfeld, M.J., Randerson, J.T., McClain, C.R., Feldman, G.C., Los, S.O., Tucker, C.J., Falkowski, P.G., Field, C.B., Frouin, R., Esaias, W.E., et al. (2001). Biospheric primary production during an ENSO transition. Science *291*, 2594–2597. https://doi.org/10.1126/science.1055071.
2. Rousseaux, C., and Gregg, W. (2013). Interannual variation in phytoplankton primary production at a global scale. Remote Sens. *6*, 1–19. https://doi.org/10.3390/rs6010001.
3. Hildebrandt, T.M., Nunes Nesi, A., Araújo, W.L., and Braun, H.-P. (2015). Amino acid catabolism in plants. Mol. Plant *8*, 1563–1579. https://doi.org/10.1016/j.molp.2015.09.005.
4. Pfister, B., and Zeeman, S.C. (2016). Formation of starch in plant cells. Cell. Mol. Life Sci. *73*, 2781–2807. https://doi.org/10.1007/s00018-016-2250-x.
5. Hölzl, G., and Dörmann, P. (2019). Chloroplast lipids and their biosynthesis. Annu. Rev. Plant Biol. *70*, 51–81. https://doi.org/10.1146/annurev-arplant-050718-100202.
6. Lange, B.M., Rujan, T., Martin, W., and Croteau, R. (2000). Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. Proc. Natl. Acad. Sci. USA *97*, 13172–13177. https://doi.org/10.1073/pnas.240454797.
7. Zrenner, R., Stitt, M., Sonnewald, U., and Boldt, R. (2006). Pyrimidine and purine biosynthesis and degradation in plants. Annu. Rev. Plant Biol. *57*, 805–836. https://doi.org/10.1146/annurev.arplant.57.032905.105421.
8. Nomura, H., Komori, T., Uemura, S., Kanda, Y., Shimotani, K., Nakai, K., Furuichi, T., Takebayashi, K., Sugimoto, T., Sano, S., et al. (2012). Chloroplast-mediated activation of plant immune signalling in Arabidopsis. Nat. Commun. *3*, 926. https://doi.org/10.1038/ncomms1926.
9. Keeling, P.J. (2010). The endosymbiotic origin, diversification and fate of plastids. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 729–748. https://doi.org/10.1098/rstb.2009.0103.
10. Leister, D. (2003). Chloroplast research in the genomic age. Trends Genet. *19*, 47–56. https://doi.org/10.1016/S0168-9525(02)00003-3.
11. Dunkley, T.P.J., Hester, S., Shadforth, I.P., Runions, J., Weimar, T., Hanton, S.L., Griffin, J.L., Bessant, C., Brandizzi, F., Hawes, C., et al. (2006). Mapping the Arabidopsis organelle proteome. Proc. Natl. Acad. Sci. USA *103*, 6518–6523. https://doi.org/10.1073/pnas.0506958103.
12. Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., et al. (2010). AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol. Cell. Proteomics *9*, 1063–1084. https://doi.org/10.1074/mcp.M900325-MCP200.
13. Terashima, M., Specht, M., Naumann, B., and Hippler, M. (2010). Characterizing the anaerobic response of *Chlamydomonas reinhardtii* by

quantitative proteomics. Mol. Cell. Proteomics *9*, 1514–1532. https://doi.org/10.1074/mcp.M900421-MCP200.

14. Karpowicz, S.J., Prochnik, S.E., Grossman, A.R., and Merchant, S.S. (2011). The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J. Biol. Chem. *286*, 21427–21439. https://doi.org/10.1074/jbc.M111.233734.
15. Emanuelsson, O., Nielsen, H., and von Heijne, G.V. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. *8*, 978–984. https://doi.org/10.1110/ps.8.5.978.
16. Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G., et al. (2012). PredAlgo: A new subcellular localization prediction tool dedicated to green algae. Mol. Biol. Evol. *29*, 3625–3639. https://doi.org/10.1093/molbev/mss178.
17. Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N., and Millar, A.H. (2017). SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. Nucleic Acids Res. *45*, D1064–D1074. https://doi.org/10.1093/nar/gkw1041.
18. Leister, D., and Kleine, T. (2008). Towards a comprehensive catalog of chloroplast proteins and their interactions. Cell Res. *18*, 1081–1083. https://doi.org/10.1038/cr.2008.297.
19. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691. https://doi.org/10.1038/nature02026.
20. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*, eaal3321. https://doi.org/10.1126/science.aal3321.
21. Kobayashi, Y., Misumi, O., Odahara, M., Ishibashi, K., Hirono, M., Hidaka, K., Endo, M., Sugiyama, H., Iwasaki, H., Kuroiwa, T., et al. (2017). Holliday junction resolvases mediate chloroplast nucleoid segregation. Science *356*, 631–634. https://doi.org/10.1126/science.aan0038.
22. Wise, R.R., and Hoober, J.K. (2006). The Structure and Function of Plastids (Springer), pp. 3–21.
23. Ytterberg, A.J., Peltier, J.-B., and van Wijk, K.J. (2006). Protein profiling of plastoglobules in chloroplasts and chromoplasts. A surprising site for differential accumulation of metabolic enzymes. Plant Physiol. *140*, 984–997. https://doi.org/10.1104/pp.105.076083.
24. Ferro, M., Salvi, D., Brugière, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., Joyard, J., and Rolland, N. (2003). Proteomics of chloroplast envelope membranes from *Arabidopsis thaliana*. Mol. Cell. Proteomics *2*, 325–345. https://doi.org/10.1074/mcp.M300030-MCP200.
25. Friso, G., Giacomelli, L., Ytterberg, A.J., Peltier, J.B., Rudella, A., Sun, Q., and Wijk, K.J. (2004). In-depth analysis of the thylakoid membrane proteome of Arabidopsis thaliana chloroplasts: new proteins, new functions, and a plastid proteome Datebase. Plant Cell *16*, 478–499. https://doi.org/10.1105/tpc.017814.
26. Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., Watkins, K.P., Barkan, A., and Wijk, K.J. (2011). Nucleoid-Enriched Proteomes in Developing Plastid and Chloroplasts from Maize leaves: A New Conceptual Framework for Nucleoid Functions. Plant Physiol. *158*, 159–189. https://doi.org/10.1104/pp.111.188474.
27. Lundquist, P.K., Poliakov, A., Bhuiyan, N.H., Zybailov, B., Sun, Q., and van Wijk, K.J. (2012). The functional network of the *Arabidopsis* plastoglobule proteome based on quantitative proteomics and genome-wide coexpression analysis. Plant Physiol. *158*, 1172–1192. https://doi.org/10.1104/pp.111.193144.
28. Mackinder, L.C.M., Chen, C., Leib, R.D., Patena, W., Blum, S.R., Rodman, M., Ramundo, S., Adams, C.M., and Jonikas, M.C. (2017). A spatial interactome reveals the protein organization of the algal CO_2-concentrating mechanism. Cell *171*, 133–147.e14. https://doi.org/10.1016/j.cell.2017.08.044.
29. Gutman, B.L., and Niyogi, K.K. (2004). Chlamydomonas and Arabidopsis. A dynamic duo. Plant Physiol. *135*, 607–610. https://doi.org/10.1104/pp.104.041491.
30. Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science *318*, 245–250. https://doi.org/10.1126/science.1143609.
31. Fauser, F., Vilarrasa-Blasi, J., Onishi, M., Ramundo, S., Patena, W., Millican, M., Osaki, J., Philp, C., Nemeth, M., Salomé, P.A., et al. (2022). Systematic characterization of gene function in a photosynthetic organism alga *Chlamydomonas reinhardtii*. Nat. Genet. *54*, 705–714. https://doi.org/10.1038/s41588-022-01052-9.
32. Iwai, M., Takizawa, K., Tokutsu, R., Okamuro, A., Takahashi, Y., and Minagawa, J. (2010). Isolation of the elusive supercomplex that drives cyclic electron flow in photosynthesis. Nature *464*, 1210–1213. https://doi.org/10.1038/nature08885.
33. Depège, N., Bellafiore, S., and Rochaix, J.-D. (2003). Role of chloroplast protein kinase Stt7 in LHCII phosphorylation and state transition in Chlamydomonas. Science *299*, 1572–1575. https://doi.org/10.1126/science.1081397.
34. Minai, L., Wostrikoff, K., Wollman, F.-A., and Choquet, Y. (2006). Chloroplast biogenesis of photosystem II cores involves a series of assembly-controlled steps that regulate translation. Plant Cell *18*, 159–175. https://doi.org/10.1105/tpc.105.037705.
35. Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nat. Biotechnol. *20*, 87–90. https://doi.org/10.1038/nbt0102-87.
36. Hopp, T.P., Prickett, K.S., Price, V.L., Libby, R.T., March, C.J., Pat Cerretti, D.P., Urdal, D.L., and Conlon, P.J. (1988). A short polypeptide marker sequence useful for recombinant protein identification and purification. Nat. Biotechnol. *6*, 1204–1210. https://doi.org/10.1038/nbt1088-1204.
37. Zhang, R., Patena, W., Armbruster, U., Gang, S.S., Blum, S.R., and Jonikas, M.C. (2014). High-throughput genotyping of green algal mutants reveals random distribution of mutagenic insertion sites and endonucleolytic cleavage of transforming DNA. Plant Cell *26*, 1398–1409. https://doi.org/10.1105/tpc.114.124099.
38. Armbrust, E.V., Ferris, P.J., and Goodenough, U.W. (1993). A mating type-linked gene cluster expressed in Chlamydomonas zygotes participates in the uniparental inheritance of the chloroplast genome. Cell *74*, 801–811. https://doi.org/10.1016/0092-8674(93)90460-8.
39. Zhan, Y., Marchand, C.H., Maes, A., Mauries, A., Sun, Y., Dhaliwal, J.S., Uniacke, J., Arragain, S., Jiang, H., Gold, N.D., et al. (2018). Pyrenoid functions revealed by proteomics in *Chlamydomonas reinhardtii*. PLoS One *13*, e0185039. https://doi.org/10.1371/journal.pone.0185039.
40. Atteia, A., Adrait, A., Brugière, S., Tardif, M., van Lis, R., Deusch, O., Dagan, T., Kuhn, L., Gontero, B., Martin, W., et al. (2009). A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the -proteobacterial mitochondrial ancestor. Mol. Biol. Evol. *26*, 1533–1548. https://doi.org/10.1093/molbev/msp068.
41. Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J. Cell Biol. *170*, 103–113. https://doi.org/10.1083/jcb.200504008.
42. Scebba, F., De Bastiani, M., Bernacchia, G., Andreucci, A., Galli, A., and Pitto, L. (2007). PRMT11: a new Arabidopsis MBD7 protein partner with arginine methyltransferase activity. Plant J. *52*, 210–222. https://doi.org/10.1111/j.1365-313X.2007.03238.x.
43. Wang, Y., Ries, A., Wu, K., Yang, A., and Crawford, N.M. (2010). The Arabidopsis prohibitin gene *PHB3* functions in nitric oxide–mediated responses and in hydrogen peroxide–induced nitric oxide accumulation. Plant Cell *22*, 249–259. https://doi.org/10.1105/tpc.109.072066.

44. Major, L.L., Wolucka, B.A., and Naismith, J.H. (2005). Structure and function of GDP-mannose-3',5'-epimerase: an enzyme which performs three chemical reactions at the same active site. J. Am. Chem. Soc. *127*, 18309–18320. https://doi.org/10.1021/ja056490i.

45. Castellana, M., Wilson, M.Z., Xu, Y., Joshi, P., Cristea, I.M., Rabinowitz, J.D., Gitai, Z., and Wingreen, N.S. (2014). Enzyme clustering accelerates processing of intermediates through metabolic channeling. Nat. Biotechnol. *32*, 1011–1018. https://doi.org/10.1038/nbt.3018.

46. Küken, A., Sommer, F., Yaneva-Roder, L., Mackinder, L.C.M., Höhne, M., Geimer, S., Jonikas, M.C., Schroda, M., Stitt, M., Nikoloski, Z., et al. (2018). Effects of microcompartmentation on flux distribution and metabolic pools in *Chlamydomonas reinhardtii* chloroplasts. eLife *7*, e37960. https://doi.org/10.7554/eLife.37960.

47. O'Connell, J.D., Zhao, A., Ellington, A.D., and Marcotte, E.M. (2012). Dynamic reorganization of metabolic enzymes into intracellular bodies. Annu. Rev. Cell Dev. Biol. *28*, 89–111. https://doi.org/10.1146/annurev-cellbio-101011-155841.

48. Pareek, V., Sha, Z., He, J., Wingreen, N.S., and Benkovic, S.J. (2021). Metabolic channeling: predictions, deductions, and evidence. Mol. Cell *81*, 3775–3785. https://doi.org/10.1016/j.molcel.2021.08.030.

49. Engel, B.D., Schaffer, M., Kuhn Cuellar, L., Villa, E., Plitzko, J.M., and Baumeister, W. (2015). Native architecture of the Chlamydomonas chloroplast revealed by in situ cryo-electron tomography. eLife *4*, e04889. https://doi.org/10.7554/eLife.04889.

50. Tanaka, R., Oster, U., Kruse, E., Rüdiger, W., and Grimm, B. (1999). Reduced activity of geranylgeranyl reductase leads to loss of chlorophyll and tocopherol and to partially geranylgeranylated chlorophyll in transgenic Tobacco plants expressing antisense RNA for geranylgeranyl reductase. Plant Physiol. *120*, 695–704. https://doi.org/10.1104/pp.120.3.695.

51. Tanaka, R., Rothbart, M., Oka, S., Takabayashi, A., Takahashi, K., Shibata, M., Myouga, F., Motohashi, R., Shinozaki, K., Grimm, B., et al. (2010). LIL3, a light-harvesting-like protein, plays an essential role in chlorophyll and tocopherol biosynthesis. Proc. Natl. Acad. Sci. USA *107*, 16721–16725. https://doi.org/10.1073/pnas.1004699107.

52. Sasaki, Y., and Nagano, Y. (2004). Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. Biosci. Biotechnol. Biochem. *68*, 1175–1184. https://doi.org/10.1271/bbb.68.1175.

53. Elias, B.A., and Givan, C.V. (1977). Alpha-ketoglutarate supply for amino acid synthesis in higher plant chloroplasts: intrachloroplastic localization of NADP-specific isocitrate dehydrogenase. Plant Physiol. *59*, 738–740. https://doi.org/10.1104/pp.59.4.738.

54. Lauersen, K.J., Willamme, R., Coosemans, N., Joris, M., Kruse, O., and Remacle, C. (2016). Peroxisomal Microbodies are at the crossroads of acetate assimilation in the green microalga *Chlamydomonas reinhardtii*. Algal Res. *16*, 266–274. https://doi.org/10.1016/j.algal.2016.03.026.

55. Willeford, K.O., Gombos, Z., and Gibbs, M. (1989). Evidence for chloroplastic succinate dehydrogenase participating in the chloroplastic respiratory and photosynthetic electron transport Chains of *Chlamydomonas reinhardtii*. Plant Physiol. *90*, 1084–1087. https://doi.org/10.1104/pp.90.3.1084.

56. Karcher, D., Köster, D., Schadach, A., Klevesath, A., and Bock, R. (2009). The Chlamydomonas chloroplast HLP protein is required for nucleoid organization and genome maintenance. Mol. Plant *2*, 1223–1232. https://doi.org/10.1093/mp/ssp083.

57. Takusagawa, M., Kobayashi, Y., Fukao, Y., Hidaka, K., Endo, M., Sugiyama, H., Hamaji, T., Kato, Y., Miyakawa, I., Misumi, O., et al. (2021). HBD1 protein with a tandem repeat of two HMG-box domains is a DNA clip to organize chloroplast nucleoids in *Chlamydomonas reinhardtii*. Proc. Natl. Acad. Sci. USA *118*, e2021053118. https://doi.org/10.1073/pnas.2021053118.

58. Bottomley, M.J., Collard, M.W., Huggenvik, J.I., Liu, Z., Gibson, T.J., and Sattler, M. (2001). The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. Nat. Struct. Biol. *8*, 626–633. https://doi.org/10.1038/89675.

59. Deruère, J., Römer, S., d'Harlingue, A., Backhaus, R.A., Kuntz, M., and Camara, B. (1994). Fibril assembly and carotenoid overaccumulation in chromoplasts: a model for supramolecular lipoprotein structures. Plant Cell *6*, 119–133. https://doi.org/10.1105/tpc.6.1.119.

60. Lois, L.M., Campos, N., Putra, S.R., Danielsen, K., Rohmer, M., and Boronat, A. (1998). Cloning and characterization of a gene from *Escherichia coli* encoding a transketolase-like enzyme that catalyzes the synthesis of D-1-deoxyxylulose 5-phosphate, a common precursor for isoprenoid, thiamin, and pyridoxol biosynthesis. Proc. Natl. Acad. Sci. USA *95*, 2105–2110. https://doi.org/10.1073/pnas.95.5.2105.

61. Osafune, T., Yokota, A., Sumida, S., and Hase, E. (1990). Immunogold localization of ribulose-1,5-bisphosphate carboxylase with reference to pyrenoid morphology in chloroplasts of synchronized *Euglena gracilis* cells. Plant Physiol. *92*, 802–808. https://doi.org/10.1104/pp.92.3.802.

62. Wunder, T., Oh, Z.G., and Mueller-Cajar, O. (2019). CO_2-fixing liquid droplets: towards a dissection of the microalgal pyrenoid. Traffic *20*, 380–389. https://doi.org/10.1111/tra.12650.

63. Meyer, M.T., Itakura, A.K., Patena, W., Wang, L., He, S., Emrich-Mills, T., Lau, C.S., Yates, G., Mackinder, L.C.M., and Jonikas, M.C. (2020). Assembly of the algal CO_2-fixing organelle, the pyrenoid, is guided by a RuBisCO-binding motif. Sci. Adv. *6*, eabd2408. https://doi.org/10.1126/sciadv.abd2408.

64. Fujiwara, M.T., Yasuzawa, M., Sasaki, S., Nakano, T., Niwa, Y., Yoshida, S., Abe, T., and Itoh, R.D. (2017). The Arabidopsis minD mutation causes aberrant FtsZ1 ring placement and moderate heterogeneity of chloroplasts in the leaf epidermis. Plant Signal. Behav. *12*, e1343776. https://doi.org/10.1080/15592324.2017.1343776.

65. Suzuki, K., Nakanishi, H., Bower, J., Yoder, D.W., Osteryoung, K.W., and Miyagishima, S.Y. (2009). Plastid chaperonin proteins Cpn60α and Cpn60β are required for plastid division in *Arabidopsis thaliana*. BMC Plant Biol. *9*, 38. https://doi.org/10.1186/1471-2229-9-38.

66. Freeman Rosenzweig, E.S., Xu, B., Kuhn Cuellar, L., Martinez-Sanchez, A., Schaffer, M., Strauss, M., Cartwright, H.N., Ronceray, P., Plitzko, J.M., Förster, F., et al. (2017). The eukaryotic CO_2-concentrating organelle is liquid-like and exhibits dynamic reorganization. Cell *171*, 148–162.e19. https://doi.org/10.1016/j.cell.2017.08.008.

67. Bracher, A., Sharma, A., Starling-Windhof, A., Hartl, F.U., and Hayer-Hartl, M. (2015). Degradation of potent RuBisCO inhibitor by selective sugar phosphatase. Nat. Plants *1*, 14002. https://doi.org/10.1038/nplants.2014.2.

68. Pearce, F.G. (2006). Catalytic by-product formation and ligand binding by ribulose bisphosphate carboxylases from different phylogenies. Biochem. J. *399*, 525–534. https://doi.org/10.1042/BJ20060430.

69. Pérez-Amador, M.A., Abler, M.L., De Rocher, E.J., Thompson, D.M., van Hoof, A., LeBrasseur, N.D., Lers, A., and Green, P.J. (2000). Identification of BFN1, a bifunctional nuclease induced during leaf and stem senescence in Arabidopsis. Plant Physiol. *122*, 169–180. https://doi.org/10.1104/pp.122.1.169.

70. Zhan, Y., Dhaliwal, J.S., Adjibade, P., Uniacke, J., Mazroui, R., and Zerges, W. (2015). Localized control of oxidized RNA. J. Cell Sci. *128*, 4210–4219. https://doi.org/10.1242/jcs.175232.

71. Walker, K.A., and Harwood, J.L. (1985). Localization of chloroplastic fatty acid synthesis *de novo* in the stroma. Biochem. J. *226*, 551–556. https://doi.org/10.1042/bj2260551.

72. Tsibris, J.C., McCormick, D.B., and Wright, L.D. (1966). Studies on the binding and function of flavin phosphates with flavin mononucleotide-dependent enzymes. J. Biol. Chem. *241*, 1138–1143. https://doi.org/10.1016/S0021-9258(18)96813-4.

73. Takahashi, H., Schmollinger, S., Lee, J.-H., Schroda, M., Rappaport, F., Wollman, F.-A., and Vallon, O. (2016). PETO Interacts with other effectors

of cyclic electron flow in Chlamydomonas. Mol. Plant *9*, 558–568. https://doi.org/10.1016/j.molp.2015.12.017.

74. Pronina, N.A., and Semenenko, V.E. (1990). Membrane-bound carbonic anhydrase takes part in CO_2 concentration in algal cells. Curr. Res. Photosynth. *4*, 489–492. https://doi.org/10.1007/978-94-009-0511-5_739.
75. Raven, J.A. (1997). CO_2-concentrating mechanisms: a direct role for thylakoid lumen acidification? Plant Cell & Environment *20*, 147–154. https://doi.org/10.1046/j.1365-3040.1997.d01-67.x.
76. Terashima, M., Petroutsos, D., Hüdig, M., Tolstygina, I., Trompelt, K., Gäbelein, P., Fufezan, C., Kudla, J., Weinl, S., Finazzi, G., et al. (2012). Calcium-dependent regulation of cyclic photosynthetic electron transfer by a CAS, ANR1, and PGRL1 complex. Proc. Natl. Acad. Sci. USA *109*, 17717–17722. https://doi.org/10.1073/pnas.1207118109.
77. Mesnage, S., Dellarole, M., Baxter, N.J., Rouget, J.-B., Dimitrov, J.D., Wang, N., Fujimoto, Y., Hounslow, A.M., Lacroix-Desmazes, S., Fukase, K., et al. (2014). Molecular basis for bacterial peptidoglycan recognition by LysM domains. Nat. Commun. *5*, 4269. https://doi.org/10.1038/ncomms5269.
78. Hirano, T., Tanidokoro, K., Shimizu, Y., Kawarabayasi, Y., Ohshima, T., Sato, M., Tadano, S., Ishikawa, H., Takio, S., Takechi, K., et al. (2016). Moss chloroplasts are surrounded by a peptidoglycan wall containing D-amino acids. Plant Cell *28*, 1521–1532. https://doi.org/10.1105/tpc.16.00104.
79. Carrie, C., and Whelan, J. (2013). Widespread dual targeting of proteins in land plants: when, where, how and why. Plant Signal. Behav. *8*, e25034. https://doi.org/10.4161/psb.25034.
80. Krupinska, K., Blanco, N.E., Oetke, S., and Zottini, M. (2020). Genome communication in plants mediated by organelle–nucleus-located proteins. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190397. https://doi.org/10.1098/rstb.2019.0397.
81. Isemer, R., Krause, K., Grabe, N., Kitahata, N., Asami, T., and Krupinska, K. (2012a). Plastid located WHIRLY1 enhances the responsiveness of Arabidopsis seedlings toward abscisic acid. Front. Plant Sci. *3*, 283. https://doi.org/10.3389/fpls.2012.00283.
82. Isemer, R., Mulisch, M., Schäfer, A., Kirchner, S., Koop, H.-U., and Krupinska, K. (2012b). Recombinant Whirly1 translocates from transplastomic chloroplasts to the nucleus. FEBS Lett. *586*, 85–88. https://doi.org/10.1016/j.febslet.2011.11.029.
83. Carrie, C., Giraud, E., and Whelan, J. (2009). Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. FEBS Journal *276*, 1187–1195. https://doi.org/10.1111/j.1742-4658.2009.06876.x.
84. Krath, B.N., and Hove-Jensen, B. (1999). Organellar and cytosolic localization of four phosphoribosyl diphosphate synthase isozymes in spinach. Plant Physiol. *119*, 497–506. https://doi.org/10.1104/pp.119.2.497.
85. Jarvis, P., and Robinson, C. (2004). Mechanisms of protein import and routing in chloroplasts. Curr. Biol. *14*, R1064–R1077. https://doi.org/10.1016/j.cub.2004.11.049.
86. Ishida, H., Yoshimoto, K., Izumi, M., Reisen, D., Yano, Y., Makino, A., Ohsumi, Y., Hanson, M.R., and Mae, T. (2008). Mobilization of Rubisco and stroma-localized fluorescent proteins of chloroplasts to the vacuole by an ATG gene-dependent autophagic process. Plant Physiol. *148*, 142–155. https://doi.org/10.1104/pp.108.122770.
87. Wolfe, G.R., Park, H., Sharp, W.P., and Hoober, J.K. (1997). Light-harvesting complex apoproteins in cytoplasmic vacuoles in *Chlamydomonas reinhardtii* (Chlorophyta). J. Phycol. *33*, 377–386. https://doi.org/10.1111/j.0022-3646.1997.00377.x.
88. Sontag, E.M., Samant, R.S., and Frydman, J. (2017). Mechanisms and functions of spatial protein Quality Control. Annu. Rev. Biochem. *86*, 97–122. https://doi.org/10.1146/annurev-biochem-060815-014616.
89. Aksoy, M., Pootakham, W., and Grossman, A.R. (2014). Critical function of a *Chlamydomonas reinhardtii* putative polyphosphate polymerase subunit during nutrient deprivation. Plant Cell *26*, 4214–4229. https://doi.org/10.1105/tpc.114.129270.
90. Docampo, R., de Souza, W., Miranda, K., Rohloff, P., and Moreno, S.N.J. (2005). Acidocalcisomes conserved from bacteria to man. Nat. Rev. Microbiol. *3*, 251–261. https://doi.org/10.1038/nrmicro1097.
91. Komine, Y., Eggink, L.L., Park, H., and Hoober, J.K. (2000). Vacuolar granules in *Chlamydomonas reinhardtii*: polyphosphate and a 70-kDa polypeptide as major components. Planta *210*, 897–905. https://doi.org/10.1007/s004250050695.
92. Ruiz, F.A., Marchesini, N., Seufferheld, M., Govindjee, and Docampo, R. (2001). The polyphosphate bodies of *Chlamydomonas reinhardtii* possess a proton-pumping pyrophosphatase and are similar to acidocalcisomes. J. Biol. Chem. *276*, 46196–46203. https://doi.org/10.1074/jbc.M105268200.
93. Huang, G., Ulrich, P.N., Storey, M., Johnson, D., Tischer, J., Tovar, J.A., Moreno, S.N.J., Orlando, R., and Docampo, R. (2014). Proteomic analysis of the acidocalcisome, an organelle conserved from bacteria to human cells. PLoS Pathog. *10*, e1004555. https://doi.org/10.1371/journal.ppat.1004555.
94. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.
95. Steinegger, M., Mirdita, M., and Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods *16*, 603–606. https://doi.org/10.1038/s41592-019-0437-4.
96. Tsai, Y.C., Mueller-Cajar, O., Saschenbrecker, S., Hartl, F.U., and Hayer-Hartl, M. (2012). Chaperonin cofactors, Cpn10 and Cpn20, of green algae and plants function as hetero-oligomeric ring complexes. J. Biol. Chem. *287*, 20471–20481. https://doi.org/10.1074/jbc.M112.365411.
97. Wietrzynski, W., Traverso, E., Wollman, F.A., and Wostrikoff, K. (2021). The state of oligomerization of RuBisCO controls the rate of synthesis of the RuBisCO large subunit in *Chlamydomonas reinhardtii*. Plant Cell *33*, 1706–1727. https://doi.org/10.1093/plcell/koab061.
98. Mackinder, L.C.M., Meyer, M.T., Mettler-Altmann, T., Chen, V.K., Mitchell, M.C., Caspari, O., Freeman Rosenzweig, E.S., Pallesen, L., Reeves, G., Itakura, A., et al. (2016). A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. Proc. Natl. Acad. Sci. USA *113*, 5958–5963. https://doi.org/10.1073/pnas.1522866113.
99. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.
100. Kropat, J., Hong-Hermesdorf, A., Casero, D., Ent, P., Castruita, M., Pellegrini, M., Merchant, S.S., and Malasarn, D. (2011). A revised mineral nutrient supplement increases biomass and growth rate in *Chlamydomonas reinhardtii*. Plant J. *66*, 770–780. https://doi.org/10.1111/j.1365-313X.2011.04537.x.
101. Li, X., Patena, W., Fauser, F., Jinkerson, R.E., Saroussi, S., Meyer, M.T., Ivanova, N., Robertson, J.M., Yue, R., Zhang, R., et al. (2019). A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. Nat. Genet. *51*, 627–635. https://doi.org/10.1038/s41588-019-0370-6.
102. Allmer, J., Naumann, B., Markert, C., Zhang, M., and Hippler, M. (2006). Mass spectrometric genomic data mining: novel insights into bioenergetic pathways in *Chlamydomonas reinhardtii*. Proteomics *6*, 6207–6220. https://doi.org/10.1002/pmic.200600208.
103. Cross, F.R. (2015). Tying down loose ends in the Chlamydomonas genome: functional significance of abundant upstream open reading frames. G3 (Bethesda) *6*, 435–446. https://doi.org/10.1534/g3.115.023119.

104. Yamano, T., Iguchi, H., and Fukuzawa, H. (2013). Rapid transformation of *Chlamydomonas reinhardtii* without cell-wall removal. J. Biosci. Bioeng. *115*, 691–694. https://doi.org/10.1016/j.jbiosc.2012.12.020.

105. Neupert, J., Karcher, D., and Bock, R. (2009). Generation of Chlamydomonas strains that efficiently express nuclear transgenes. Plant J. *57*, 1140–1150. https://doi.org/10.1111/j.1365-313X.2008.03746.x.

106. Wang, L., Yamano, T., Takane, S., Niikawa, Y., Toyokawa, C., Ozawa, S.I., Tokutsu, R., Takahashi, Y., Minagawa, J., Kanesaki, Y., et al. (2016). Chloroplast-mediated regulation of CO_2-concentrating mechanism by Ca^{2+}-binding protein CAS in the green alga *Chlamydomonas reinhardtii*. Proc. Natl. Acad. Sci. USA *113*, 12586–12591. https://doi.org/10.1073/pnas.1606519113.

107. Picariello, T., Hou, Y., Kubo, T., McNeill, N.A., Yanagisawa, H.A., Oda, T., and Witman, G.B. (2020). TIM, a targeted insertional mutagenesis method utilizing CRISPR/Cas9 in Chlamydomonas reinhardtii. PLoS One *15*, e0232594. https://doi.org/10.1371/journal.pone.0232594.

108. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). HuggingFace's transformers: state-of-the-art natural language processing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.03771.

109. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat. Protoc. *1*, 2856–2860. https://doi.org/10.1038/nprot.2006.468.

110. Sainsbury, F., Thuenemann, E.C., and Lomonossoff, G.P. (2009). pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. Plant Biotechnol. J. *7*, 682–693. https://doi.org/10.1111/j.1467-7652.2009.00434.x.

111. Dorfer, V., Strobl, M., Winkler, S., and Mechtler, K. (2021). MS Amanda 2.0: advancements in the standalone implementation. Rapid Commun. Mass Spectrom. *35*, e9088. https://doi.org/10.1002/rcm.9088.

112. Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

113. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. *75*, 4646–4658. https://doi.org/10.1021/ac0341261.

114. Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389–403. https://doi.org/10.1016/j.cell.2009.04.042.

115. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat anti-Rabbit IgG(H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Invitrogen	Cat# A11034; RRID: AB_2576217
Rabbit polyclonal anti-Cre24.g755197	This paper	RRID: AB_2941034
Rabbit polyclonal anti-Cre06.g278195	This paper	RRID: AB_2941035
Rabbit polyclonal anti-Cre12.g519900	This paper	RRID: AB_2941036
Rabbit polyclonal anti-Cre09.g394550	This paper	RRID: AB_2941037
Rabbit polyclonal anti-Cre01.g047800	This paper	RRID: AB_2941038
Rabbit polyclonal anti-Cre06.g278150	This paper	RRID: AB_2941039
Rabbit polyclonal anti-Cre08.g372950	This paper	RRID: AB_2941040
Rabbit polyclonal anti-Cre03.g172850	This paper	RRID: AB_2941041
Rabbit polyclonal anti-Cre16.g665650	This paper	RRID: AB_2941042
Rabbit polyclonal anti-Cre01.g028150	This paper	RRID: AB_2941043
Rabbit polyclonal anti-Cre01.g013150	This paper	RRID: AB_2941044
Rabbit polyclonal anti-Cre01.g050950	This paper	RRID: AB_2941045
Rabbit polyclonal anti-Cre07.g344550	This paper	RRID: AB_2941046
Rabbit polyclonal anti-Cre24.g149250	This paper	RRID: AB_2941047
Rabbit polyclonal anti-Cre12.g519350	This paper	RRID: AB_2941048
Rabbit polyclonal anti-Cre03.g172550	This paper	RRID: AB_2941049
Rabbit polyclonal anti-Cre01.g019250	This paper	RRID: AB_2941050
Rabbit IgG (H+L) in Goat, polyclonal secondary HRP	Invitrogen	Cat# PI31466; RRID: AB_228341
Chemicals, Peptides, and Recombinant proteins		
Digitonin, Water soluble	Research Products International	Cat# 11024-24-1
cOmplete, EDTA-free Protease inhibitor	Roche	Cat# 5056489001
Anti-FLAG M2 Magnetic Beads	Sigma-Aldrich	Cat# M8823
3×FLAG peptide	Sigma-Aldrich	Cat# F4799
4×Laemmli sample buffer	Bio-Rad	Cat# 1610747
Dimethyl sulfoxide	Sigma-Aldrich	Cat# 67-68-5
Betaine	Sigma-Aldrich	Cat# 107-43-7
Formaldehyde solution 4 %, pH6.9	Sigma-Aldrich	Cat# 1004960700
Bovine Serum Albumin	Sigma-Aldrich	Cat# A3059
VECTASHIELD Mounting Medium	Vector Laboratories	Cat# H-1300
UltraPure Low-Melting Point Agarose	Invitrogen	Cat# 16500100
4-15 % Criterion TGX Precast Midi Protein Gel	Bio-Rad	Cat# 5671084
MAX efficiency transformation Reagent for Algae	Invitrogen	Cat# A24229
DL-Dithiothreitol	Sigma-Aldrich	Cat# D0632
Trypsin Gold, Mass Spectrometry Grade	Promega	Cat# V5280
Critical Commercial Assays		
Phusion High-Fidelity DNA polymerase	New England BioLabs	Cat# M0530L
Alt-R S.p. Cas9 Nuclease V3	Integrated DNA Technologies	Cat# 1081058
MinElute Gel Extraction Kit	QIAGEN	Cat# 28606
Gibson Assembly Master Mix	New England BioLabs	Cat# E2611L
Invitrogen Gateway Clonase II	Invitrogen	Cat#11791020
QIAprep Spin Miniprep Kit	QIAGEN	Cat# 27106

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
C.reinhardtii: wild-type CC-4453	Chlamydomonas Resource Center	CC-4533 cw15
E. coli Stellar Competent Cells	Takara	Cat# 636763
Chlamydomonas strains expressing tagged proteins listed in Table S6	This paper, Chlamydomonas Resource Center	https://www.chlamycollection.org/
Oligonucleotides and Recombinant DNA		
pLM005	Mackinder et al.[98]; GenBank	KX077945.1
pLM006	Mackinder et al.[98]; GenBank	KX077949.1
pENTR223	ABRC	https://www.arabidopsis.org/
pEarleyGate 103	ABRC	https://www.arabidopsis.org/
Mitochondria marker (CD3-991)	ABRC	https://www.arabidopsis.org/
Plasmid constructs generated and listed in Table S1	This paper, Chlamydomonas Resource	https://www.chlamycollection.org/
Software and Algorithms		
Fiji	Schindelin et al.[99]	https://imagej.net/software/fiji/downloads
Thermo Proteome Discoverer 2.5	Thermo Scientific	Cat# OPTON-30945
Scaffold 5	Proteome Software	https://www.proteomesoftware.com/products/scaffold-5
Others		
Electroporation Cuvette, 2mm gap	Bulldog Bio.	Cat# 12358-346
Ibidi USA μ–Slide 8 well, Glass bottom	Ibidi	Cat# NC0704855
Poly-L-lysine coated glass slides	Sigma-Aldrich	Cat# P0425
Kontes Duall #22 homogenizer	Kimble	Cat# KT885450-0022
Immobilon-P PVDF membrane	Mllipore	Cat# IPVH00010
Avanti J-26X with 8.1000 rotor	Beckman Coulter	N/A
Cryomill	Retsch	Part NO. 20.749.0001
Electroporator	NEPA GENE	NEPA21 type II
SP5 Confocal Microscope	Leica	TCS SP5
Singer Rotor HAD	Singer Instruments	Cat# ROT-001
Typhoon FLA9500 fluorescence scanner	GE Healthare	N/A
Colony picker	Norgren Systems	Part NO.12461
Nikon Confocal laser scanning microscope	Nikon	A1Rsi
Nikon Confocal microscope	Nikon	A1R-STED

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Martin C. Jonikas (mjonikas@princeton.edu).

Materials availability

- Protein localization images are available at https://www.chlamylibrary.org/.
- Fluorescently tagged strains and plasmid constructs are available at https://www.chlamycollection.org/
- All unique/stable reagents generated in this study are available from the lead contact upon request.

Data and code availability

- The manuscript reports code for a machine-learning classifier of protein localization. As described in the STAR Methods, our script for training and evaluation is available here: https://github.com/clairemcwhite/transformer_infrastructure/hf_classification.py
- The trained model files are available here: https://huggingface.co/wpatena/PB-Chlamy/tree/main.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strains and culture conditions

The *Chlamydomonas reinhardtii* strain CC-4533 (cMJ030) was used for wild-type (hereafter WT) in all experiments. All strains were maintained on Tris-acetate-phosphate (TAP) solid medium with 1.5 % agar at 22 °C under dim light (<10 μmol photons m^{-2} s^{-1}). All media used revised trace element solution.[100]

METHOD DETAILS

Target genes selection

Target genes were selected from seven sources (Figure S1A), including 1,093 genes encoding proteins identified in Arabidopsis chloroplast proteomics,[12] 644 genes encoding proteins identified in Chlamydomonas chloroplast proteomics,[13] 154 genes encoding proteins identified in Chlamydomonas pyrenoid proteomics,[39] 3,317 genes encoding PredAlgo predicted chloroplast proteins and 858 genes encoding proteins with low PredAlgo score in non-chloroplast organelles,[16] 510 genes encoding GreenCut2 proteins,[14] and 303 genes encoding candidate proteins required for photosynthesis suggested in mutant screening.[101] In addition, we also selected 777 genes because of their potential association with chloroplast function suggested either in their Phytozome annotation (https://phytozome-next.jgi.doe.gov/) or in related reports, such as the TEF proteins present in thylakoid enriched fraction[102] and FTT proteins interacting with well-known chloroplast proteins.[28] To avoid duplicating effort, we removed the overlapping genes across the seven sources above and genes encoding proteins which had been localized in Mackinder et al.[28] Altogether, we obtained 5,874 target genes.

Plasmid Construction and Cloning

We designed our primers according to gene sequences present in the v5.5 *Chlamydomonas reinhardtii* genome. Cross et al.[103] identified upstream ATGs in many of these gene sequences, and supplementary data in Mackinder et al.[28] indicate that for genes that include such upstream ATGs, using the original ATG leads to lower localization success rates, suggesting that the Cross et al.[103] upstream ATGs more frequently correspond to the native translation start site. Therefore, wherever an upstream ATG had been identified by Cross et al.,[103] we used this ATG instead of the one annotated in the genome, leading to our usage of a corrected upstream ATG in 1,213 of our target genes (Table S1).

The cloning pipeline was based on that used in Mackinder et al.,[28] with some modifications. The open reading fames were amplified from Chlamydomonas WT genomic DNA by PCR using Phusion High-Fidelity DNA polymerase (New England BioLabs) with additives of 6 % DMSO (v/v) (Sigma-Aldrich) and 1 M Betaine (Sigma-Aldrich). The PCR products were gel purified using MinElute Gel Extraction Kit (QIAGEN) and then cloned in-frame with a C-terminal Venus-3×FLAG in pLM005 by Gibson assembly (New England BioLabs). Primers were designed to amplify the open reading frame until but excluding the stop codon, and with adaptors to allow efficient assembly into *HpaI*-cut pLM005. Considering the PCR limitations to amplification of large genes, we mainly focused on genes smaller than 8 kb in this study. For genes larger than 6 kb, we split them into multiple fragments (< 3 kb) for PCR amplification and then reassembled the fragments together during the final Gibson assembly step. The fragment size was verified by restriction enzyme digestion. A pilot study showed that 334/334 (100 %) of genes had correct junctions as verified by Sanger sequencing.

Cloning of Chlamydomonas genes is known to be challenging due to high GC content, repetitive sequences, and gene length.[28] In total, we successfully cloned 3,116 genes (53 %) (Figure S1A), a similar fraction to the 48 % in Mackinder et al.[28] Interestingly, the cloning success of genes smaller than 500 bp is 66.2 %, which is lower than 86.5 % of genes with size between 1,000∼2,000 bp (Figure S1H).

Chlamydomonas transformation

Constructs were linearized by *EcoRV*, *DraI*, *AflII*, or *BsaI* prior to the electroporation into WT Chlamydomonas strain CC-4533. WT cells were pre-cultured in TAP liquid medium at 22 °C under light with a photon flux density of 150 μmol photons m^{-2} s^{-1} until the cell density reached to ∼2 ×10^6 cells mL-1.For each transformation, 150 ng of cut plasmid was mixed with 60 μL of 2×10^8 cells mL^{-1} suspended in MAX Efficiency Transformation reagent (Invitrogen) in an ice-cold 0.2 cm gap electroporation cuvette (Bulldog Bio.) and transformed into WT strains by electroporation using a NEPA21 electroporator (NEPA GENE).[104] The settings were: Poring Pulse: 250.0 Volts, 8.0 ms pulse length, 50.0 ms pulse interval, 2 pulses, 10 % decay rate, + polarity; Transfer Pulse: 20.0 Volts, 50.0 ms pulse length), 50.0 ms pulse interval), 10 pulses, 40 % decay rate), +/- polarity. For recovery, cells were transferred to 10 mL TAP liquid medium plus 40 mM sucrose and incubated with gentle shaking under dim light (<10 μmol photons m^{-2} s^{-1}) overnight. The transformants were plated on TAP agar medium supplied with 20 μg mL^{-1} paromomycin. After 7 days incubation under dim light (<10 μmol photons m^{-2} s^{-1}), 48 transformants from each plate were arrayed on a new rectangular TAP agar PlusPlate (Singer Instruments) using a colony Picker (Norgren Systems). The transformants were replicated manually onto a fresh TAP agar PlusPlate using a 96-Long pin pad (Singer Instruments). The TAP plates with arrayed transformants were screened for fluorescence using a Typhoon FLA9500 fluorescence scanner (GE Healthcare) with the following settings: Venus, 532 nm excitation with 555/20 nm emission. The colonies with positive fluorescence signals were isolated and maintained in 96 arrays using a Singer Rotor propagation robot (Singer Instruments).

Transformation of constructs and localization of proteins in Chlamydomonas are known to be inefficient,[28] possibly due to several mechanisms that fight foreign DNA.[37,105] Our transformation and localization success rate (34 %) was lower than that in Mackinder

et al.[28] (49 %), possibly because the genes targeted in the present study were overall expressed at lower levels. To generate dual-tag lines, pLM006 harboring an mCherry-6×HIS tag was used as the backbone, and TAP agar medium supplied with 20 μg mL^{-1} hygromycin was used for selection.

Confocal Microscopy

For confocal imaging, colonies were transferred to a 96-well microtiter plate with 100 μL TP liquid medium and 5 μg mL^{-1} antibiotics in each well and then pre-cultured in air under 150 μmol photons m^{-2} s^{-1} on an orbital shaker with gentle agitation of 600 RPM. After ~16 hr of growth, 10 μL cells were transferred onto an μ-Slide 8-well glass-bottom plate (Ibidi) and 200 μL of 1 % TP low-melting-point agarose at ~35 °C was overlaid to restrict cell movement. All imaging except for Fluorescence Recovery After Photobleaching (FRAP) assays was conducted using a Leica SP5 confocal microscope with the following settings: Venus, 514 nm excitation with 530/10 nm emission; mCherry, 561 nm excitation with 610/30 nm emission; and chlorophyll, 514 nm excitation with 685/40 nm emission. All confocal microscopy images were analyzed using Fiji.[99] For each strain, a confocal section through a cell showing the predominant localization pattern was captured and analyzed. To minimize the bias in determining the localization patterns, each localization image was independently analyzed by two researchers. Localization patterns for 31 proteins where there was clear disagreement or insufficient signal were categorized as Ambiguous.

FRAP assays were performed using a Nikon A1R-STED confocal microscope with the following setting: Venus 514 nm excitation with 530/10 nm emission; and chlorophyll, 514 excitation with 685/40 nm emission. One baseline image was acquired before FRAP was performed. The selected puncta were bleached by a high-intensity laser beam (514 nm wavelength). The recovery of fluorescence at the bleached puncta was imaged every 30 s for 10 min.

Indirect Immunofluorescence Assay

Indirect immunofluorescence was performed as described previously.[106] Briefly, Cells were harvested by centrifugation and rinsed with PBS buffer twice. Then 100 μL of cells was spotted onto Poly-L-lysine coated glass slides (Sigma-Aldrich). Cells were fixed with 4 % (w/v) formaldehyde (Sigma-Aldrich) in PBS for 20 min and then incubated with 100 % ice-cold methanol for 20 min to remove chlorophyll. Purified antibodies (Yenzyme) against Cre01.g028150, Cre01.g013150, Cre12.g519350, Cre03.g172550, Cre01.g019250, Cre07.g344550, Cre06.g278195, Cre24.g755197, Cre24.g149250, Cre12.g519900, Cre09.g394550, Cre01.g047800, Cre06.g278150, Cre08.g372950, Cre03.g172850, Cre16.g665650, Cre01.g028150, Cre01.g013150, and Cre01.g050950 were used at a dilution of 1:200. The purified antibodies were generated using the following peptides: C-Ahx-PDQPPRILTTRRE-amide (Cre01.g028150), C-Ahx-TWDVKAPINKHYNFH-cooh (Cre01.g013150), C-Ahx-YLPNTGNMLMQVNPNQ-cooh (Cre12.g519350), C-Ahx-RGQVKNTQQYRMR-cooh (Cre03.g172550), C-Ahx-KGVDATKYSHSTIVQT-amide (Cre01.g019250), C-Ahx-NKYTEEVIYKEENMDYDK-amide (Cre07.g344550), C-Ahx-RLKLPRFLEDEQPKDPEQTKQD-amide (Cre06.g278195), C-Ahx-RGNKLTRNDPKSMVTRKE-amide (Cre24.g755197), C-Ahx-GADPNYRRNMSLR-amide (Cre24.g149250), C-Ahx-QDQGRTPLWFEPYLQKAQRDNKR-amide (Cre12.g519900), C-Ahx-YPDEGAWKRFHYQFKSEGYPE-amide (Cre09.g394550), C-Ahx-DRIVKENPIVVFVKGTRQQP-amide (Cre01.g047800), C-Ahx-PGVNLSQLVAKHPRLLSEYR-amide (Cre06.g278150), C-Ahx-HKLAHGELKETTGWLNPGKP-amide (Cre08.g372950), C-Ahx-NSDEPKYVKNDK-amide (Cre03.g172850), C-Ahx-RTHRRQYREKRSSTP-amide (Cre16.g665650), and C-Ahx-SVNSKSVNVSFGSKANEER-amide (Cre01.g050950). After washing the slides 4 times, each with 50 mL PBS-T (supplied with 0.1% Tween 20 (v/v)) in Coplin jar, Alexa Fluor 488 goat anti-rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody (Invitrogen) was used at a dilution of 1:500. The slides were washed 4 times, each with 50 mL PBS-T. Fluorescence and bright-field images were acquired using a confocal microscope (Leica, SP5).

Immunoblotting Analysis

Total protein was extracted from wild-type cells by sonication followed by centrifugation at 18,000 g for 10 min. 30 μL of supernatant was mixed with 9.75 μL 4×SDS-PAGE buffer (Bio-Rad) containing 100 mM DTT (Sigma-Aldrich) followed by denaturation by heating at 70 °C for 10 min. Then, 34 μL denatured protein sample was loaded into a well of a 4-15 % Criterion TGX Precast Midi Protein Gel (BioRad) for electrophoresis at 0.06 A for 60 min until the protein front reached to the gel bottom. Next, proteins were transferred to Immobilon-P PVDF membrane (Millipore) using a semidry blotting system. Membranes were blocked with 5% (w/v) Non-fat Dry Milk (LabScientific) in TBS-T buffer which contained 0.1% (v/v) Tween 20 (Sigma-Aldrich). Blocked membranes were then washed with TBS-T and treated with the following antibodies: anti-Cre03.g172550 (1:5,000 dilution), anti-Cre12.g519350 (1:5,000), anti-Cre01.g019250 (1:5,000), anti-Cre07.g344550 (1:5,000), anti-Cre06.g278195 (1:5,000), anti-Cre09.g394550 (1:5,000), anti-Cre08.g372950 (1:5,000), anti-Cre03.g172850 (1:5,000), anti-Cre16.g665650 (1:5,000), anti-Cre01.g013150 (1:5,000), anti-Cre24.g755197 (1:200), anti-Cre03.g149250 (1:200), anti-Cre12.g519900 (1:5,000), anti-Cre01.g047800 (1:5,000), anti-Cre06.g278150 (1:5,000), and anti-Cre01.g028150 (1:200). To recognize the primary antibody, a horseradish peroxidase-conjugated goat anti-rabbit IgG antibody (Invitrogen) was used as a secondary antibody in a dilution of 1:7,500.

Mutant generation by CRISPR-Cas9

The *chlp1* and *icl2* insertional mutants were generated using CRISPR-Cas9 system as described previously[107] with modifications. We designed the guide RNA sequences using CHOPCHOP (https://chopchop.cbu.uib.no/). The single guide RNAs (sgRNAs)

were generated to target the following sequences: AGACCCTCAGCGAGACCGAGTGG (Cre01.g050950,CHLP1) and CATGAGAAA CAGCCGCACCTCGG (Cre03.g149250, ICL2). To assemble Cas9-guide RNA ribonucleoprotein (RNP), we incubated 24 pmol of each Alt-R CRISPR-Cas9 sgRNA (Integrated DNA Technologies) with 61 pmol Alt-R S.p. Cas9 Nuclease V3 (Integrated DNA Technologies) at 37 °C for 30 min. WT or CC-125 cells were pre-cultured in TAP liquid medium at 22 °C under light with a photon flux density of 150 μmol photons m^{-2} s^{-1} until the cell density reached ~5 ×10^6 cells mL^{-1}. We then harvested cells, incubated the cells in autolysin medium for 1.5 h in room temperature, and resuspend the cells in ½ TAP+80 mM sucrose. For each reaction, the RNP together with 1 μg donor DNA containing *AphII* cassette was mixed with 125 μL autolysin-treated cells of 2×10^8 cells mL^{-1} and delivered by electroporation using a NEPA21 electroporator (NEPA GENE) in a 0.2 cm gap electroporation cuvette (Bulldog Bio.). The settings were: Poring Pulse: 250.0 Volts, 8.0 ms pulse length, 50.0 ms pulse interval, 2 pulses, 10 % decay rate, + polarity; Transfer Pulse: 20.0 Volts, 50.0 ms pulse length), 50.0 ms pulse interval), 10 pulses, 40 % decay rate), +/- polarity. For recovery, cells were transferred to 10 mL TAP liquid medium plus 80 mM sucrose and incubated with gentle shaking under dim light (<10 μmol photons m^{-2} s^{-1}) overnight. The mutant candidates were plated on TAP agar medium supplied with 10 μg mL^{-1} Hygromycin and incubated under dim light (<10 μmol photons m^{-2} s^{-1}) for 7 days until the colonies reached reasonable size. Then, the mutant candidates were screened by PCR amplification, confocal imaging, or growth test. The primers sets for mutant verification were as follows: P1, 5-GACCTGCCCATGGAGATTATT-3; P2, 5-TCGTAGCTGTTGTAGTGGATGG-3; P3, 5-GTTTATCAATTGAGCTTGTGCG-3; P4, 5-ACCAGCAGGATGTCGTTACC-3.

Protein localization prediction

For each subcellular localization, we trained a protein language model to predict protein localization from protein sequence. Protein language models are first trained on large numbers of sequences, and then these pretrained models can be retrained for a specific prediction task, in this case subcellular location prediction (Figure 7A). For our pretrained model, we used ProtBertBFD, a protein language model pre-trained on billions of protein sequences[94] (https://huggingface.co/Rostlab/prot_bert_bfd). Given a protein sequence, ProtBertBFD outputs numeric vectors, or embeddings, that capture features of each amino acid in that sequence, as well as an embedding that represents the whole sequence. These amino acid embeddings contain information on biochemical and structural properties.[94] We use the sequence-level embedding (CLS) as an input to a linear classifier to distinguish if a protein is localized to a particular cellular compartment or not. Specifically, we use the model architecture BertForSequenceClassification from the huggingface python package.[108] Our script for running training and evaluation is https://github.com/clairemcwhite/transformer_infrastructure/hf_classification.py. For each compartment (chloroplast, mitochondrial and secretory) we used proteins found to localize to the compartment as positive cases, and proteins not found to localize to the compartment as negative cases. We used a random 60% of these positive and negative cases to train the model, 20% for performance validation during training, and 20% as a fully withheld test set to evaluate model performance on unseen examples. These sets are listed in Table S7.

The raw score distributions, PR and ROC curves and summary measures compared to PredAlgo are shown in Figure S6; for the purpose of comparisons with PredAlgo, we used the testing sets minus any proteins that were included in PredAlgo training data. In Figure 7B, the measure displayed is precision (what % of predicted positives are correct). Each separate localization predictor had its own test set, with sizes as follows: 111 chloroplast and 97 non-chloroplast proteins, 12 mitochondrial and 194 non-mitochondrial, 18 secretory and 149 non-secretory. The error bars are 90% confidence intervals; their large size for mitochondrial and secretory data is due to those testing datasets being relatively small. We used Fisher's exact test to compare the numbers of correct and incorrect positives between PB-Chlamy and PredAlgo; the results were p=0.000036 for chloroplast, p=0.010 for mitochondrial, and p=0.033 for secretory data. We then used the trained models to predict protein localizations for the entire Chlamydomonas proteome (Table S7).

We downloaded Chlamydomonas protein sequences from Phytozome (https://phytozome-next.jgi.doe.gov/info/Creinhardtii_v5_6, genome version 5.6); we only used primary transcripts for training and for localization prediction. We adjusted the protein sequences to use the new start codons described by Cross et al.[103] - they are included in Table S7.

The training command and environment setup for the chloroplast were as follows, with analogous commands for the other localizations:

```
module load cudatoolkit
bash make_hf-transformers_conda_env.sh
conda activate hf-transformers
python hf_classification.py -m prot_bert_bfd/ -tr chloro_train.csv -v chloro_val.csv -te chloro_test.csv -o results_chloro -maxl 1150 -n chloro -e 10 -tbsize 1 -vbsize 1 -s 3
```

The input files containing the training/validation/test sets were plaintext, formatted as follows:

```
Entry_name,sequence,label
Cre03.g155400,MHKTPCLHGGSSLSAGRAPLARLCCASQRVRGPAPAQAFWKQSGASAGKSGKARPGAKAQQPKQKAGGGKQGGGGGGGGLMDSEVPVYAEAFDINKCVDLYLRFFKWVSSPVTGGSGKK,Chloroplast
```

The trained model files are available (https://huggingface.co/wpatena/PB-Chlamy/tree/main).

Affinity Purification and Mass Spectrometry

Each affinity purification-mass spectrometry (AP-MS) experiment was performed twice from independently grown samples of the same strain. Cells expressing Venus-3×FLAG-tagged proteins were pre-cultured in 50 mL TAP medium with 5 μg mL^{-1} paromomycin

until the cell density reach to ~2-4 ×10^6 cells mL^{-1}. Then, cells were harvested by centrifugation at 1,000 g for 5 min and the pellets were suspended in 1,000 mL TP liquid medium. Cells were grown with air bubbling and constant stirring of 210 RPM under 150 μmol photons m^{-2} s^{-1} light until the cell density reached ~2-4 ×10^6 cells mL^{-1}. Cells were collected by centrifugation at 3,000 g for 4 min in an Avanti J-26X centrifuge with an 8.1000 rotor (Beckman) at 4 °C. The pellets were washed in 35 mL ice-cold washing buffer (25 mM HEPES, 25 mM KOAc, 1 mM $Mg(OAc)_2$, 0.5 mM $CaCl_2$, 100 mM Sorbitol, 1mM NaF, 0.3 mM Na_3VO_4, and cOmplete EDTA-free protease inhibitor (1 tablet/500 mL)) and then resuspended in a 1:1 (v/w) ratio of ice-cold 2×IP buffer (50 mM HEPES, 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM $Cacl_2$, 200 mM Sorbitol, 1mM NaF, 0.3 mM Na_3VO_4, and cOmplete EDTA-free protease inhibitor (1 tablet/50 mL)). 3 mL cell slurry was immediately added to liquid nitrogen to form small popcorn pellets which were stored at -80 °C until needed. Cells were lysed by cryogenic grinding using a Cryomill (Retsch) at frequency of 25 oscillations per second for 20 min. The ground powder was defrosted on ice for 45 min and dounced 25 times on ice with a Kontes Duall #22 homogenizer (Kimble). 1mL homogenized cells of each sample was used for the following processes. Membrane proteins were solubilized by incrementally adding an equal volume of ice-cold 1×IP buffer plus 2 % digitonin (RPI) followed by an incubation of 45 min with nutation at 4 °C. The cell debris were removed by spinning at 12,700 g for 30 min at 4°C. The supernatant was then mixed with 50 μL anti 3×FLAG magnetic beads (Sigma) which had been previously washed sequentially with 1×IP buffer 3 times and 1×IP buffer plus 0.1 % digitonin 2 times. The mixture was incubated with nutation at 4 °C for 1.5 hr, followed by the removal of supernatant. The beads were washed 4 times with 1×IP buffer plus 0.1 % digitonin followed by a 30 min competitive elution with 45 μL of 1×IP buffer plus 0.25 % digitonin and 2 μg/μL 3×FLAG peptide (Sigma-Aldrich). After elution, 30 μL protein samples were mixed with 9.75 μL 4×SDS-PAGE buffer (Bio-Rad) containing 100 mM DTT (Sigma-Aldrich) followed by denaturation by heating at 70 °C for 10 min. Then, 30 μL denatured protein sample was loaded into a well of a 4-15 % Criterion TGX Precast Midi Protein Gel (BioRad) for electrophoresis at 50 V for 40 min until the protein front moved ~2.5 cm. ~2.0 cm gel slice containing target proteins with molecular weight >= 10 kDa (to exclude the 3xFLAG peptide) were excised and stored at 4 °C until processing for in-gel digestion. To decrease cross-contamination from samples in neighboring wells, samples were loaded in every other well. To further avoid carry-over contamination of mass spectrometry and contamination from sequential samples, we performed two biological repeats of AP-MS and changed the order of samples in the two biological repeats.

In-gel digestion of protein bands using trypsin was performed as previously.[109] Trypsin digested samples were dried completely in a SpeedVac and resuspended with 20 μL of 0.1 % formic acid pH 3 in water. 2 μL (~ 360 ng) was injected per run using an Easy-nLC 1,200 UPLC system. Samples were loaded directly onto a 15 cm long, 75 μm inner diameter nanocapillary column packed with 1.9 μm C18-AQ resin (Dr. Maisch, Germany) mated to a metal emitter in-line with an Orbitrap Fusion Lumos (Thermo Scientific, USA). The column temperature was set at 45 °C and a half-hour gradient method with 300 nL per minute flow was used. The mass spectrometer was operated in data dependent mode with a 120,000 resolution MS1 scan (positive mode, profile data type, AGC gain of 4e5, maximum injection time of 54 s and mass range of 375-1,500 m/z) in the Orbitrap followed by HCD fragmentation in ion trap with 35 % collision energy. A dynamic exclusion list was invoked to exclude previously sequenced peptides for 60s and a maximum cycle time of 3 s was used. Peptides were isolated for fragmentation using the quadrupole (1.2 m/z isolation window). The ion trap was operated in Rapid mode.

Transient expression of Arabidopsis gene in Tobacco leaf

We first cloned the full-length cDNA of AT4g16060 into pENTR223 (ABRC). Then the AT4G16060 was cloned in-frame with a C-terminal GFP in pEarleyGate 103 by LR recombination reaction (Invitrogen Gateway Clonase II). The construct of AT4G16060-GFP and mitochondria mCherry marker CD3-991 (ABRC) were then separately transformed into Agrobacterium tumefaciens by heat shock. A mixture (OD600 = 0.125) of the Agrobacterium tumefaciens harboring At4G16060-GFP, the Agrobacterium tumefaciens carrying CD3-991, and the p19 protein of tomato bushy stunt virus (TBSV) in a 2:1:2 ratio co-infiltrated into 4-weeks old tobacco plants (*Nicotiana tabacum*).[110] Three days after infiltration, the abaxial epidermis of the leaves were imaged using Nikon A1Rsi confocal laser scanning microscopy (CLSM) with Nikon 60x Apo (NA1.40) objective. The imaging was performed with the following setting: GFP, 488 nm excitation with 525/25 nm emission; mCherry, 561 nm excitation with 595/25 nm emission; and Chlorophyll, 647 nm excitation with 699/37 nm emission. Image acquisition and analysis were performed using the Nikon NIS Elements software (version 5.21.03).

QUANTIFICATION AND STATISTICAL ANALYSIS

Peptide identification

Raw files were searched using MSAmanda 2.0[111] and Sequest HT algorithms[112] within the Proteome Discoverer 2.5.0 suite (Thermo Scientific, USA). 10 ppm MS1 and 0.4 Da MS2 mass tolerances were specified. Carbamidomethylation of cysteine was used as fixed modification, oxidation of methionine, deamidation of asparagine and glutamine were specified as dynamic modifications. Pyro glutamate conversion from glutamic acid and glutamine are set as dynamic modifications at peptide N-terminus. Acetylation was specified as dynamic modification at protein N-terminus. Trypsin digestion was selected with a maximum of 2 missed cleavages allowed. Files were searched against the UP000006906 Chlamydomonas database downloaded from Uniprot.org (https://www.uniprot.org/).

Scaffold (version Scaffold 5.1.0, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0 % probability by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at greater than 99.9 % probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm.[113]

Calculating WD-scores

The WD-scores of MS data were calculated using the ComPASS method, which analyzes spectral counts based on the specificity of the prey, spectral count number and reproducibility.[114,115] Instead of using the spectral counts from two technical repeats, we used the spectral counts from two biological replicas with different neighbors for each sample. First, we generate a Stats table containing all the bait proteins and interactors as below,

Stats Table

	Bait 1	Bait2	Bait3	Bait4	Bait *k*	
Interactor 1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	$X_{k,1}$	$\overline{X}_1$
Interactor 2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	$X_{k,2}$	$\overline{X}_2$
Interactor 3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	$X_{k,3}$	
Interactor 4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{4,4}$	$X_{k,4}$	
Interactor *m*	$X_{1,m}$	$X_{2,m}$	$X_{3,m}$	$X_{4,m}$	$X_{k,m}$	$\overline{X}_m$

$X_{i,j}$ is the average spectral counts from two biological replicas for interactor j from bait i.
m is the total number of unique prey proteins identified (11,911).
k is the total number of unique baits (67).
We calculated the WD-scores using the equations[114] below,

$$\mathrm{WDi,j} = \sqrt{\left(\frac{k}{\sum_{i=k}^{j=1} f_{i,j}}\omega_j\right)^{p}} X_{i,j} \qquad \text{(Equation 1)}$$

$$\omega_j = \left(\frac{\sigma_j}{\overline{X}_j}\right), \overline{X}_j = \frac{\sum_{i=1}^{i=k} X_{i,j}}{k}, n = 1,2,\ldots m, \begin{array}{l}\text{if } \omega_j \le 1 \Rightarrow \omega_j = 1 \\ \text{if } \omega_j > 1 \Rightarrow \omega_j = \omega_j\end{array}$$

$$f_{i,j} = \begin{cases} 1; & X_{i,j} > 0 \\ X_{i,j} & \end{cases}$$

p is the number of replicates runs in which the interactor is present $f_{i,j}$.

The minimum WD score values for high-confidence interactions will be different for each study because the WD score depends on the specific proteins and methods used in each study. In Mackinder et al.,[28] we set the high-confidence WD cut-off based on WD-scores of prey proteins that localized to a different compartment than the bait. Because the vast majority of the baits in the present study localized to the same compartment (the chloroplast) we could not use the same approach to set the WD cut-off. We therefore set our WD cutoff at 3.7 % of all interactions based on the corresponding value in Mackinder et al.[28] of 3.78 %. This rationale led to a WD cut-off for the present study of 20.367, with 411 of the 11,911 interactions above this threshold (Table S5). We defined high-confidence protein-protein interactions as those having a WD score above the cutoff of 20.367 and where the prey was detected in both biological repeats, which resulted in ~274 high-confidence protein-protein interactions (Table S5).

Data visualization

The calculation of WD-score and assembly of bait-prey matrix were performed in Microsoft Exel. The alignment of amino acid sequence was conducted using Clustal Omega with default settings (https://www.ebi.ac.uk/Tools/msa/clustalo/).

Transmembrane prediction and Protein homology prediction

Protein transmembrane domains were predicted using TMHMM2.0 (https://services.healthtech.dtu.dk/service.php?TMHMM-2.0). Protein homologies were predicted using Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The repeats in protein sequences were predicated using RADAR (https://www.ebi.ac.uk/Tools/pfa/radar/).

Statistical tests

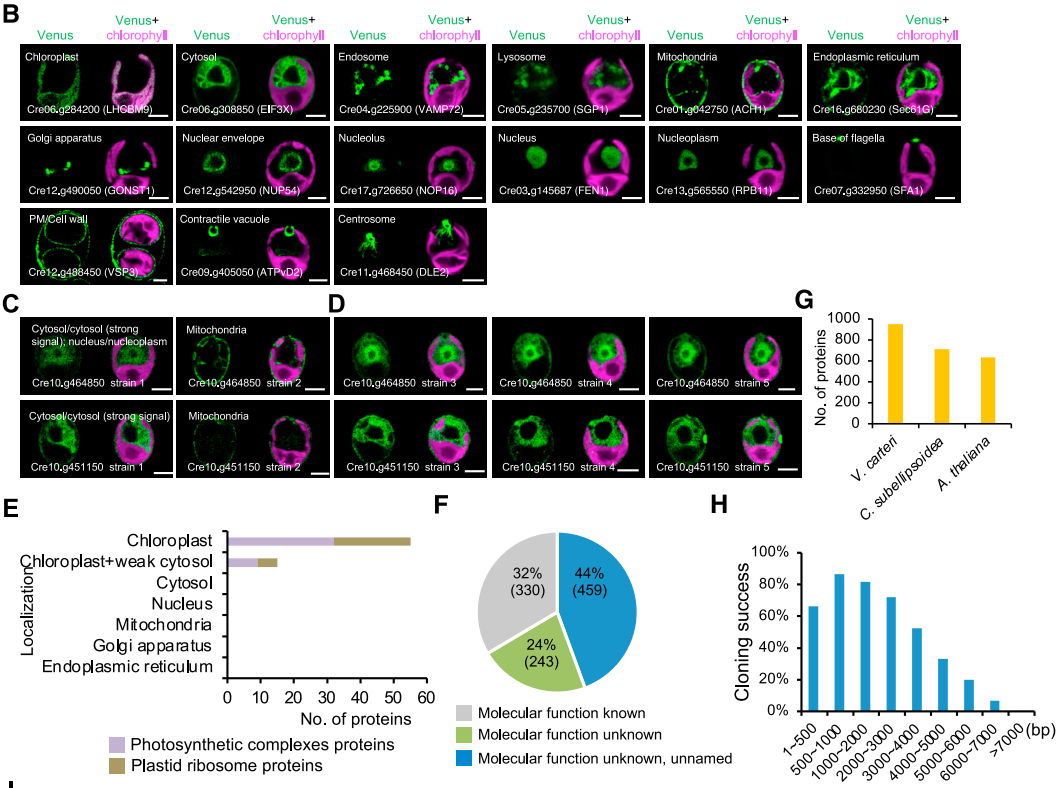
Statistical tests comparing PredAlgo and PB-Chlamy were performed in Python using scipy.stats and rpy2. All other statistical tests were performed in Microsoft Excel.

CellPress
OPEN ACCESS

Supplemental figures

A

Order	Sources of candidate	Number of candidates	Number of cloned candidates	Number of localized candidates	References
(1)	Chloroplast proteomics in *Arabidopsis thaliana*	1,093	516	295	Ferro *et al.*[12]
(2)	Chloroplast proteomics in *Chlamydomonas reinhardtii*	644	384	235	Terashima *et al.*[13]
(3)	Pyrenoid proteomics	154	80	44	Zhan *et al.*[39]
(4)	PredAlgo predicted chloroplast proteins	4,175	2,143	697	Tardif *et al.*[16]
(5)	GreenCut2 proteins	510	329	155	Karpowicz *et al.*[14]
(6)	Candidate proteins required for photosynthesis	303	132	47	Li *et al.*[101]
(7)	Others	777	554	200	-
	In total	**5,874**	**3,117**	**1,034**	-



I

Venus-tagged protein	Average number of puncta	Average distance to pyrenoid center (μm)	Average puncta size (μm, diameter)
Cre13.g576400	4.4 ± 0.5	6.11 ± 0.74	0.337 ± 0.10
Cre12.g519900	5.4 ± 0.9	6.33 ± 1.14	0.367 ± 0.10
Cre05.g234050	4.8 ± 0.8	5.00 ± 1.52	0.292 ± 0.06
Cre24.g755197	4.0 ± 0.7	5.44 ±1.63	0.246 ± 0.06
Cre13.g603500	2.8 ± 0.8	2.99 ± 0.59	0.375 ± 0.07
Cre07.g344550	2.2 ± 0.4	2.63 ± 0.70	0.237 ± 0.07
Cre01.g005534	5.0 ± 1.0	4.13 ± 0.81	0.344 ± 0.13
Cre01.g050950	26.6 ± 3.2	5.99 ± 0.83	0.166 ± 0.06
Cre03.g145507	18.6 ± 0.9	5.48 ± 0.24	0.120 ± 0.03
Cre03.g197650	19.2 ± 3.4	5.95 ± 0.52	0.115 ± 0.09
Cre06.g256850	7.4 ± 1.7	5.30 ± 0.37	0.293 ± 0.09
Cre16.g672300	7.2 ± 1.5	4.98 ± 0.90	0.313 ± 0.06
Cre06.g285401	6.8 ± 1.6	4.73 ± 0.77	0.295 ± 0.05
Cre06.g295350	1.8 ± 0.4	5.50 ± 1.94	0.264 ± 0.11
Cre06.g278195	20.0 ± 4.0	6.68 ± 2.00	0.243 ± 0.08

Figure S1. We localized 1,034 proteins from 5,874 target proteins, related to Figures 1 and 2

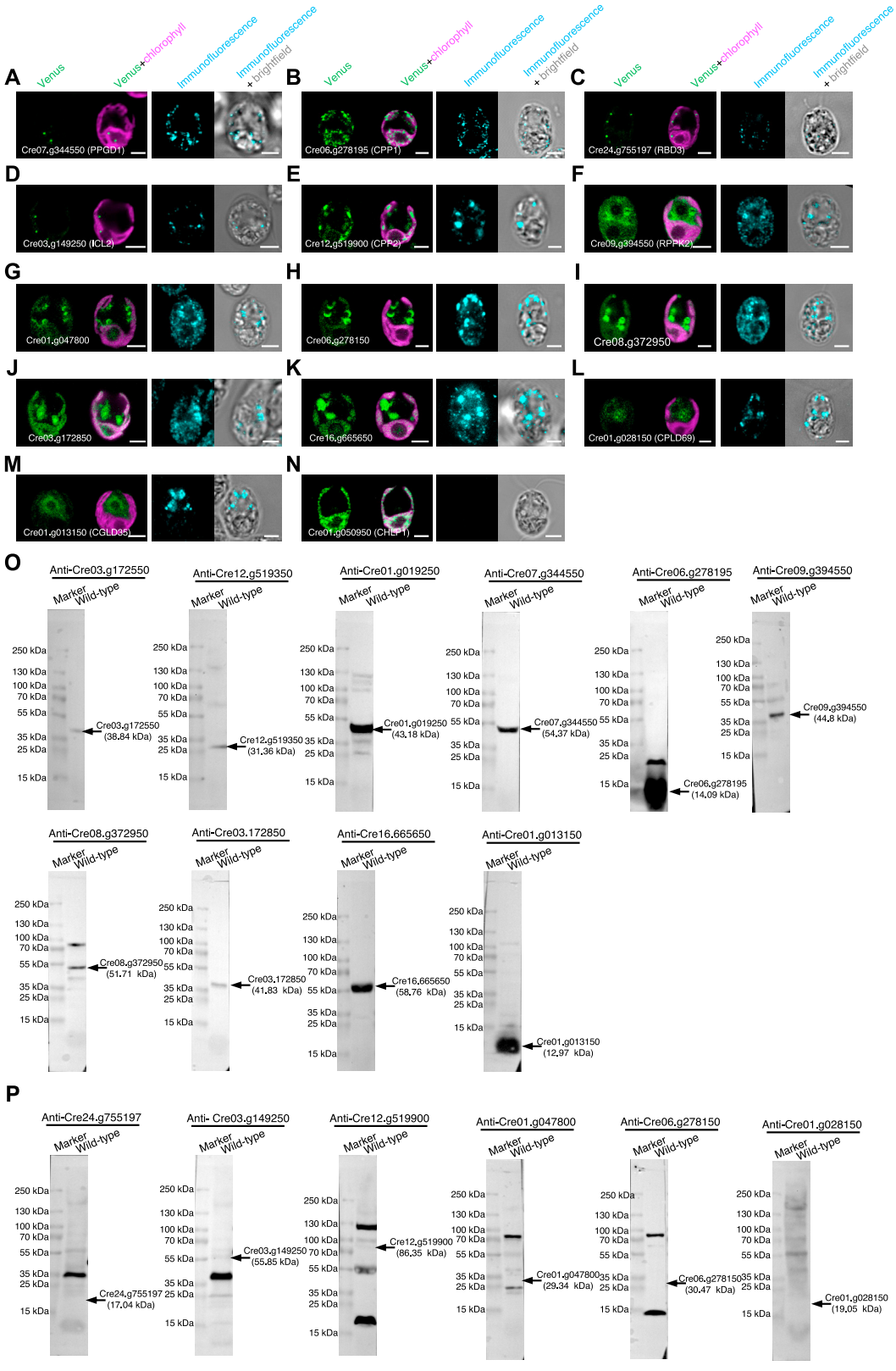


Figure S2. We verified our localization dataset using immunofluorescence, related to Figures 1, 2, and 6

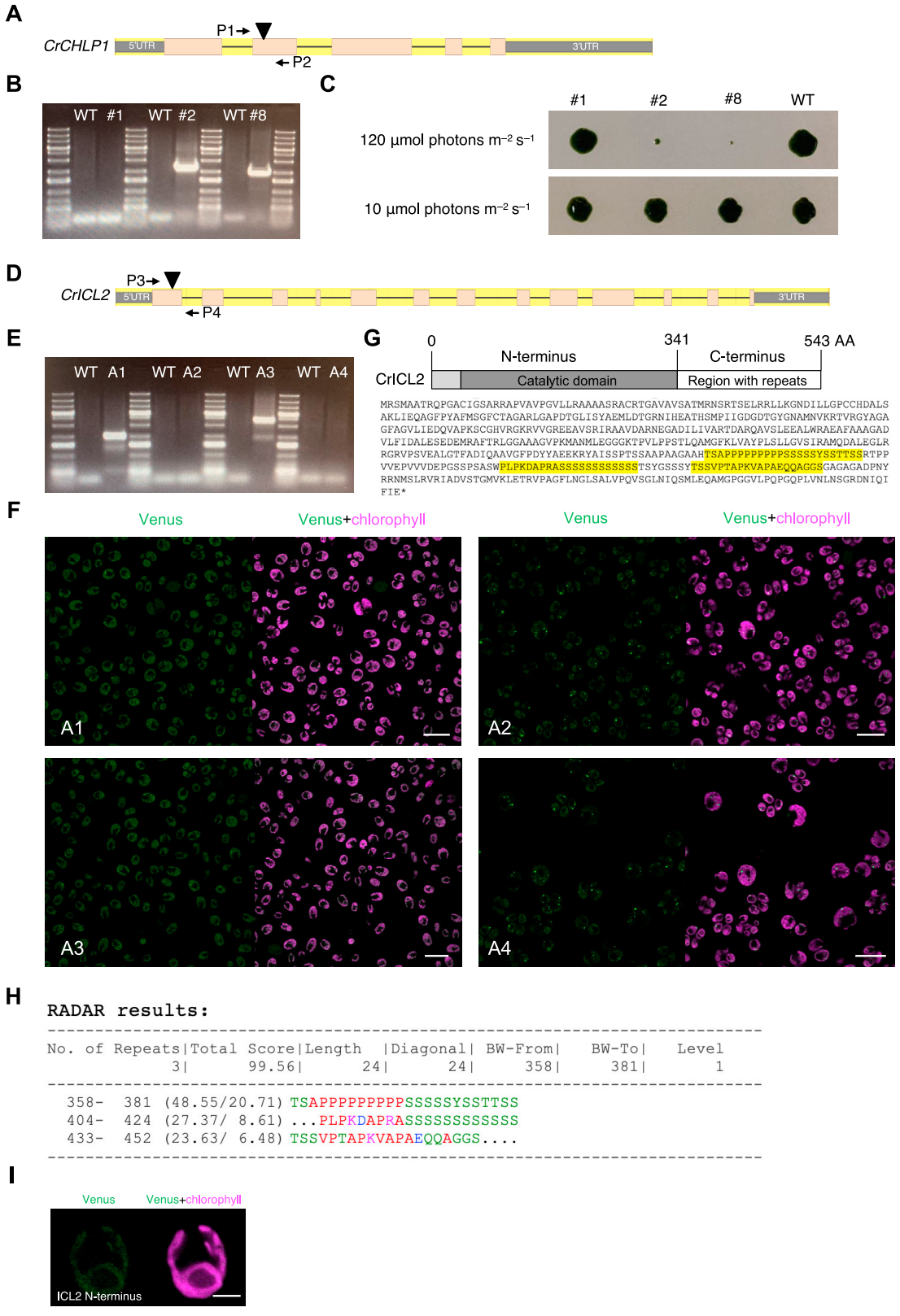


Figure S3. The disruption of *CHLP1* leads to a growth defect under light, and ICL2 is essential for the formation of RBD3 puncta, related to Figure 2

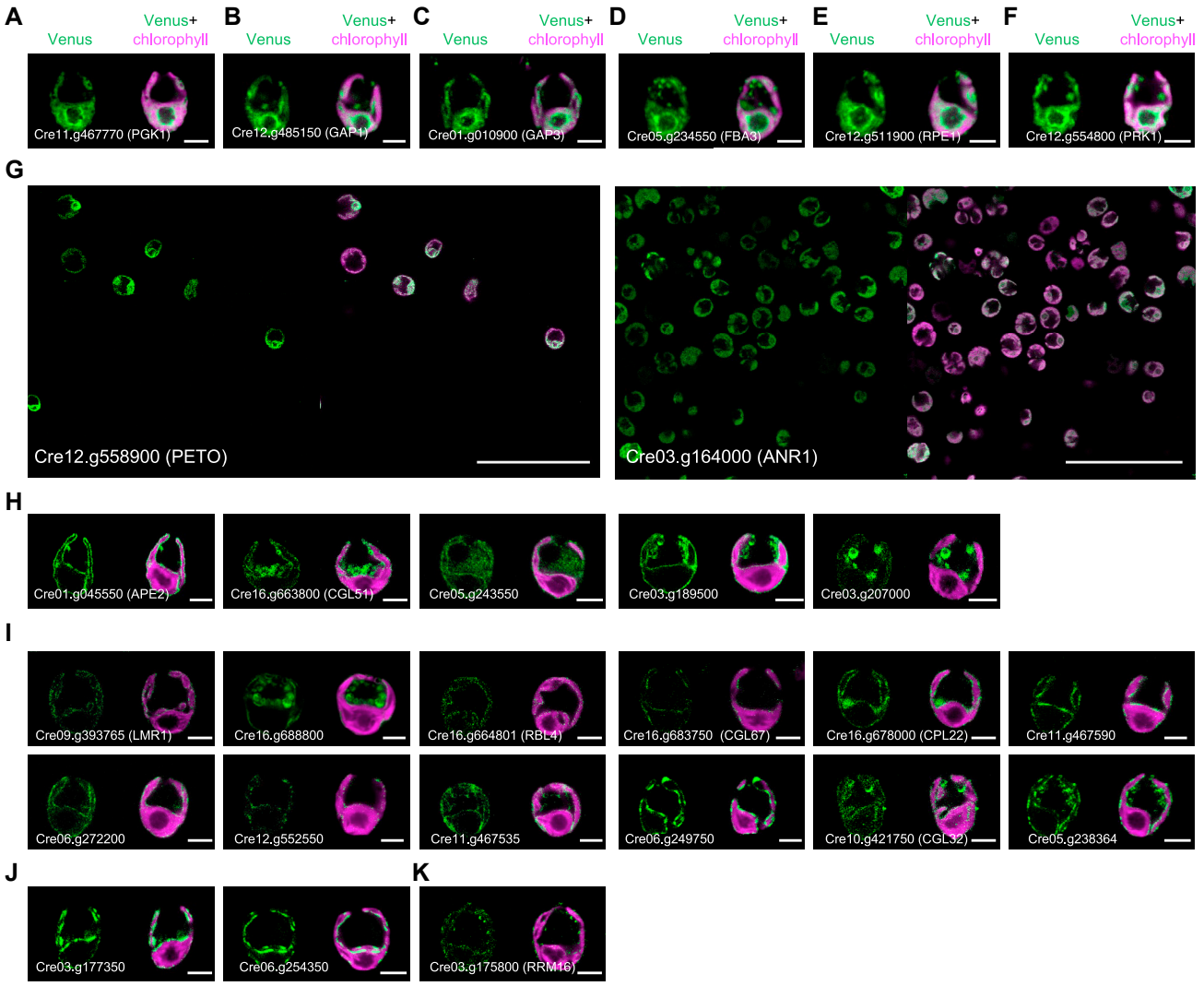


Figure S4. The Calvin-Benson cycle enzymes showed enrichment around pyrenoid, PETO showed different localization patterns from ANR1, and the diverse localizations of proteins localized to the chloroplast envelope, related to Figures 4 and 5

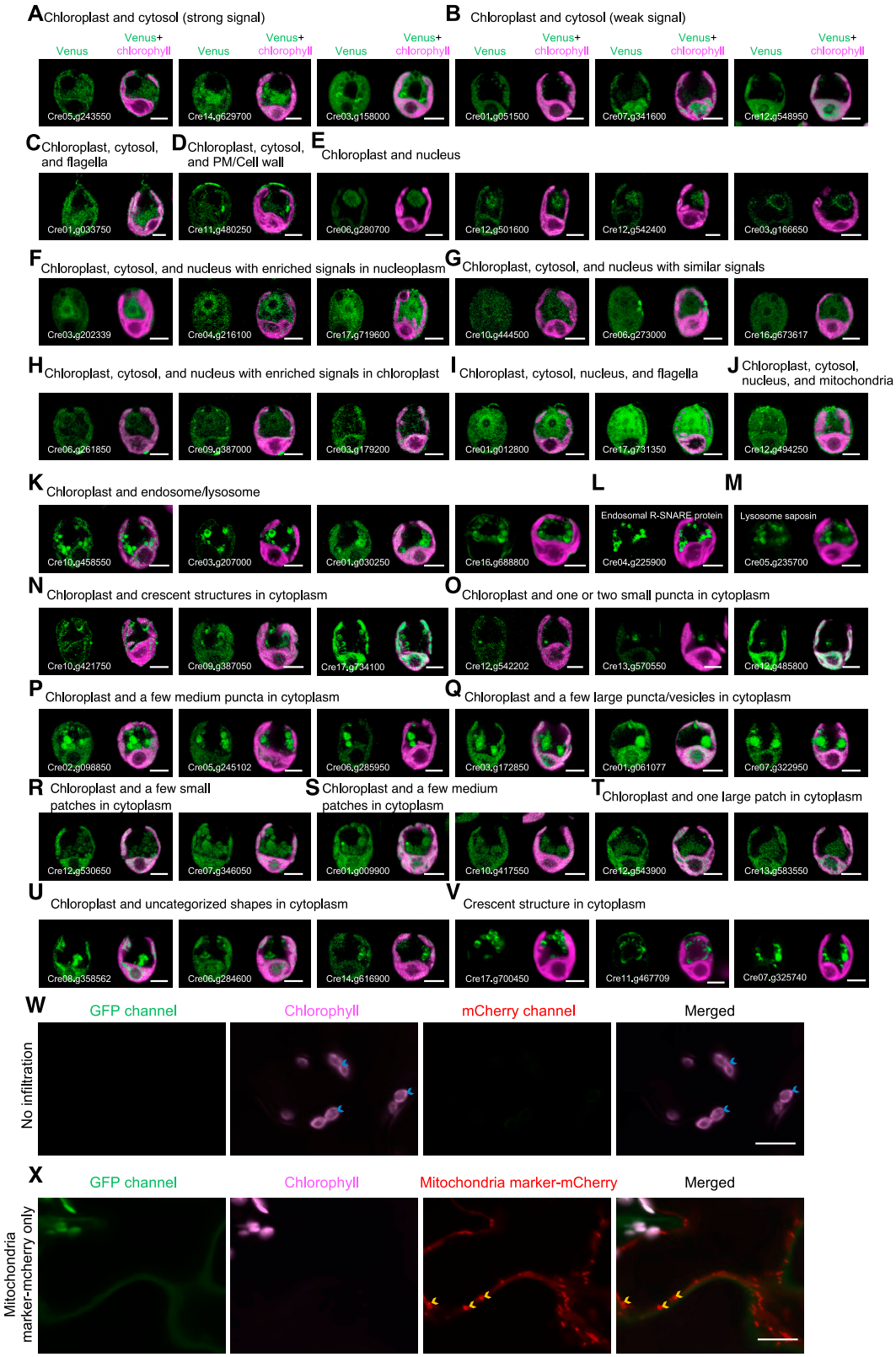


Figure S5. Diverse localizations of representative proteins localized to the chloroplast and other organelles, related to Figure 6

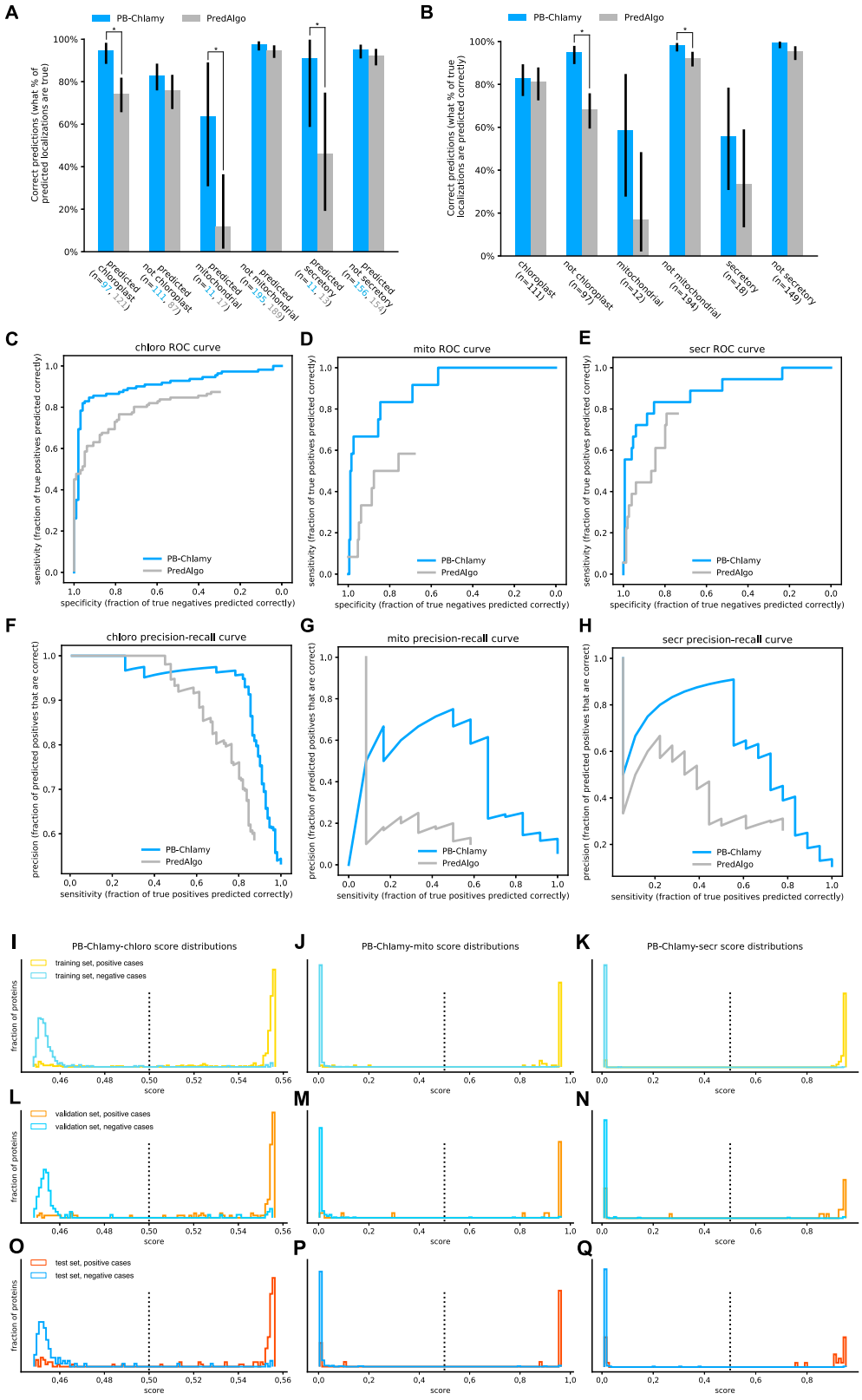


Figure S6. PB-Chlamy reliably predicts *Chlamydomonas* protein localizations, related to Figure 7