Received: 27 April 2023 | Revised: 3 August 2023 | Accepted: 11 August 2023

DOI: 10.1002/pld3.527

RESEARCH ARTICLE

Identification and preliminary characterization of conserved uncharacterized proteins from *Chlamydomonas reinhardtii*, *Arabidopsis thaliana*, and *Setaria viridis*

Eric P. Knoshaug[1] | Peipei Sun[2] | Ambarish Nag[3] | Huong Nguyen[2,4] | Erin M. Mattoon[2,5] | Ningning Zhang[2] | Jian Liu[6] | Chen Chen[6] | Jianlin Cheng[6] | Ru Zhang[2] | Peter St. John[1] | James Umen[2]

[1]Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado, USA

[2]Donald Danforth Plant Science Center, St. Louis, MO, USA

[3]Computational Sciences Center, National Renewable Energy Laboratory, Golden, Colorado, USA

[4]Institute of Genomics for Crop Abiotic Stress Tolerance, Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas, USA

[5]Plant and Microbial Biosciences Program, Division of Biology and Biomedical Sciences, Washington University in Saint Louis, St. Louis, Missouri, USA

[6]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA

Correspondence
James Umen, Donald Danforth Plant Science Center, St. Louis, MO 63132, USA.
Email: jumen@danforthcenter.org

Funding information
Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program, Grant/Award Number: DE-SC0020400

Abstract

The rapid accumulation of sequenced plant genomes in the past decade has outpaced the still difficult problem of genome-wide protein-coding gene annotation. A substantial fraction of protein-coding genes in all plant genomes are poorly annotated or unannotated and remain functionally uncharacterized. We identified unannotated proteins in three model organisms representing distinct branches of the green lineage (Viridiplantae): *Arabidopsis thaliana* (eudicot), *Setaria viridis* (monocot), and *Chlamydomonas reinhardtii* (Chlorophyte alga). Using similarity searching, we identified a subset of unannotated proteins that were conserved between these species and defined them as Deep Green proteins. Bioinformatic, genomic, and structural predictions were performed to begin classifying Deep Green genes and proteins. Compared to whole proteomes for each species, the Deep Green set was enriched for proteins with predicted chloroplast targeting signals predictive of photosynthetic or plastid functions, a result that was consistent with enrichment for daylight phase diurnal expression patterning. Structural predictions using AlphaFold and comparisons to known structures showed that a significant proportion of Deep Green proteins may possess novel folds. Though only available for three organisms, the Deep Green genes and proteins provide a starting resource of high-value targets for further investigation of potentially new protein structures and functions conserved across the green lineage.

KEYWORDS
Arabidopsis, Deep Green conserved proteins, functional annotation, protein structure, Setaria

Eric P. Knoshaug and Peipei Sun co-first authors.

1 | INTRODUCTION

The genome sequencing revolution of the past two decades has removed a major barrier to identifying and describing the genetic toolkits used by green lineage organisms. The number of sequenced plant genomes is growing rapidly, but the resources for comprehensive, experimental structural and functional protein annotation are lagging (Ellens et al., 2017). Homology-based annotations are a simple means of predicting protein function in the absence of any other functional data. In homology-based annotations, new sequences are searched for similarity to proteins in other species, some of which may already have known functions that may be assigned to the newly predicted protein "by proxy." It is generally assumed that conservation at the sequence level implies conservation of function, though this correlation is imperfect (Blaby-Hass & de Crecy-Lagard, 2011). New sequences can also be searched for the presence of conserved domains using sensitive Hidden Markov Models (HMM) and/or multiple sequence alignments (Soding, 2005). The Protein ANalysis THrough Evolutionary Relationships (PANTHER), Protein Families (Pfam), InterPro, and Clusters of Orthologous Groups of proteins (COG) and the eukaryote-specific version EuKaryotic Orthologous Groups (KOG) are powerful classification tools that leverage growing sets of data to improve annotation based on sequence similarity (Blum et al., 2020; Bolger et al., 2017; El-Gebali et al., 2018; Finn et al., 2016; Koonin et al., 2004; Mi et al., 2012, 2016; Tatusov et al., 2003). The above approaches will identify functions and/or structural domains for around half of all proteins in newly sequenced plant genomes, with some variation dependent on genome size and complexity as well as taxonomic position (Hanson et al., 2010). Even the well-annotated genomes from budding yeast and humans still contain approximately 30% unannotated proteins, and 30–40% of these unannotated proteins (∼10% total) are likely to have an uncharacterized catalytic function (Ellens et al., 2017). Likewise, every genome contains predicted proteins that are unannotated because they either have no similarity to characterized proteins or have limited information available beyond the possible presence or absence of domains. Unannotated proteins have typically accounted for approximately 40–60% in plants and algal genomes (Berardini et al., 2015; Blaby-Haas & Merchant, 2019; Niehaus et al., 2015) with functional assignments slowly increasing. The potential for the discovery of new structures, catalytic activities, and biological functions among unknown proteins is enormous, but these proteins also represent a huge challenge due to their overwhelming numbers (Fox et al., 2008; Hanson et al., 2010).

Functional annotation of plant proteins lags behind that of animals, fungi, and prokaryotes though this situation is improving with dedicated portals and tool development on platforms such as Phytozome, PLAZA, and Gramene (Goodstein et al., 2012; Proost et al., 2015; Tello-Ruiz et al., 2018; Van Bel et al., 2022). Currently, approximately 12% of predicted proteins have an experimentally determined structure or structural data derived from related proteins in the Protein Data Bank (PDB) for *Arabidopsis thaliana*, compared to more than 25% in *Saccharomyces cerevisiae* and more than 30% in *Homo sapiens* (Callaway, 2022).

One approach for prioritizing unknown/unannotated genes and proteins for further functional characterization is to demand sequence conservation across one or more taxa (aka phylogenomics). While this does not directly help with annotation, it ensures that information obtained about that gene or associated protein from one species will be impactful as it can likely be applied across species. Phylogenomics approaches have been used successfully to obtain new biological information in multiple contexts. For example, the GreenCut proteins were defined based on conservation in multiple photosynthetic species, but not in non-photosynthetic eukaryotes (Karpowicz et al., 2011; Merchant et al., 2007). Indeed, many of the original unknowns in the GreenCut list were found to have key functions in photosynthetic processes (Arthur et al., 2019; Wakao et al., 2021). The criteria used to define GreenCut proteins provided a strong filter, but it may have also excluded proteins with faster divergence times.

Here we took a similar but less stringent approach to create the Deep Green list of unannotated green lineage proteins. The Deep Green list is based on identification and curation of conserved unknown proteins in three green lineage (Viridiplantae) model organisms; *A. thaliana*, *Chlamydomonas reinhardtii*, and *Setaria viridis*, and is significantly expanded compared with the GreenCut protein list. We report the curation and preliminary characterization of Deep Green proteins and genes using different informatics tools and published data sets. These analyses revealed additional similarities among Deep Green proteins in diurnal expression patterning and in predicted localization. Finally, structural predictions indicated that many of the Deep Green proteins may have novel tertiary folds.

2 | RESULTS AND DISCUSSION

Our objective was to identify and begin characterizing a set of poorly annotated or unannotated conserved green lineage proteins using genetically tractable reference species; *A. thaliana* (Arabidopsis), *S. viridis* (Setaria), and *C. reinhardtii* (Chlamydomonas). Arabidopsis has the best studied and well-annotated genome of any angiosperm species (Berardini et al., 2015), while Setaria (green foxtail millet) is an emerging model C_4 grass and bioenergy feedstock model that has a small stature, short generation time, and is genetically tractable (Hu et al., 2018; Huang et al., 2016; Pant et al., 2016; Zhu et al., 2017). Together, Arabidopsis and Setaria represent two major branches of angiosperms, eudicots, and monocots, respectively. The unicellular green alga Chlamydomonas is a well-established model for investigating cellular processes in photosynthetic eukaryotes due to its fast growth rates, haploid genome, low levels of gene duplication (see below), and availability of high-throughput genetics and genomics tools (Sasso et al., 2018).

2.1 | Identification of conserved protein families

To reduce search complexity for unknown proteins, we developed a down-selection strategy (Figure 1a). The first step involved grouping

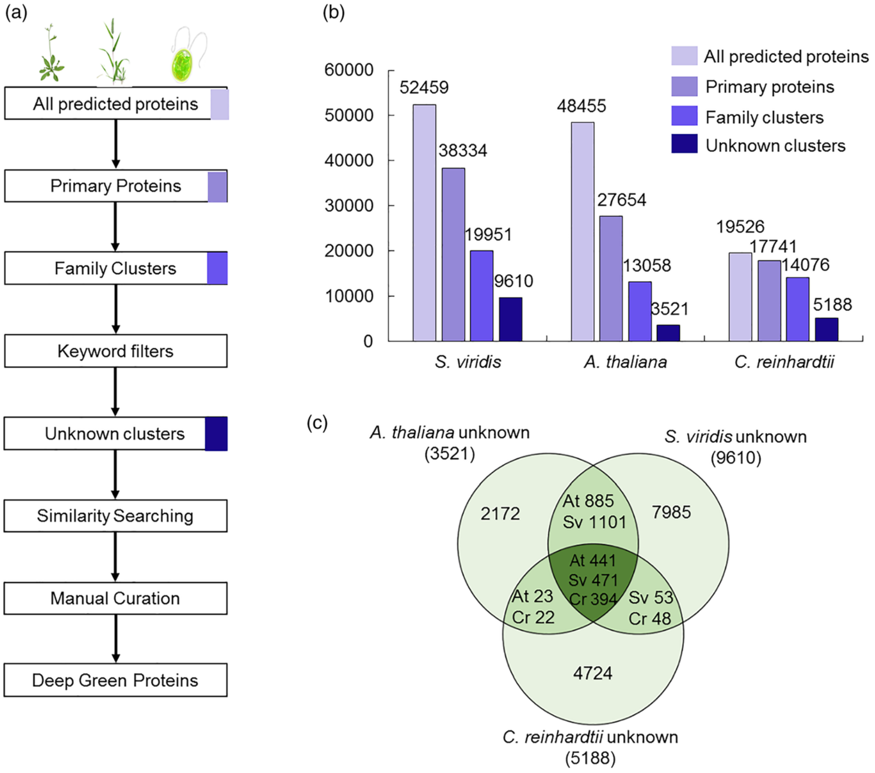

FIGURE 1 (a) Deep Green down-selection flowchart, (b) number of proteins in each species at each of the down-selection steps, (c) Venn diagram showing overlaps between conserved proteins, which define the Deep Green set.

the predicted proteins in each species into families with highly similar sequences (>30% sequence identity, >50% overlap), while at the same time merging annotations among family members (Tables S1–S3). This paralog grouping reduced sequence search space from 27,654, 38,334, and 17,741 predicted primary proteins to 13,058, 19,951, and 14,076 in Arabidopsis, Setaria, and Chlamydomonas, respectively (Figure 1b), and ensured that unannotated members of a paralog family were not included when one of their family members was already annotated. The h3-cd-hit algorithm, which was used for grouping into families with one or more members, also selected a lead protein in each multi-protein cluster, which served as a representative for similarity searching between proteomes of the three focal species (Tables S4–S6).

The next step was identifying protein families that were shared between the species. Using the basic local alignment search tool for protein (BLASTP) among lead proteins from each species, the top high-scoring segment pair (hsp) was identified, and those that passed a similarity threshold (e-value cutoff threshold of 10^{-3}; >40% of the sequence length aligned [qcovs] for at least one of the two proteins) were used to further group proteins into superfamilies shared across two or more of the focal species (Table S7). Two-way comparisons of protein family sizes also revealed the relative compactness of the Chlamydomonas protein family sizes, which are typically single copies compared with land plant families that often have multiple paralogs (Figure 2). We compared our inter-specific homolog search process to results obtained from INPARANOID (Remm et al., 2001) and OrthoFinder (Emms & Kelly, 2019) (Figure S1). Our method showed comparable results, though there were instances where either method did not group the same proteins with the same orthologous groups or did not identify an ortholog (Tables S8–S10). Orthofinder is one of the most sensitive methods for identifying ortho groups and included 85 conserved unknown proteins missing from our Deep Green list. An additional eight conserved unknown proteins were found by INPARANOID but not Orthofinder. Of these 93 proteins not included in our list, 23 were false positives as they are in families with members that have known functions and 62 were rejected due to short coverage length (qcovs) for the region of similarity (<40% for one of the 2 genes being compared). The remaining eight were not missed by our method but rather had different best homologs with higher match scores (Table S10). Conversely, our custom method identified many matches that were missed by OrthoFinder or INPARANOID (Figure S1). Overall the three methods were largely in agreement and the differences among them represented less than 10% of the total families.

2.2 | Identification of unannotated/unknown proteins

For each protein family, Phytozome annotations were collected, merged, and used to identify unannotated or poorly annotated families (Figure 1a). We defined poorly or unannotated proteins in these species as those without any annotation in the Phytozome database (Goodstein et al., 2012), and those with only limited annotation (Tables S4–S6). If a family in one species was well annotated or characterized, its conserved counterparts in the other two species were also considered annotated and removed from consideration. There was some subjectivity regarding proteins with existing but limited annotation, andwe focused our efforts on those where the domain

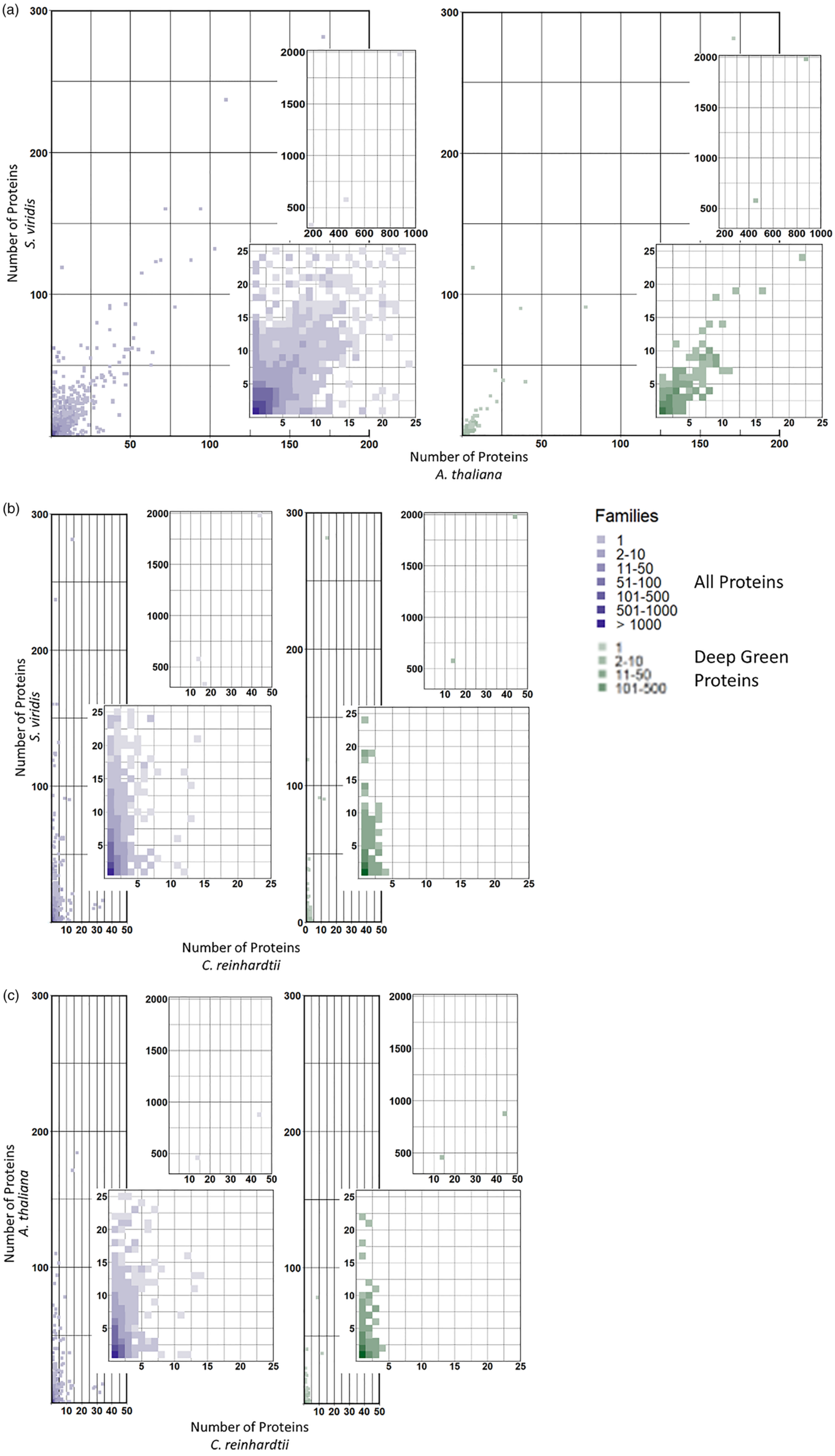

FIGURE 2 Comparison of protein family sizes in (a) Arabidopsis and Setaria, (b) Chlamydomonas and Setaria, and (c) Chlamydomonas and Arabidopsis. Insets show in greater detail families containing 25 or fewer proteins or those containing greater than 300 members. Axes indicate the number of proteins in a family and color scale denotes the number of families at each position in the graph. For example, the darkest point in each graph (1:1) represents >1,000 protein families with a single family member in each of the two species.

was small and the region of similarity extended beyond that small domain. If the definition line (defline) was blank or contained the terms "unknown," "undefined," "uncharacterized," "hypothetical," "domain of unknown function," "expressed protein," "transmembrane," "function unknown," "predicted protein" and "conserved in plant or green lineage," "anykrin," "fbox," "tetra- or penta- tricopeptide," the family was retained. We also incorporated proteins from previously compiled lists of unannotated Arabidopsis proteins (Cheng et al., 2017) (https://conf.phoenixbioinformatics.org/pages/viewpage.action?pageId=22807120).

The conserved unknown proteins within each list were further curated by manual inspection of annotations and by searching for the protein or gene IDs in publications. If a protein had been functionally characterized (e.g., published mutant phenotype) it was removed from the list. Manual searching for ambiguous or poor-quality annotations among the lists of annotated clusters (e.g., Arabidopsis clusters 7,885 and 9,084, which were only annotated as "plant/protein") was also done to enable the inclusion of proteins that did not fit the defined criteria above. In parallel, we updated the GreenCut2 protein list with new gene IDs based on the Chlamydomonas v5.5 genome assembly and by removing those in the list that had been characterized since publication (Arthur et al., 2019; Karpowicz et al., 2011). The remaining 204 uncharacterized GreenCut2 proteins were merged into the final list of unannotated proteins. This final manual curation led to 3,521, 9,610, and 5,188 uncharacterized or poorly characterized families in Arabidopsis, Setaria, and Chlamydomonas, respectively (Figure 1b). Finally, we sorted the unknown protein lists to identify overlaps between each of the species (Figure 1c, Tables S4–S6). Criteria for inclusion in the Deep Green list was the presence of a homolog in Chlamydomonas, as well as in at least one of the two land plant species. As expected, most of the list members had homologs in both plant species since the latter two diverged from each other much

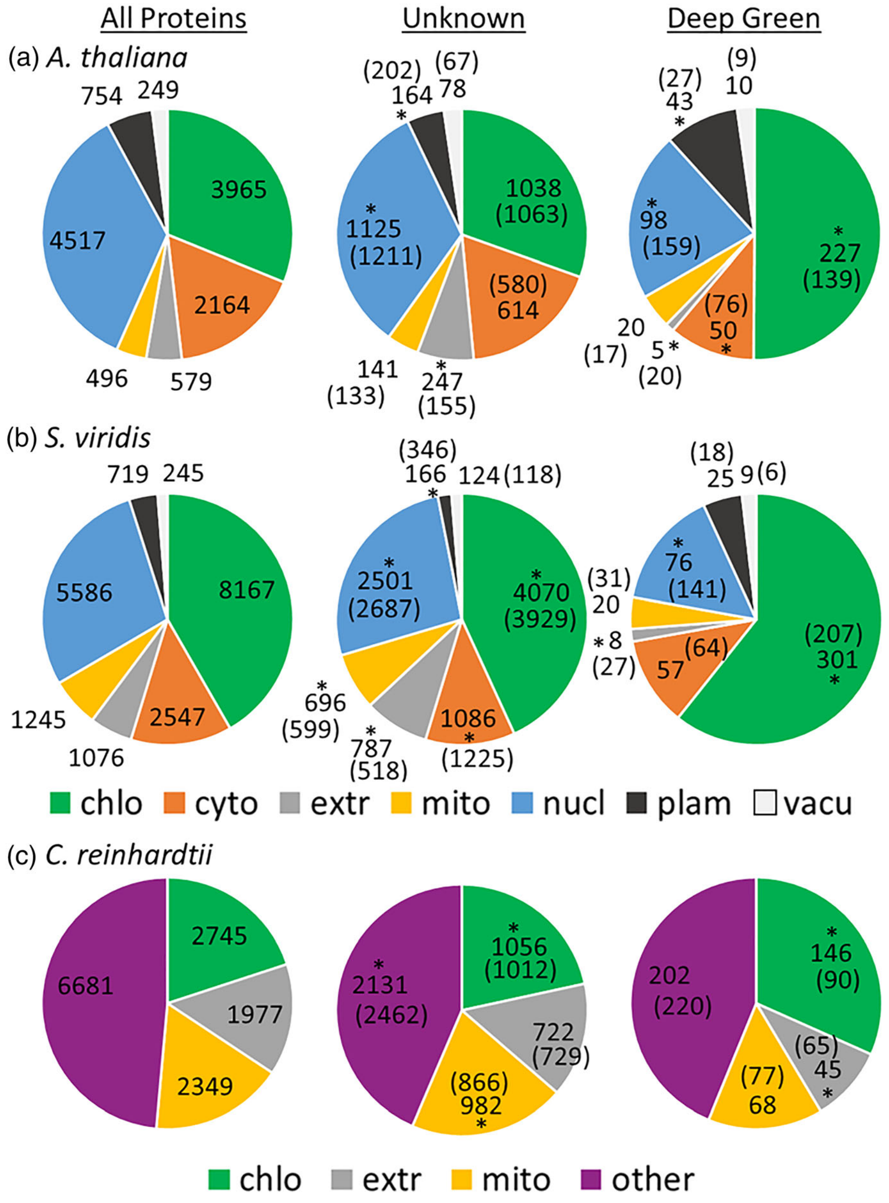

FIGURE 3 Predicted protein localization for (a) Arabidopsis, (b) Setaria, and (c) Chlamydomonas. Predictions for Arabidopsis and Setaria were done using WoLFPSORT. Predictions for Chlamydomonas were done using Predalgo, which was trained on Chlamydomonas. Numbers in parentheses for the unknown and Deep Green protein sets are the expected number using Fisher's Exact Test (background size 13,058, 19,952, and 14,076 for the 3 species, respectively) based on the total number of proteins and their predicted localization. Significant enrichment or depletion (*FDR corrected p-value < .05) is indicated. Cellular localization abbreviations: chlo, chloroplast; cyto, cytosol; extr, extracellular; mito, mitochondria; nucl, nucleus; plam, plasma membrane; vacu, vacuole.

more recently than their common ancestor with Chlamydomonas, and any missing proteins were likely due to gene loss in one of the two land plants. The larger group of unknown proteins shared just between Arabidopsis and Setaria is also of interest but was not further characterized here.

2.3 | Characterization of unknown and Deep Green proteins

We interrogated the Deep Green list in several ways to provide preliminary information on potential functions. Predicted subcellular targeting for each Deep Green protein was compared with that of the entire predicted proteome for each species using Wolf P-Sort (Horton et al., 2007) for plant proteins and Predalgo (Tardif et al., 2012) for Chlamydomonas. Deep Green proteins in all three species were predicted to be strongly enriched for chloroplast targeting, and the plant members of this set were also depleted for nuclear targeting (Figure 3). Importantly, the strong chloroplast localization enrichment for predicted proteins in each species was not seen in the set of all unknown proteins where there was just a slight enrichment. These findings suggest that there are a significant number of conserved proteins with chloroplast functions that have yet to be characterized. Deep Green and unknown proteins were also significantly enriched in proteins predicted to contain one or more transmembrane domain(s) as compared to all predicted protein families (Figure S2).

We next performed co-expression analysis for Chlamydomonas Deep Green genes using published transcriptome data sets. An important resource was a previously described high-resolution diurnal data set for synchronized Chlamydomonas cultures (Zones et al., 2015). In that study, around 80% of genes with detectable expression (∼12,000) showed strong periodic diurnal or cell-cycle-controlled expression patterns (Figure 4a). Genes coding for the Deep Green proteins showed higher overall average expression levels compared with all expressed genes of 3.63 versus 1.91 $\log_2$RPKM, respectively (Figure 4b). We also investigated the distribution and enrichment of Deep Green genes in 18 diurnal clustered and unclustered expression groups and found them to be significantly over-represented in clusters 2, 4, 6, and 7, which all have peak expression in the light phase when most of the chloroplast or photosynthesis-related genes are also expressed (Figure 4c, Figure S3). Deep Green genes were also significantly over-represented in the non-differentially expressed cluster (i.e., constitutively expressed genes) and are under-represented in the non-expressed group (FPKM<1 at all time points). Finally, Deep Green genes were also significantly under-represented in the dark phase clusters 13 and 15, which are enriched for cell motility and protein post-translation modification, respectively. These results, combined with the enrichment for Deep Green proteins targeted to the chloroplast (Figure 3), suggest that approximately 60% of Deep Green genes may have important fundamental roles in chloroplast function or biogenesis.

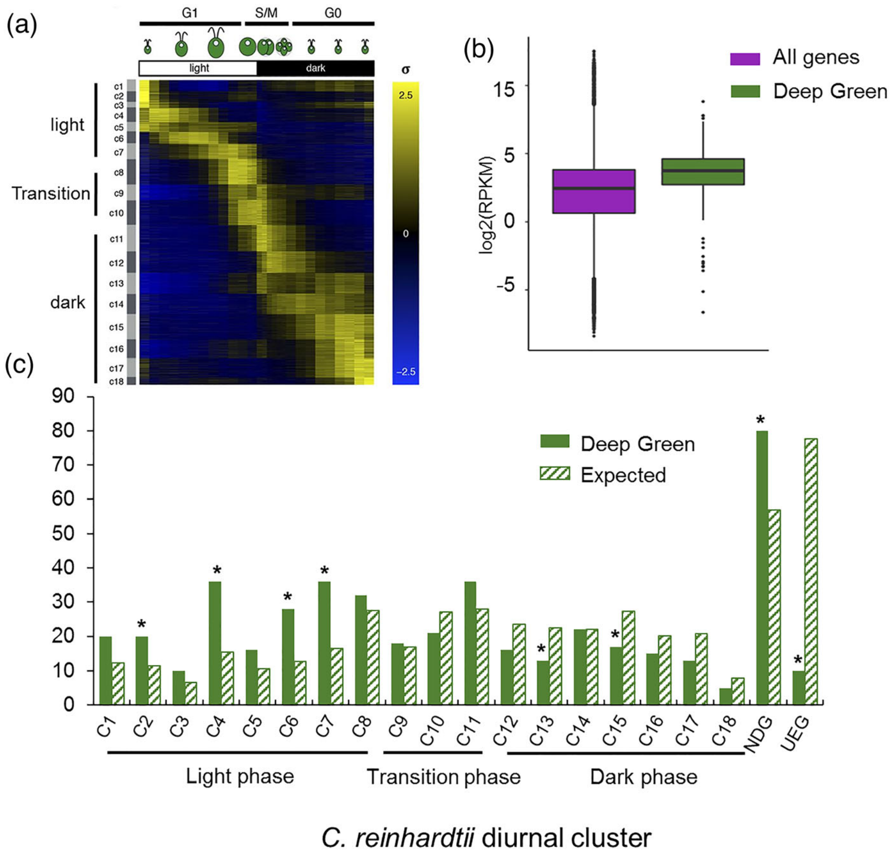

FIGURE 4 Relative expression levels and diurnal expression patterning of Deep Green genes. (a) Expression heatmap of differentially expressed genes in diurnally synchronized cultures as described previously (Zones et al., 2015, reproduced with permission from the authors). (b) Average transcript abundance of Deep Green genes in the diurnal transcriptome compared to all genes (p-value = 3.0 e-37). (c) Enrichment of Deep Green genes in 18 diurnal expression clusters shown in panel A with peak expression times shown in the same order. Non-differentially expressed (NDG), and unexpressed (UEG) groups in the diurnal transcriptome are on the right side. Significant enrichment or depletion determined by Fisher's Exact Test (background size 17,737) is indicated (*FDR corrected p-value < .05).

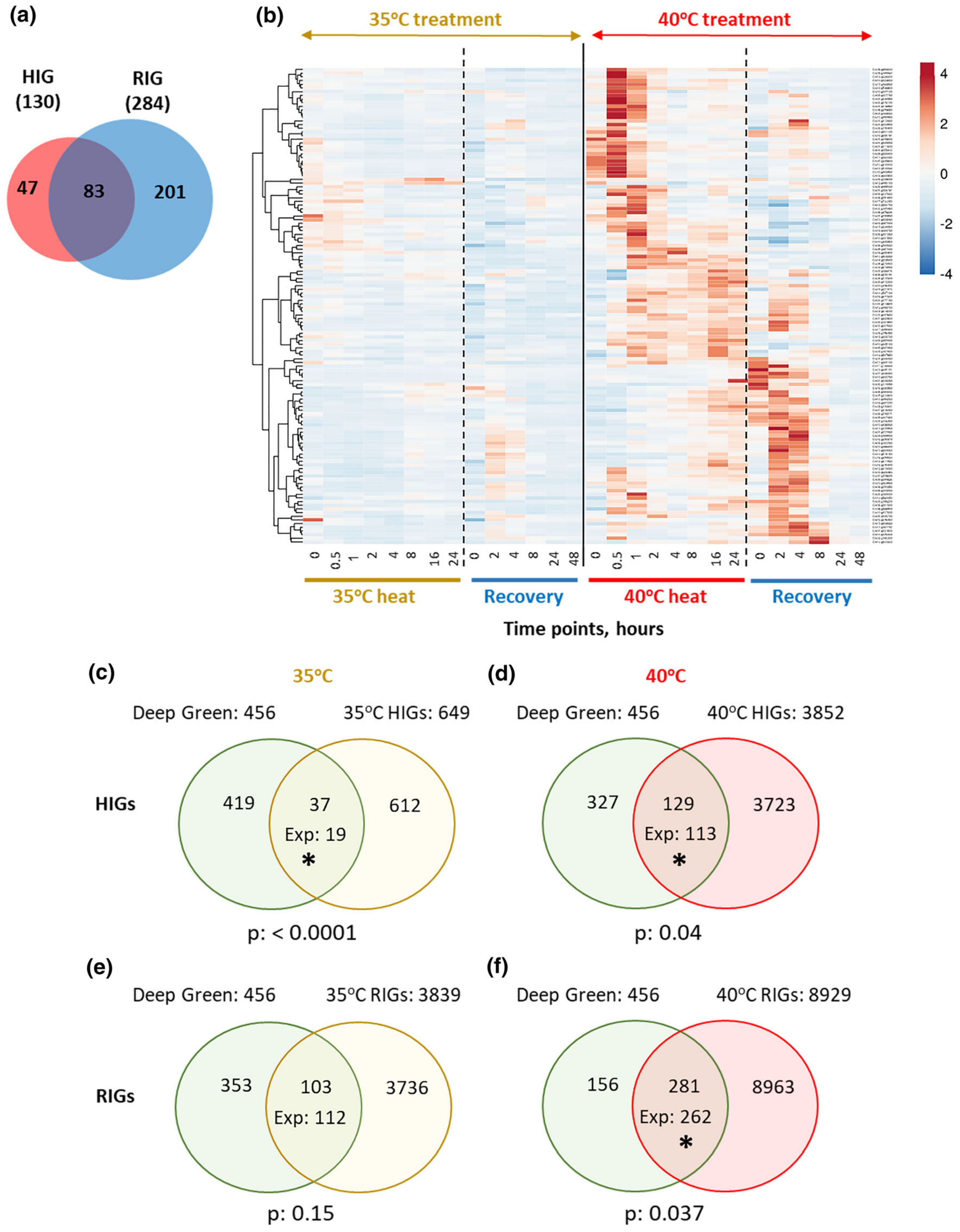

FIGURE 5 Deep Green genes are significantly enriched for heat-inducible genes in Chlamydomonas. (a) Differentially expressed Deep Green genes during and after heat treatments (35 °C or 40 °C) described in Zhang et al. (2022). Deep Green genes with a $\log_2$(foldchange) ≥ 1 and an FDR corrected p-value < .05 at a minimum of one-time point during heat treatment of 35 °C or 40 °C or during recovery were identified as heat-inducible genes (HIGs) or recovery inducible genes (RIGs), respectively. (b) Heatmaps of differentially expressed Deep Green gens during and after heat treatments. Color bars represent log2(foldchange) of transcripts as compared to the preheat time point, with red colors for up-regulation and blue colors for down-regulation, white color for no differential expression. The black solid line separates the 35 °C and 40 °C treatments. The black dashed lines indicate the end of 24 h heat treatments. Time points indicate the length of time at the respective temperature starting from 0 hours (h) when the sample had reached the target (35 °C or 40 °C) or recovery (25 °C) temperature. Each horizontal row represents a Deep Green gene. (c–f) Deep Green genes were significantly enriched for HIGs during (c, d) and RIGs after (e, f) heat treatments (Fisher's Exact Test, background size 15,541, *p-value < .05). Exp, expected overlapping numbers based on random chances. (e, f) *C. reinhardtii* Deep Green genes were significantly enriched for RIGs after 40 °C heat treatment, (Fisher's Exact Test, background size 15,541, *p-value < .05). Exp, expected overlapping numbers based on random chances.

We also examined the expression of Chlamydomonas Deep Green genes in a set of data that identified genes upregulated during and recovery from heat stress (Zhang et al., 2022). Deep Green genes were identified in transcriptome data of wild-type Chlamydomonas cells in response to 24 h high-temperature treatments of 35 °C or 40 °C followed by recovery at 25 °C (Figure 5). Of the Deep Green genes present in the RNA-seq dataset, 130 (29%) and 284 (62%) were significantly up-regulated during heat treatments (heat-induced genes, HIGs) and recovery phase (recovery-induced genes, RIGs), respectively (Figure 5a). Among them, 83 (18%) were significantly up-regulated during both heat treatment and recovery while only 47 (10%) were up-regulated during heat treatments but not during the recovery phase. For the Deep Green genes that were represented in the HIGs and RIGs, expression was much stronger in the 40 °C treatment than in the 35 °C treatment suggesting a connection between heat stress response and Deep Green gene function (Figure 5b). In the 40 °C heat stress experiment, more than 50% of the heat-inducible Deep Green genes changed their expression immediately in the first 2 or 4 h. Finally, we looked at enrichment of Deep Green genes among HIGs and RIGs from the 35 °C and 40 °C experiments and found significant over-representation for HIGs from both the 35 °C and 40 °C high-temperature treatments (Figure 5c,d), which suggests conservation of temperature responsive genes in the green lineage. In contrast, Deep Green genes were enriched for RIGs after 40 °C but not 35 °C heat treatments (Figure 5e,f), suggesting some Deep Green genes may have potential functions in recovery from acute heat stress.

As a more general test of co-expression, we examined the distributions of Deep Green genes in ChlamyNET, a web-based tool to explore gene co-expression networks based on published transcriptome data (Romero-Campero et al., 2016). ChlamyNET has 9 major co-expression clusters among 9,171 Chlamydomonas genes and captures co-expression relationships established under 25 different growth conditions. Among all 464 Chlamydomonas Deep Green genes, 240 are also represented in ChlamyNet. Deep Green genes were significantly enriched in clusters containing proteins associated with protein assembly and degradation and translation and lipid metabolism in the co-expression network (Figure 6). Both of these ChlamyNet clusters with over-representation of Deep Green genes were themselves over-represented for light phase or light–dark transition phase genes as previously defined (Matt & Umen, 2018; Zones et al., 2015) (Figure S4). Taken together, our results suggest conserved but unexplored functions for many Chlamydomonas Deep Green proteins in photosynthetic biology and stress responses.

We also performed enrichment testing of Deep Green genes from Arabidopsis by making use of published transcriptome data from a study on responses to combinations of stresses (Rasmussen et al., 2013). Under the combined stress conditions, which included salt, cold, heat, high light, and flg22 peptide treatment (a stimulator of pathogen responses), nine significant co-expression modules were defined. Arabidopsis Deep Green genes were enriched in module 2 (p-value = .027), which shows association with both single abiotic stresses and combined abiotic stresses. To examine the diurnal expression pattern of Arabidopsis Deep Green genes, we interrogated a microarray data set (Blasing et al., 2005) since we could not find a publicly available Arabidopsis diurnal data set from an RNA-seq experiment. The microarray data had a much smaller dynamic detection range than a typical transcriptome study, and thus there were fewer DEGs in Arabidopsis than in the Chlamydomonas diurnal data. We grouped all the DEGs (2342) into 12 clusters based on gene expression patterns. The low number of DEGs in each Arabidopsis cluster limited the statistical power of finding enrichment among the Deep Green subset of genes, but we noted that cluster 2, which peaked during the light phase, had the strongest enrichment of Deep Green genes (p-value = .068) (Figure S5).

2.4 | Structural properties of Deep Green proteins

Tertiary structures for each Deep Green protein were predicted using AlphaFold v2.1 (Jumper et al., 2021) and are available as the Deep Green protein set (https://data.nrel.gov/submissions/216; DOI: 10.7799/1970473) (Figure 7). Predicted structures for all proteins in the larger unannotated/unknown protein set for each organism were also deposited. Deep Green proteins and proteins in the unknown sets with predicted local-distance difference test (pLDDT) confidence score higher than 50 (1,338 from among the three species in the Deep

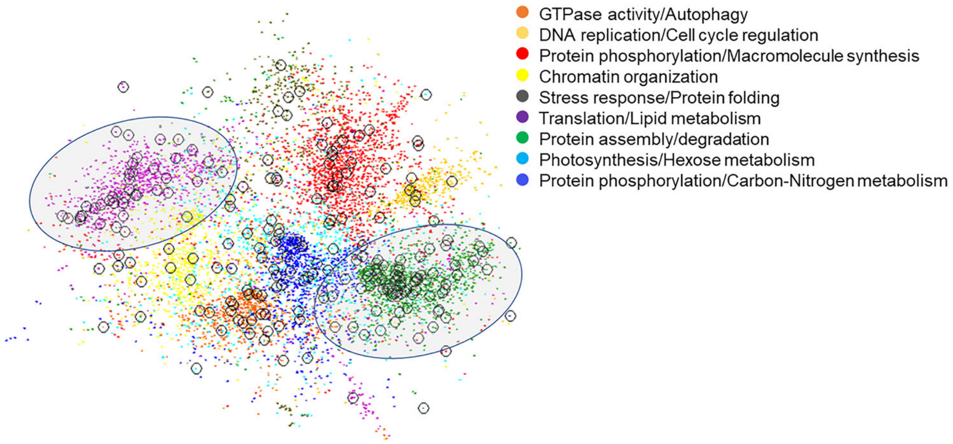

FIGURE 6 Distribution of Deep Green genes in the ChlamyNet cluster network. Two-dimensional graph showing gene expression clusters that are color-coded, with each point representing a gene in ChlamyNet whose subcluster enrichment profile is shown in the color key. Deep Green genes are circled, and the two large subclusters with enriched representation of Deep Green genes are demarcated by shaded ovals.

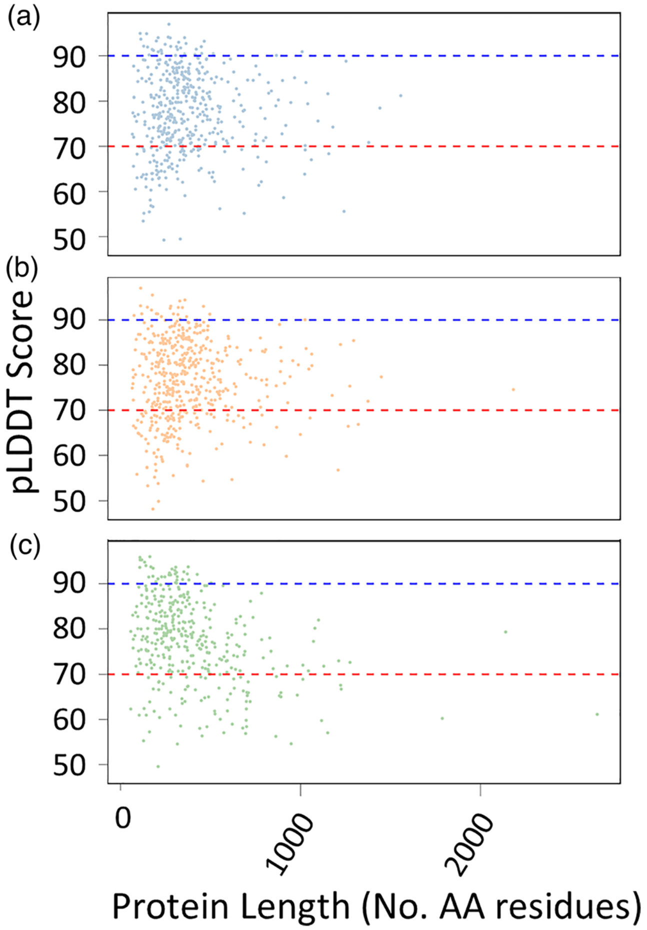

FIGURE 7 Average local-distance difference test (pLDDT) scores for Deep Green protein structures predicted using AlphaFold v2.1 for (a) Arabidopsis, (b) Setaria, and (c) Chlamydomonas. Scores >90 (blue dashed line) are considered to be highly accurate. Scores between 90 and 70 (red dashed line) are considered to indicate a generally correct backbone structure. Scores between 70 and 50 are considered to be low confidence.

Green protein set and 2,568, 6,764, 3,221 in the unknown sets for Arabidopsis, Setaria, and Chlamydomonas, respectively) were selected to perform structural matching in the Protein Data Bank (PDB) using Foldseek (van Kempen et al., 2022) (Tables S11, S12). Foldseek creates a template modeling (TM) score between 0 and 1 that reflects structural differences between an input query protein and its best structural match in PDB. TM scores below .5 indicate significantly different structures while those close to 1 are near perfect matches. The TM score distribution between the predicted tertiary structures and extant protein models in the PDB showed 777 out of 1,338 (58.1%) Deep Green proteins from all three species (268 out of 455 for Arabidopsis, 220 out of 381 for Setaria, and 289 out of 502 for Chlamydomonas) to have TM scores less than .5, a common threshold for protein structural comparison, and therefore are likely to have structures with novel folds (Xu & Zhang, 2010) (Figure 8a). In addition, 9 selected Deep Green proteins having potentially novel folds showed high structural similarity among the three interspecific homologs (Table 1, Figure 8b). Among the larger functionally unannotated protein set in each organism, 1,304, 3,118, and 2,425 proteins in Arabidopsis, Setaria, and Chlamydomonas, respectively had a TM-score < .5 suggesting novel folds. Finally, approximately 60% of the Arabidopsis proteome currently has either experimentally determined structures or structures through association with related proteins in the PDB (12%), with the remaining majority (48%) having been predicted using AlphaFold (https://alphafold.ebi.ac.uk/)(Callaway, 2022).

To better understand the structural complexity of the Deep Green protein set, two measurements describing order versus disorder were used. IUPred3 predicts the likelihood of individual residues being in a structured region based on thermodynamic properties of each residue (Erdos et al., 2021) (Figure 9a). An alternative predictor for disorder is the percentage of alanine (A), glycine (G), and proline (P) (%AGP) in each region of a protein (Cock et al., 2009) (Figure 9b). The distribution of the percentage of disorder values for the Deep Green proteins in each of the three organisms indicates that the Deep Green proteins have higher representation at lower percentage disorder values and lower representation at higher percentage disorder values compared to all proteins suggesting they are overall more ordered than average (Figure 9a). The %AGP for the Deep Green proteins also reflected less disorder than that for the set of all predicted (Figure 9b). We also note, as previously observed (Basile et al., 2017), a correlation between GC content in each of the three genomes and the overall amount of proteome structural disorder predicted from % AGP since the three AGP codons are represented by GC-rich triplets. Chlamydomonas has the greatest amount of disorder in its predicted proteome and the highest genomic GC content (66%), while Arabidopsis has the lowest predicted disorder and the lowest GC content (36%) with Setaria in between (46% GC). Nonetheless, within each species, the Deep Green proteins were predicted to be more structured than average proteins.

3 | SUMMARY AND PERSPECTIVES

The goal of this study was to identify conserved unknown proteins in genetically tractable green lineage representatives to enable follow on studies aimed at assigning function to the large fraction of proteins currently uncharacterized. As a first step in this direction, stable structures for the Deep Green proteins were predicted, and, excitingly, many did not have significant matches in the PDB, meaning that they likely represent new families of structural folds, which were partly validated by the observed agreement found between predicted structures of Deep Green orthologs in the three species from which they were selected. Using only three species may miss some proteins conserved across the greater green lineage, however, the conservation of Deep Green proteins implies that they may encode important agronomic traits. For example, 4 Setaria loci (Sevir.1G224300, Sevir.5G282600, Sevir.5G335650, Sevir.9G583700) identified as being linked to temperature and precipitation extremes (Mamidi et al., 2020) and 107 Setaria genes whose expression changed in response to aphid infection (Dangol et al., 2022) were part of our Setaria Deep Green protein set and are excellent candidates for

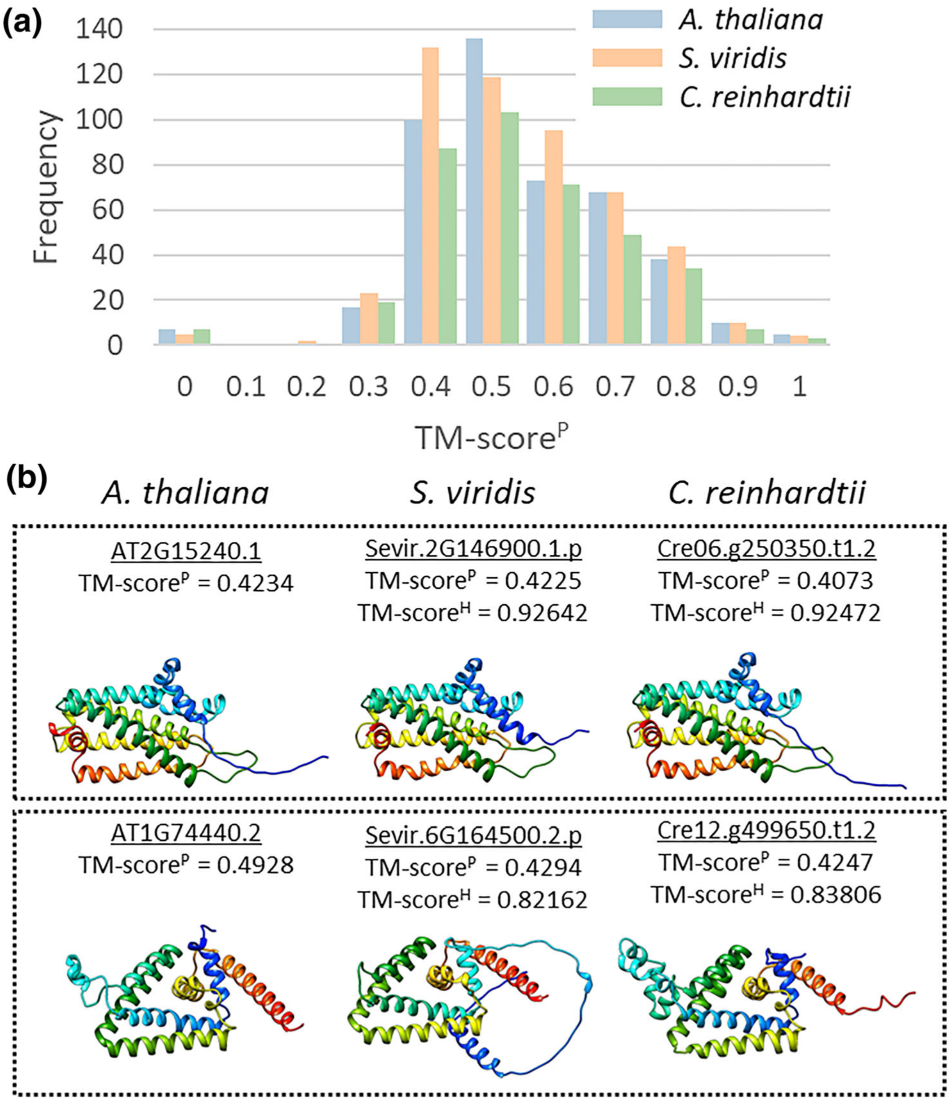

FIGURE 8 Identifying Deep Green proteins with novel structural folds. (a) TM-score^P distributions are shown for the best match in the protein data Bank (PDB) for each predicted Deep Green protein structure, with data for each species identified in the legend. A TM-score^P below .5 is considered a new fold. (b) Top and bottom boxes show structural predictions for two separate Deep Green protein families with novel structures. Ribbon diagrams show secondary structures. Best TM-scores for PDB matches are shown by TM-score^P and matches between Deep Green homologs are shown by TM-score^H compared to the Arabidopsis structure.

TABLE 1 TM-score^H for interspecific Deep Green protein structures having a potentially novel fold. Arabidopsis versus Setaria (Ara-Set), Arabidopsis versus Chlamydomonas (Ara-Chl), and Setaria versus Chlamydomonas (Set-Chl).

			TM-score^H		
Arabidopsis	**Setaria**	**Chlamydomonas**	**Ara-Set**	**Ara-Chl**	**Set-Chl**
AT2G41770.1	Sevir.3G262100.1.p	Cre06.g260650.t1.1	.708	.446	.487
AT1G28140.1	Sevir.9G010200.1.p	Cre04.g228450.t1.2	.847	.781	.794
AT1G74440.2	Sevir.6G164500.2.p	Cre12.g499650.t1.2	.822	.838	.680
AT3G28720.2	Sevir.4G242300.1.p	Cre02.g081650.t1.2	.826	.774	.649
AT5G11960.1	Sevir.5G293000.1.p	Cre10.g419700.t1.1	.913	.732	.650
AT1G21370.1	Sevir.7G112400.1.p	Cre02.g116000.t1.1	.795	.533	.452
AT1G04190.1	Sevir.3G398100.1.p	Cre02.g106850.t1.2	.797	.532	.951
AT2G15240.1	Sevir.2G146900.1.p	Cre06.g250350.t1.2	.926	.925	.659
AT3G60810.1	Sevir.9G003500.1.p	Cre11.g468750.t1.2	.722	.734	.681

further functional characterization. The inclusion of the alga Chlamydomonas provided a strong filter for identifying deeply conserved proteins and against proteins that arose more recently in angiosperms or that evolved rapidly, and thus associated many unknown proteins with multiple paralogs from plants that had a single copy or low copy number paralogs in Chlamydomonas (Figure 2). Our search not only captured previously identified GreenCut2 proteins but expanded the list of unknowns to 464 genes (2.3-fold of the unknowns in GreenCut2).

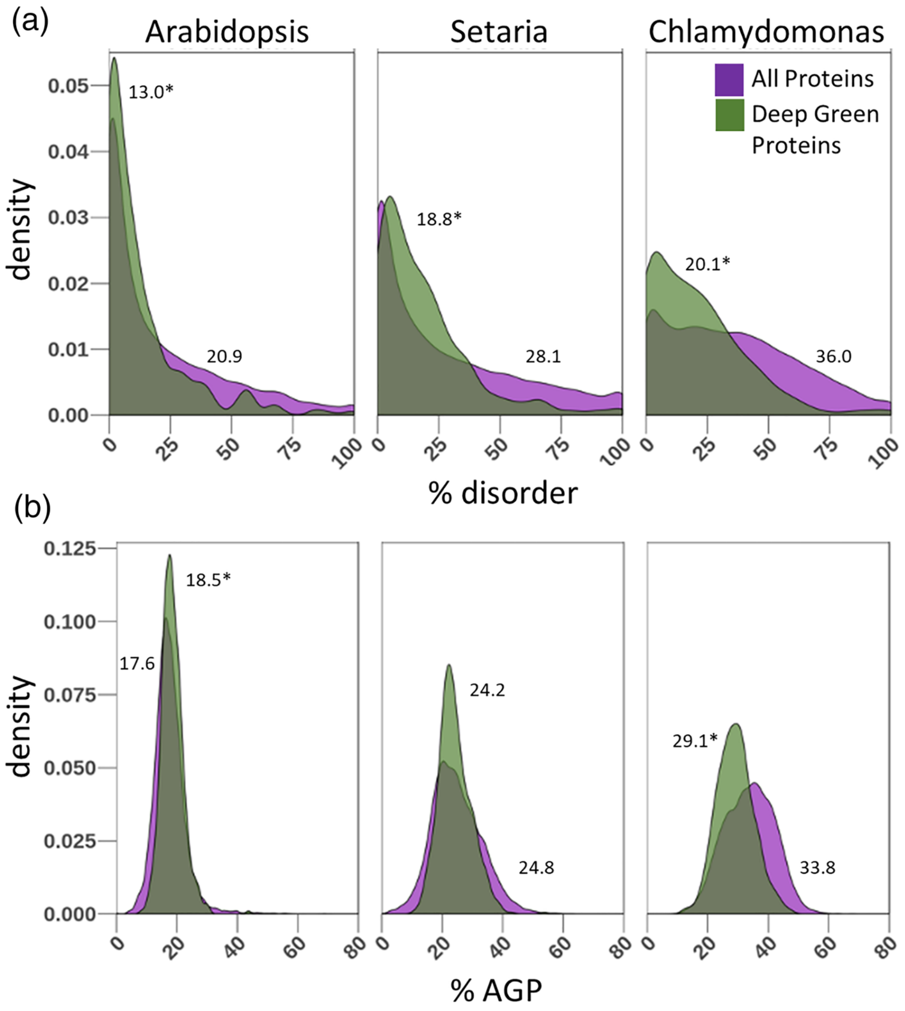

FIGURE 9 Distributions of (a) predicted % disorder or (b) % residues correlated with disorder, alanine, glycine, and proline (AGP). Each panel shows distribution of all predicted primary proteins in the indicated species and the subset of Deep Green proteins including the mean % value positioned near the respective dataset. *p-value < .05.

The Chlamydomonas Deep Green genes are good future targets for functional genomics studies since there are reverse genetics resources available including an indexed mutant library (Li et al., 2016) and gene editing tools (Ferenczi et al., 2017) to enable rapid functional characterization. In summary, the Deep Green gene/protein list that has been created and characterized here will be an impactful starting point for applying functional genomics and structural studies that will help shed light on unexplored areas of biology in photosynthetic eukaryotes.

4 | MATERIALS AND METHODS

4.1 | Datasets

Current protein lists for Arabidopsis, Setaria, and Chlamydomonas were downloaded from Phytozome 13 (https://phytozome.jgi.doe.gov/pz/portal.html). The files downloaded and used in our analysis were:

Arabidopsis v447_Araport11: Athaliana_447_Araport11.annotation_info.txt Athaliana_447_Araport11.define.txt.
Athaliana_447_Araport11.protein_primaryTranscriptOnly.fa.
Setaria v2.1: Sviridis_500_v2.1.annotation_info.txt.
Sviridis_500_v2.1.defline.txt.
Sviridis_500_v2.1.protein_primaryTranscriptOnly.fa.
Chlamydomonas v5.6: Creinhardtii_281_v5.6.annotation_info.txt.
Creinhardtii_281_v5.6.defline.txt.
Creinhardtii_281_v5.6.description.txt.
Creinhardtii_281_v5.6.protein_primaryTranscriptOnly.fa.

For the Arabidopsis protein set, Araport11 was chosen over TAIR10 because it is a comprehensive re-annotation of the Col-0 genome using 113 public RNA-seq data sets and other annotation contributions from the National Center for Biotechnology Information (NCBI), Uniprot, and labs conducting Arabidopsis research (https://www.araport.org/data/araport11). The initial, primary transcript-only protein lists contained 27,654, 38,334, and 17,741 proteins for Arabidopsis, Setaria, and Chlamydomonas, respectively.

4.2 | Protein family clustering and ortholog analysis

To reduce the overall number of proteins and generate a non-redundant protein set, the three-step hierarchical clustering algorithm

Cluster Database at High Identity with Tolerance (cd-hit, h3-version; http://weizhong-lab.ucsd.edu/webMGA/server/) was used on the protein list derived from the primary transcript only lists from each of the three organisms to identify those proteins that cluster together with ≥30% primary sequence identity to group closely related protein families with a representative sequence (Huang et al., 2010). To identify orthologs, proteins in each organism were searched against each other using BLASTP to identify those proteins with an e-value cutoff of 10^{-3} and a qcovs score ≥40% using both BLOSUM 45 and 62 matrices (Altschul et al., 1990). Orthologs were further identified using INPARANOID (Remm et al., 2001) and OrthoFinder (v2.5.5) using the default settings (Emms & Kelly, 2019).

4.3 | Protein localization prediction

To characterize the functionally unknown protein sets, analyses of the primary amino acid sequences were performed. Intra- and extracellular localization and signal peptide cleavage sites were predicted using TargetP 2.0 (Armenteros, Salvatore, et al., 2019), WoLF PSORT (Horton et al., 2007), and PredAlgo ((Tardif et al., 2012), more accurate only for *C. reinhardtii*), transmembrane domains were predicted using Phobius ((Kall et al., 2007), https://phobius.sbc.su.se/). Enrichment in cellular localization and transmembrane predictions of the unknown and Deep Green protein sets were performed using the hypergeometric or Fisher's exact test.

4.4 | Co-expression analysis

Co-expression analyses were performed on Chlamydomonas Deep Green proteins using two different datasets, as follows: (1) 18 diurnally expressed clusters and two unclustered groups (non-differentially expressed and non-expressed clusters) described in Chlamydomonas (Zones et al., 2015); (2) High-temperature and recovery inducible genes (HIGs and RIGs, respectively) identified previously (HIGs and RIGs are defined as transcripts that were induced for at least one-time point during high temperatures and recovery, respectively) (Zhang et al., 2022); Deep Green genes that are present in this RNA-seq datasets were used for the enrichment analysis. Clustvis heat map clustering (Metsalu & Jaak, 2015) was performed via correlation distance, completed clustering with tightest cluster first for rows and no clustering for columns. Gene co-expression networks in Chlamydomonas genes are derived from ChlamyNET (Romero-Campero et al., 2016). Graphical representation of the ChlamyNet cluster networks was performed using Cytoscape with an organic layout method (Smoot et al., 2011). This algorithm consists of a variant of the force-directed layout. Nodes produce repulsive forces, whereas edges induce attractive forces. Nodes are then placed such that the sum of these forces is minimized. The organic layout has the effect of exposing the clustering structure of a network. In particular, this layout tends to locate tightly connected nodes with many interactions or *hub nodes* together in central areas of the network. The over-representation hypher.test analysis was performed using the R programming language with a significant level of .05.

4.5 | Structural predictions

We used AlphaFold v2.1 to predict tertiary structures of the Deep Green proteins from their amino acid sequences. Five structural models were generated per protein and the models were ranked using the predicted local-distance difference test (pLDDT) scores (Jumper et al., 2021). The model with the highest pLDDT score was accepted as the most accurate structural prediction. Computations were carried out on NREL's Eagle High-Performance Computing (HPC) cluster. Structural predictions for protein sequences with more than 1,100 amino acid residues were run on graphics processing unit (GPU) nodes (with 16 GB Tesla V100 accelerators), while shorter sequences were run on central processing unit (CPU) nodes due to memory limitations. Twenty Arabidopsis, 26 Setaria, and 18 Chlamydomonas Deep Green protein sequences were predicted by SignalP-5.0 to contain signal peptides (Armenteros, Tsirigos, et al., 2019). For these sequences, the signal peptides were truncated *in-silico* prior to structure prediction. AlphaFold v2.1 structural prediction ran successfully on 457 out of 458 Arabidopsis, all 504 Setaria, and 382 out of 384 Chlamydomonas Deep Green proteins. AlphaFold v2.1 runtime errors occurred for the Arabidopsis protein AT1G21650.3 (1806 aa) and for two Chlamydomonas proteins, Cre04.g216050.t1.1 (3,691 aa after removal of predicted signal peptide) and Cre07.g314900.t1.1 (732 aa). These structures could not be predicted due to runtime errors of the HHBlits software that AlphaFold v2.1 uses for fast iterative protein sequence searching by HMM-HMM alignment. AlphaFold v2.1 has been documented to fold proteins that are at least 16 and at most 2,700 amino acid residues long. To perform structural homology analysis on the Deep Green proteins, proteins with a predicted tertiary structure having a confidence score higher than .5 were selected and FoldSeek (van Kempen et al., 2022) was used to generate structural alignments with proteins in the Protein Data Bank (PDB, version on 2021-06-01) using the parameters: --alignment-type 1 --tmscore-threshold 0 --max-seqs 2000.

4.6 | Protein disorder predictions and analyses

Protein order versus disorder based on overall secondary structure was quantified using a standalone version of the Intrinsically Unstructured Prediction (IUPred3) (Erdos et al., 2021) tool that was run on the NREL high-performance computing (HPC) cluster. In the current work, the long disorder prediction mode of IUPred3 was used along with the medium smoothing option that involves the Savitzky–Golay filter with parameters 19 and 5. IUPred3 returns a score, between 0 and 1, for each amino acid residue in the input protein sequence, which represents the probability of the given residue being part of a disordered region. Residues with scores equal to or exceeding .5 were considered to be disordered. Next, the percentage disorder (percentage of the total number of amino acid residues in a protein that are disordered) was quantified for each of the lead proteins from the entire proteomes of Arabidopsis, Setaria, and Chlamydomonas. The percentage disorder values for the Deep Green proteins, which constitute a subset of the set of all the lead proteins, were selected for additional analyses. The percentage of amino residues that are Ala, Pro, and Gly in each of the lead Arabidopsis, Setaria, and Chlamydomonas proteins were estimated as another measure of structural disorder using in-house Python code that involved the use of the Biopython library (Cock et al., 2009). Signal peptides were removed using SignalP5.0.

AUTHOR CONTRIBUTIONS

Research Design and Data Analysis: Overall design [JU]; Deep Green list assembly [EPK, PS, AN]; Deep Green list curation and GreenCut2 update [PS, HN, NNZ]; Localization analyses [EPK]; Chlamydomonas heat stress analyses [RZ, EMM, HN, NNZ]; diurnal and abiotic stress analyses [PS]; ChlamyNet analyses [CC]. Protein disorder predictions [AN]; Alphafold2 predictions [PSJ]; Structural comparisons and curation [JL, JC]. Manuscript writing: main draft [EPK, PS, JU, RZ, EMM]; editorial input and revisions [all authors].

ACKNOWLEDGMENTS

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding is provided by the Department of Energy's Office of Science Biological and Environmental Research. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work or allow others to do so, for the U.S. Government purposes. This research is supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Sciences Program grant no. DE-SC0020400. Erin Mattoon was supported by the William H. Danforth Fellowship in Plant Sciences, DDPSC start-up funding (to Ru Zhang), and Washington University in St. Louis.

CONFLICT OF INTEREST STATEMENT

The authors did not report any conflict of interest.

DATA AVAILABILITY STATEMENT

All of the data for this work has been provided in the Supplemental Tables accompanying this article. Additional data consisting of all of the tertiary structures for the AlphaFold structural predictions are also available on the NREL data catalog: https://data.nrel.gov/submissions/216 (DOI:10.7799/1970473).

ORCID

Eric P. Knoshaug https://orcid.org/0000-0002-5709-914X
Peipei Sun https://orcid.org/0000-0001-6448-4620
James Umen https://orcid.org/0000-0003-4094-9045

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance*, *2*, e201900429. https://doi.org/10.26508/lsa.201900429

Armenteros, J. J. A., Tsirigos, K. D., Sonderby, C. K., Petersen, T. N., Winther, O., Brunak, S., van Heijne, G., & Nielsen, H. (2019). SignalP5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*, 420–423. https://doi.org/10.1038/s41587-019-0036-z

Arthur, G., Emanuel, S. L., Heng, Y., & Wenqiang, Y. (2019). Building the GreenCut2 suite of proteins to unmask photosynthetic function and regulation. *Microbiology*, *165*, 697–718.

Basile, W., Oxana, S., Light, S., & Elofsson, A. (2017). High GC content causes orphan proteins to be intrinsically disordered. *PLoS Computational Biology*, *13*, e1005375. https://doi.org/10.1371/journal.pcbi.1005375

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The *Arabidopsis* information resource: Making and mining the 'gold standard' annotated reference plant genome. *Genesis*, *53*, 474–485. https://doi.org/10.1002/dvg.22877

Blaby-Haas, C. E., & Merchant, S. S. (2019). Comparative and functional algal genomics. *Annual Review of Plant Biology*, *70*, 605–638. https://doi.org/10.1146/annurev-arplant-050718-095841

Blaby-Hass, C. E., & de Crecy-Lagard, V. (2011). Mining high-throughput experimental data to link gene and function. *Trends in Biotechnology*, *29*, 174–182. https://doi.org/10.1016/j.tibtech.2011.01.001

Blasing, O. E., Gibon, Y., Gunther, M., Hohne, M., Morcuende, R., Osuna, D., Thimm, O., Usadel, B., Scheible, W. R., & Stitt, M. (2005). Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *The Plant Cell*, *17*, 3257–3281. https://doi.org/10.1105/tpc.105.035261

Blum, M., Chang, H., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., … Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*, D344–D354.

Bolger, M. E., Arsova, B., & Usadel, B. (2017). Plant genome and transcriptome annotations: From misconceptions to simple solutions. *Briefings in Bioinformatics*, *3*, bbw135–bbw113. https://doi.org/10.1093/bib/bbw135

Callaway, E. (2022). What's next for the AI protein folding revolution – AlphaFold, software that can predict the 3D shape of proteins, is already changing biology. *Nature*, *604*, 234–238. https://doi.org/10.1038/d41586-022-00997-5

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, *89*, 789–804. https://doi.org/10.1111/tpj.13415

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Dangol, A., Shavit, R., Yaakov, B., Strickler, S. R., Jander, G., & Tzin, V. (2022). Characterizing serotonin biosynthesis in *Setaria viridis* leaves and its effect on aphids. *Plant Molecular Biology*, *109*, 533–549. https://doi.org/10.1007/s11103-021-01239-4

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*, D427–D432.

Ellens, K. W., Christian, N., Singh, C., Satagopam, V. P., May, P., & Linster, C. L. (2017). Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Research*, *45*, 11495–11514. https://doi.org/10.1093/nar/gkx937

Emms, D. M., & Kelly, S. (2019). OrthoFinder: [hylogenetoc orthology inference for comparative genomics]. *Genome Biology*, *20*, 238. https://doi.org/10.1186/s13059-019-1832-y

Erdos, G., Pajkos, M., & Dosztanyi, Z. (2021). IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, *49*, W297–W303. https://doi.org/10.1093/nar/gkab408

Ferenczi, A., Pyott, D. E., Xipnotou, A., & Molnar, A. (2017). Efficient targeted DNA editing and replacement in *Chlamydomonas reinhardtii* using Cpf1 ribonucleoproteins and single-stranded DNA. *PNAS*, *114*, 13567–13572. https://doi.org/10.1073/pnas.1710597114

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*, D279–D285. https://doi.org/10.1093/nar/gkv1344

Fox, B. G., Goulding, C., Malkowski, M. G., Stewart, L., & Deacon, A. (2008). Structural genomics: From genes to structures with valuable materials and many questions in between. *Nature Methods*, *5*, 129–132. https://doi.org/10.1038/nmeth0208-129

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Hanson, A. D., Pribat, A., Waller, J. C., & de Crécy-Lagard, V. (2010). 'Unknown' proteins and 'orphan' enzymes: The missing half of the engineering parts list – And how to find it. *The Biochemical Journal*, *425*, 1–11. https://doi.org/10.1042/BJ20091328

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: Protein localization predictor. *Nucleic Acids Research*, *35*, W585–W587. https://doi.org/10.1093/nar/gkm259

Hu, H., Mauro-Herrera, M., & Doust, A. N. (2018). Domestication and improvement in the model C4 grass, Setaria. *Frontiers in Plant Science*, *9*, 11034. https://doi.org/10.3389/fpls.2018.00719

Huang, P., Shyu, C., Coelho, C. P., Cao, Y., & Brutnell, T. P. (2016). *Setaria viridis* as a model system to advance millet genetics and genomics. *Frontiers in Plant Science*, *7*, e99940. https://doi.org/10.3389/fpls.2016.01781

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT suite: A web server for clustering and comparing biologicla sequences. *Bioinformatics*, *26*, 680–682. https://doi.org/10.1093/bioinformatics/btq003

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kall, L., Krogh, A., & Sonnhammer, E. L. L. (2007). Advantages of combined transmembrane topology and signal peptide prediciton - the Phobius web server. *Nucleic Acids Research*, *35*, W429–W432. https://doi.org/10.1093/nar/gkm256

Karpowicz, S. J., Prochnik, S. E., Grossman, A. R., & Merchant, S. S. (2011). The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. *The Journal of Biological Chemistry*, *286*, 1427–1439. https://doi.org/10.1074/jbc.M111.233734

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B., Rogozin, I. B., Smirnov, S., Sorokin, A. V., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, *5*, R7. https://doi.org/10.1186/gb-2004-5-2-r7

Li, X., Zhang, R., Patena, W., Gang, S. S., Blum, S. R., Ivanova, N., Yue, R., Robertson, J. M., Lefebvre, P. A., Fitz-Gibbon, S. T., Grossman, A. R., & Jonikas, M. C. (2016). An indexed, mapped mutant library enables reverse genetics studies of biological processes in *Chlamydomonas reinhardtii*. *Plant Cell*, *28*, 367–387. https://doi.org/10.1105/tpc.15.00465

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Lovell, J. T., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Tavares, R., Nusinow, D. A., … Kellogg, E. A. (2020). A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nature Biotechnology*, *38*, 1203–1210. https://doi.org/10.1038/s41587-020-0681-2

Matt, G. Y., & Umen, J. G. (2018). Cell-type transcriptomes of the multicellular Green AlgaVolvox carteriYield insights into the evolutionary origins of germ and somatic differentiation programs. *G3 (Bethesda)*, *8*, 531–550. https://doi.org/10.1534/g3.117.300253

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., Marshall, W. F., Qu, L.-H., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., … Grossman, A. R. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, *318*, 245–250. https://doi.org/10.1126/science.1143609

Metsalu, T., & Jaak, V. (2015). Clustvis: A web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Research*, *43*, W566–W570. https://doi.org/10.1093/nar/gkv468

Mi, H., Muruganujan, A., & Thomas, P. D. (2012). PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, *41*, D377–D386. https://doi.org/10.1093/nar/gks1118

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: Expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, *44*, D336–D342. https://doi.org/10.1093/nar/gkv1194

Niehaus, T. D., Thamm, A. M., de Crécy-Lagard, V., & Hanson, A. D. (2015). Proteins of unknown biochemical function - a persistent problem and a roadmap to help overcome it. *Plant Physiology*, *959*, 00959.2015. https://doi.org/10.1104/pp.15.00959

Pant, S. R., Irigoyen, S., Doust, A. N., Scholthof, K.-B. G., & Mandadi, K. K. (2016). *Setaria*: A food crop and translational research model for C_4 grasses. *Frontiers in Plant Science*, *7*, 1885. https://doi.org/10.3389/fpls.2016.01885

Proost, S., Van Bel, M., Vaneechoutte, D., Van De Peer, Y., Inze, D., Mueller-Roeber, B., & Vandepoele, K. (2015). PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Research*, *43*, D974–D981. https://doi.org/10.1093/nar/gku986

Rasmussen, S., Barah, P., Suarez-Rodriguez, M. C., Bressendorff, S., Frijs, P., Costantino, P., Bones, A. M., Nielsen, H. B., & Mundy, J.

(2013). Transcriptome responses to combinations of stresses in Arabidopsis. *Plant Physiology*, *161*, 1783–1794. https://doi.org/10.1104/pp.112.210773

Remm, M., Storm, C. E. V., & Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparison. *Journal of Molecular Biology*, *314*, 1041–1052. https://doi.org/10.1006/jmbi.2000.5197

Romero-Campero, F. J., Perez-Hurtado, I., Lucas-Reina, E., Romero, J. M., & Valverde, F. (2016). ChlamyNET: A *Chlamydomonas* gene co-expression network reveals global properties of the transcriptome and the early setup of key co-expression patterns in the green lineage. *BMC Genomics*, *17*, 227. https://doi.org/10.1186/s12864-016-2564-y

Sasso, S., Herwig, S., Mittag, M., & Grossman, A. R. (2018). The natural history of model organisms: From molecular manipulation of domesticated *Chlamydomonas reinhardtii* to survival in nature. *eLife*, *7*, e39233. https://doi.org/10.7554/eLife.39233

Smoot, M., Ono, K., Ruscheinski, J., Peng-Liang, W., & Ideker, T. (2011). Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics*, *27*, 431–432. https://doi.org/10.1093/bioinformatics/btq675

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*, 951–960. https://doi.org/10.1093/bioinformatics/bti125

Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugie're, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G., Vallon, O., & Cournac, L. (2012). PredAlgo: A new subcellular localization prediction tool dedicated to green algae. *Molecular Biology and Evolution*, *29*, 3625–3639. https://doi.org/10.1093/molbev/mss178

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41. https://doi.org/10.1186/1471-2105-4-41

Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., Wei, S., Preece, J., Geniza, M. J., Jiao, Y., Lee, Y. K., Wang, B., Mulvaney, J., Chougule, K., Elser, J., Al-Bader, N., Kumari, S., Thomason, J., Kumar, V., … Ware, D. (2018). Gramene 2018: Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*, *46*, D1181–D1189. https://doi.org/10.1093/nar/gkx1111

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., & Vandepoele, K. (2022). PLAZA 5.0: Extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, *50*, D1468–D1474. https://doi.org/10.1093/nar/gkab1024

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Soding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein strucure search. *bioRxiv*.

Wakao, S., Shih, P. M., Guan, K., Schackwitz, W., Ye, J., Patel, D., Shih, R. M., Dent, R. M., Chovatia, M., Sharma, A., Martin, J., Wei, C. L., & Niyogi, K. K. (2021). Discovery of photosynthesis genes through whole-genome sequencing of acetate-requiring mutants of *Chlamydomonas reinhardtii*. *PLoS Genetics*, *17*, e1009725. https://doi.org/10.1371/journal.pgen.1009725

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, *26*, 889–895. https://doi.org/10.1093/bioinformatics/btq066

Zhang, N., Mattoon, E. M., McHargue, W., Venn, B., Zimmer, D., Pecani, K., Jeong, J., Anderson, C. M., Chen, C., Berry, J. C., Xia, M., Tzeng, S.-C., Becker, E., Pazouki, L., Evans, B., Cross, F., Cheng, J., Czymmek, K. J., Schroda, M., … Zhang, R. (2022). Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*. *Communications Biology*, *5*, 460. https://doi.org/10.1038/s42003-022-03359-z

Zhu, C., Yang, J., & Shyu, C. (2017). Setaria comes of age: Meeting report on the second international Setaria genetics conference. *Frontiers in Plant Science*, *8*, 555. https://doi.org/10.3389/fpls.2017.01562

Zones, J. M., Blaby, I. K., Merchant, S. S., & Umen, J. G. (2015). High-resolution profiling of a synchronized diurnal transcriptome from *Chlamydomonas reinhardtii* reveals continuous cell and metabolic differentiation. *Plant Cell*, *27*, 2743–2769. https://doi.org/10.1105/tpc.15.00498

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Knoshaug, E. P., Sun, P., Nag, A., Nguyen, H., Mattoon, E. M., Zhang, N., Liu, J., Chen, C., Cheng, J., Zhang, R., St. John, P., & Umen, J. (2023). Identification and preliminary characterization of conserved uncharacterized proteins from *Chlamydomonas reinhardtii*, *Arabidopsis thaliana*, and *Setaria viridis*. *Plant Direct*, *7*(12), e527. https://doi.org/10.1002/pld3.527