THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 24, pp. 21427–21439, June 17, 2011
Printed in the U.S.A.

The GreenCut2 Resource, a Phylogenomically Derived Inventory of Proteins Specific to the Plant Lineage*[S]

Received for publication, February 21, 2011, and in revised form, April 11, 2011 Published, JBC Papers in Press, April 22, 2011, DOI 10.1074/jbc.M111.233734

Steven J. Karpowicz[‡1], **Simon E. Prochnik**[§], **Arthur R. Grossman**[¶2], **and Sabeeha S. Merchant**[‡‖3]

From the [‡]*Department of Chemistry and Biochemistry and* [‖]*Institute for Genomics and Proteomics, UCLA, Los Angeles, California 90095, the* [§]*United States Department of Energy Joint Genome Institute, Walnut Creek, California 94598, and the* [¶]*Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305*

The plastid is a defining structure of photosynthetic eukaryotes and houses many plant-specific processes, including the light reactions, carbon fixation, pigment synthesis, and other primary metabolic processes. Identifying proteins associated with catalytic, structural, and regulatory functions that are unique to plastid-containing organisms is necessary to fully define the scope of plant biochemistry. Here, we performed phylogenomics on 20 genomes to compile a new inventory of 597 nucleus-encoded proteins conserved in plants and green algae but not in non-photosynthetic organisms. 286 of these proteins are of known function, whereas 311 are not characterized. This inventory was validated as applicable and relevant to diverse photosynthetic eukaryotes using an additional eight genomes from distantly related plants (including *Micromonas*, *Selaginella*, and soybean). Manual curation of the known proteins in the inventory established its importance to plastid biochemistry. To predict functions for the 52% of proteins of unknown function, we used sequence motifs, subcellular localization, co-expression analysis, and RNA abundance data. We demonstrate that 18% of the proteins in the inventory have functions outside the plastid and/or beyond green tissues. Although 32% of proteins in the inventory have homologs in all cyanobacteria, unexpectedly, 30% are eukaryote-specific. Finally, 8% of the proteins of unknown function share no similarity to any characterized protein and are plant lineage-specific. We present this annotated inventory of 597 proteins as a resource for functional analyses of plant-specific biochemistry.

The plastid is an organelle in plants and algae that evolved from a photosynthetic cyanobacterium after it was engulfed by an ancestral eukaryotic cell over 1.5 billion years ago (1, 2). How the endosymbiont became integral to host cell functions and evolved into a plastid is still under debate (3), but functions localized to the present day plastid depend on both plastid- and nucleus-encoded proteins. The latter are synthesized in the cytoplasm and imported into the organelle by a specific multiprotein complex composed of the translocon of the outer and inner chloroplast envelope membrane (TOC[4] and TIC, respectively) proteins (4, 5). Over 2000 proteins are estimated to be located in the plastid with the vast majority (>90%) encoded by genes in the nucleus (6–9). Many of the nucleus-encoded proteins that function within plastids are conserved among photosynthetic organisms. These conserved proteins function in processes such as the capture and utilization of excitation energy, carbohydrate metabolism, and the synthesis of key cellular metabolites (such as lipids, isoprenoids, pigments, and amino acids). Interestingly, however, many plastid-localized proteins have not yet been assigned a specific biochemical function.

The increasing availability of sequence information from diverse organisms has allowed the application of comparative genomics, or phylogenomics, to discover proteins specific to bacteria (10–12), cyanobacteria (13–15), fungi (16, 17), metazoa (18), archaea (19), and plastids (20). Additionally, computational attempts have been made to recognize protein families that are conserved in select plant genomes (21). However, the inventory of proteins exclusive to plants was only first explored in 2007 because the number of plant genomes available before then had been limited.

A previous phylogenomics analysis of green plants attempted to identify plant proteins associated with the plastid (22). In that study, orthologs (and recent paralogs) of proteins encoded by the *Chlamydomonas reinhardtii* genome were identified in the predicted proteomes of the angiosperm *Arabidopsis thaliana* (23), the moss *Physcomitrella patens* (24), and the marine, picoplanktonic algae *Ostreococcus tauri* (25) and *Ostreococcus lucimarinus* (26) but not in non-photosynthetic organisms. An inventory of 349 conserved proteins was generated and designated the "GreenCut" because it represented all of the protein families contained in a slice through the green lineage of the phylogenetic tree. However, the GreenCut was restricted in scope because of the limited number of genomes queried. In

* This work was supported, in whole or in part, by National Institutes of Health Grant GM07185, a Ruth L. Kirschstein National Research Service Award (to S. J. K.). This work was also supported by the Office of Science of the United States Department of Energy under Contract DE-AC02-05CH11231 (to S. E. P.).

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Tables S1–S4 and Figs. S1–S4 in Files 1–4.

[1] Supported in part by a departmental Majeti-Alapati Fellowship.
[2] Supported by National Science Foundation Grant MCB-095 1094 and United States Department of Energy Grant DE-FG02-07ER64427.
[3] Supported by United States Department of Energy Cooperative Agreement DE-FC02-02ER63421 (to David Eisenberg). To whom correspondence should be addressed: Dept. of Chemistry and Biochemistry, UCLA, 607 Charles E. Young Dr. E., Los Angeles, CA 90095-1569. Tel.: 310-825-8300; Fax: 310-206-1035; E-mail: sabeeha@chem.ucla.edu.

[4] The abbreviations used are: TOC, translocon of the outer chloroplast envelope membrane; TIC, translocon of the inner chloroplast envelope membrane; str., strain; K, known; KI, known with inferred function; U, unknown; UP, unknown with predicted function; Rubisco, ribulose-bisphosphate carboxylase/oxygenase; RNA-seq, RNA sequencing; RPKM, reads per kilobase of mappable sequence per million reads; LHC, light-harvesting chlorophyll-binding protein.

addition, the inclusion of two *Ostreococcus* species, which have reduced/specialized genomes and proteomes, constrained the output from the analysis.

Several additional plant genomes have been sequenced in the last 3 years, including those of the poplar tree *Populus trichocarpa* (27), the legume *Glycine max* (28), the spike moss *Selaginella moellendorffii*, and the green algae *Ostreococcus* sp. RCC890, *Volvox carteri* (29), and *Chlorella variabilis* NC64A (30). In addition, the annotation of other plant genomes, such as *Oryza sativa* (31, 32), has been updated. This new sequence information allows for the recognition of a plant lineage-specific inventory that represents a greater diversity of all green plants.

With the availability of this new genomic information, our goal was to generate an inventory of proteins unique to plastid-containing organisms. This inventory would contain fruitful targets for experimental studies of plant processes. Therefore, we performed a phylogenomics study to derive a set of proteins that is restricted to diverse organisms of the green lineage. We compared proteins encoded by eight plant genomes, but not by nine non-photosynthetic organisms, with proteins of five other photosynthetic eukaryotes (plants and diatoms) to establish a comprehensive set of green lineage proteins, which we designated the "GreenCut2." We verified the completeness and representative character of the protein inventory by comparing it with proteins encoded by the genomes of six additional photosynthetic eukaryotes. We annotated the GreenCut2 inventory by performing a meta-analysis of gene, mRNA, and protein data to generate new hypotheses concerning the activity of proteins of unknown function in the GreenCut2 and the roles of these proteins in plastid biology. This analysis suggested potential functions/activities for some of these proteins based on the presence of specific protein domains or motifs, subcellular location, and pattern of expression of the genes that encode them, thus identifying promising targets for future experimental work. Furthermore, the analysis suggests that there is a subset of proteins that is not directly associated with photosynthetic function or plastid biochemistry but that is still specific to the green lineage. Given their conservation, these proteins are likely to be critical for plant-specific processes and activities beyond photosynthesis.

EXPERIMENTAL PROCEDURES

GreenCut2 Algorithm—Predicted protein sequences from sequenced genomes were subjected to phylogenomics analysis using methods described previously (22). Briefly, WU-BLASTP searches were conducted between the *C. reinhardtii* (JGI v3.1) predicted proteome and the predicted proteomes from a phylogenetically diverse set of organisms (listed below). A mutual best BLASTP hit (E-value $<1e-10$) was used to establish orthology to a *Chlamydomonas* protein. Additional eukaryotic proteins that were not a mutual best hit but had >50% amino acid identity to a *Chlamydomonas* protein within an ortholog cluster were selected as in-paralogs (co-orthologs throughout). In-paralogs are genes that duplicated within a species after it diverged from another species under consideration (34). In-paralogs are by definition less diverged from each other than are the orthologs in the two species that diverged at the speciation event in question. The sequence identity threshold was chosen empirically to recover closely related co-orthologs without generating overly large ortholog clusters.

Two major criteria were used to generate the inventory of GreenCut2 proteins (supplemental File 1, Fig. S1). First, the *Chlamydomonas* proteins of the GreenCut2 must have an ortholog encoded by the nuclear genomes of the green lineage organisms *A. thaliana* (TAIR v8), *P. patens* (JGI v1.1), *O. sativa* (*japonica*) (TIGR v5.0), *P. trichocarpa* (JGI v1.1), and one of the three *Ostreococcus* species with fully sequenced genomes (*O. lucimarinus* (JGI v2.0), *O. tauri* (JGI v2.0), or *Ostreococcus* sp. RCC809 (JGI v2.0)). Second, proteins with orthologs in the green lineage organisms listed above were only included in the GreenCut2 if they had no ortholog in *Pseudomonas aeruginosa* str. PA01, *Staphylococcus aureus* subsp. *aureus* str. N315, *Dictyostelium discoideum* AX4, *Phytophthora sojae*, *Neurospora crassa* OR74A, *Methanosarcina acetivorans* str. C2A, *Sulfolobus solfataricus* str. P2, *Caenorhabditis elegans*, and *Homo sapiens*. Searches for orthologs in *Cyanidioschyzon merolae* str. 10D, *Thalassiosira pseudonana* (JGI v2.0), and *Phaeodactylum tricornutum* (JGI v3.0) also were conducted, but for inclusion in the GreenCut2, we did not require that a *Chlamydomonas* protein have an ortholog in these organisms. The inventory of orthologs produced in this analysis is presented in supplemental File 2 (Table S2).

Orthologs in Other Genomes—Arabidopsis GreenCut2 proteins were used to query the genomes of *Micromonas pusilla* CCMP1545 (JGI v2.0), *Coccomyxa* sp. C-169 (JGI v2.0), *G. max* (JGI v1.0), *Sorghum bicolor* (JGI v1.0), *S. moellendorffii* (JGI v1.0), and *Fragilariopsis cylindrus* (JGI v1.0) to identify potential orthologs. When a BLASTP search (E-value $<1e-10$) indicated that a potential ortholog was not encoded by one of these genomes, a TBLASTN search (E-value $<1e-5$) against the genomic sequence was conducted using the *Arabidopsis* GreenCut2 protein as the query sequence. The GreenCut2 proteins that are not present in *M. pusilla*, *Coccomyxa* sp. C-169, *G. max*, *S. bicolor*, *S. moellendorffii*, and *F. cylindrus* are given in supplemental File 3 (Table S3).

All proteins encoded by the nuclear genome of *Chlamydomonas* were used in a BLASTP search (E-value $<1e-10$) against the chloroplast genomes of other GreenCut2 plants to detect potential orthologs. Similarly, proteins encoded by the chloroplast genome of *Chlamydomonas* were used as queries for BLASTP searches to detect potential orthologs in the nuclear genomes of the other GreenCut2 organisms. Proteins with mutual best hits were screened for conservation in non-photosynthetic organisms by BLASTP searches (E-value $<1e-10$) against the NCBI non-redundant database.

Protein Data—Subcellular localization data for *Arabidopsis* proteins determined from proteomics studies were obtained from the Plant Proteome Database (35), the Plastid Protein Database (36), the Sub-cellular Localization Database for *Arabidopsis* proteins (SUBA) (37), and AT-CHLORO (38). The *Arabidopsis* Information Resource (TAIR) (23) and SUBA assigned subcellular localizations based on GFP-hybrid protein experiments. *Chlamydomonas* proteins were assigned a mitochondrial localization based on their identification in purified mitochondria (39). Localizations also were assigned based on

reported experimental studies. Finally, *Arabidopsis* GreenCut2 proteins were used for TargetP (7) and Wolf-Psort (40) predictions. In those cases in which the two algorithms yielded different results, the localization predicted by TargetP was chosen except when TargetP yielded no prediction.

Protein Function—Proteins were assigned to one of the following function classes: known (K), known with inferred function (KI), unknown (U), or unknown with predicted function (UP). A protein was classified as K if a publication defined its function or activity. KI proteins have orthologs (with unknown function) within the green plants, but we were able to infer the function of KI proteins because they have sequence similarity (BLASTP E-value $<1e-10$) to other proteins with known functions. Proteins classified as U did not contain homology to any known protein or have domains that would suggest a biochemical function. A UP assignment for an undefined protein was based on the presence of a functional domain or on relevant literature that suggested a function based on a mutant phenotype. Literature searches used protein identifier numbers to identify recent research relating to the GreenCut2 proteins. Pfam (41) domain predictions (v24.0) for both *Arabidopsis* and *Chlamydomonas* GreenCut2 proteins were obtained from the Pfam web site. Additional domain predictions (FIGfams) were accessed at The SEED database (42).

MapMan (43) categories were retrieved using the *Arabidopsis* locus identifier numbers. MapMan bin classifiers and annotated functions were used to sort the proteins into general functional categories. Unknown proteins with an informative domain(s) were assigned to functional groups based on the potential activity associated with that domain(s) and on the characteristics of potential interacting proteins (44, 45). Every GreenCut2 protein was assigned to a single functional category for simplicity of classification, although in some cases, the assignment was arbitrary because the protein could have been assigned to more than one functional category.

False Positive/Negative—The false negative rate was determined using 21 previously characterized nucleus-encoded proteins involved in photosynthesis that are known to be conserved in all photosynthetic organisms in the green lineage. These proteins are PsbO, -P, -Q, -R, -S, -W, -X, -Y, PsaD, -E, -F, -G, -H, -K, -L, -O, plastocyanin, ferredoxin, ferredoxin-NADP reductase, Rubisco small subunit, and phosphoribulokinase. Of these 21 proteins, 19 (90%) are recovered in the GreenCut2. The two proteins not recovered, PsbR and PsbX, have mutual best BLASTP hit E-values larger than our threshold of $1e-10$ ($>2e-6$ for PsbR and $>1e-3$ for PsbX). This is a typical problem for identifying orthologous sequences of moderately divergent, small proteins. PsbR and PsbX of *Chlamydomonas* are 121 and 101 amino acids long, respectively, and exhibit only 42 and 35% sequence coverage relative to the *Arabidopsis* orthologs.

The false positive rate was determined by manual curation/analysis of the GreenCut2 proteins. Specifically, the presence of orthologous proteins in the complete non-redundant database was investigated to determine whether the protein under consideration was plant lineage-specific. GreenCut2 proteins with orthologs in non-photosynthetic organisms (a non-photosynthetic organism is among the top five BLASTP hits of the non-redundant database) were flagged as false positives. Proteins of known function that localized to subcellular compartments other than the plastid were also investigated. The 17 potential false positives identified in the complete inventory of 597 proteins and the reasons for placing them in this category are given in supplemental File 4 (Table S4).

RNA Abundance Determined from Microarrays and RNA-seq—*Arabidopsis* organ development microarray expression data normalized by gcRMA (46) were used to evaluate organ-specific abundances of *Arabidopsis* transcripts encoding GreenCut2 proteins. MATLAB software (MathWorks) was used to cluster and display microarray values as a dendrogram using default hierarchical clustering parameters. Genes whose linkage value in the hierarchical tree, based on similarity of expression patterns, was 0.7 or greater were assigned as one node in the dendrogram. Separate nodes containing data derived from the same organ or from tissues with similar phenotypic characteristics were considered members of a single expression category. For example, two nodes containing intensity values from green leaf tissues were combined into the green organ expression category. Seven expression categories were identified. To evaluate the specificity of each GreenCut2 transcript, the following procedure was implemented. Within an expression category, microarray intensities for an individual transcript were averaged. The expression category average was compared with the sum of the averaged values for that specific transcript from all of the categories. If the average intensity in one category was greater than 25% of the total summed average intensities (summed from all seven categories), then that transcript was defined as organ/tissue-specific. This threshold was determined based on intensity values of transcripts designated as organ-specific by Schmid *et al.* (46).

In addition, RNA-seq data for *Chlamydomonas* (47) and *Arabidopsis* (48) provided quantitative mRNA abundances for transcripts encoding GreenCut2 proteins in different *Arabidopsis* organ types and under different growth conditions for *Chlamydomonas*. The transcript abundance values are represented in reads per kilobase of mappable sequence per million reads (RPKM) (47). Briefly, an RPKM value for a particular transcript is derived from the number of nucleotides comprising a sequenced mRNA fragment that uniquely map to an underlying gene model, and the sum of the mapped nucleotides is normalized to transcript length and sequencing depth. For each data set, GreenCut2 transcripts were grouped into bins using RPKM values based on a $\log_{10}$ scale. For example, transcripts with 0.2–1 RPKM were placed in one bin, whereas those with 2–10 RPKM were placed into another bin and so on. Each bin was treated as a single data point to plot the distribution of transcript abundances.

Searches for Cyanobacterial Homologs—Cyanobacterial homologs of GreenCut2 proteins were determined by the best BLASTP hit (E-value $<1e-4$) of an *Arabidopsis* GreenCut2 protein to a cyanobacterial protein. The *Arabidopsis* GreenCut2 proteins were used for this analysis because the gene models underlying the predicted protein sequences are generally of the highest quality. *Arabidopsis* co-orthologs of GreenCut2 proteins generated the same relationships during searches for cyanobacterial homologs (they matched the same cyanobacterial proteins with an E-value threshold of $1e-4$ or less); thus,

each *Arabidopsis* ortholog and its co-ortholog(s) were treated as a single protein in this analysis.

Arabidopsis protein sequences were downloaded from TAIR (TAIR v8), whereas protein sequences deduced from 37 finished cyanobacterial genomes (supplemental File 1, Table S1) were downloaded from the Integrated Microbial Genomes database (September 13, 2009). Synteny between the cyanobacterial genomes was visualized on The SEED database. Yeast two-hybrid interaction partners (45) were accessed on Cyanobase (49). One-way analysis of variance tests using Origin 7.5 software (OriginLab) were used to evaluate enrichment of functional categories within the bins of cyanobacterial genomes that contain homologs to GreenCut2 proteins.

RESULTS AND DISCUSSION

Generation of Inventory

C. reinhardtii proteins (FM 3.1 set of gene models) were used to identify orthologs in *A. thaliana*, *P. patens*, *O. sativa*, *P. trichocarpa*, and *O. tauri*, *O. lucimarinus*, and *Ostreococcus* sp. RCC809. Those proteins with orthologs in all land plants and at least one of the *Ostreococcus* species were retained (see below). This set of proteins was compared with proteins in a group of non-photosynthetic organisms (see "Experimental Procedures" for the list of these organisms), and those proteins that had orthologs in any of the non-photosynthetic organisms were removed from consideration (supplemental File 1, Fig. S1).

The *Ostreococcus* species were considered useful in the phylogenomics analysis because they provide data from divergent species within the chlorophyte lineage (Fig. 1). They are cosmopolitan, marine algae found throughout the world's oceans. However, their genomes are small, and each species is adapted to their environmental niche (50). Therefore, they may have lost some biochemical functions that are present in other plants. To minimize the impact of specialization, we sampled three *Ostreococcus* genomes, which provide a broad base of prasinophyte gene representation. We required that an ortholog be encoded by a gene model in one or more of the *Ostreococcus* genomes. In effect, we attempted to sample an *Ostreococcus* "pan-genome" that represents protein-encoding genes that are present anywhere in the genus. As a result, 597 proteins were captured in the inventory. Had we required that a protein be encoded by all three *Ostreococcus* genomes, we would have lost 126 proteins, each of which is presumably dispensable in a particular marine niche occupied by a specialized *Ostreococcus* species.

This set of 597 proteins is designated the GreenCut2 (supplemental File 2, Table S2). As a consequence of whole genome duplications in *Arabidopsis*, the 597 *Chlamydomonas* GreenCut2 proteins capture 710 *Arabidopsis* co-orthologs. GreenCut2 proteins were assigned to the general categories K, KI, U, and UP (see "Experimental Procedures").

Subgroups of GreenCut2

To investigate whether GreenCut2 proteins are conserved in photosynthetic organisms that are not affiliated with the green lineage, we identified GreenCut2 orthologs encoded by the genomes of the red alga *C. merolae* (51) and the diatoms *T. pseudonana* and *P. tricornutum* (52, 53). *C. merolae* is a member of the plant kingdom whose ancestor diverged from the green plant lineage (Fig. 1). Diatoms, in contrast, are heterokonts. They acquired their plastid through a secondary endosymbiosis (54, 55), a process in which an endosymbiont-containing eukaryote is engulfed by another free-living eukaryote.

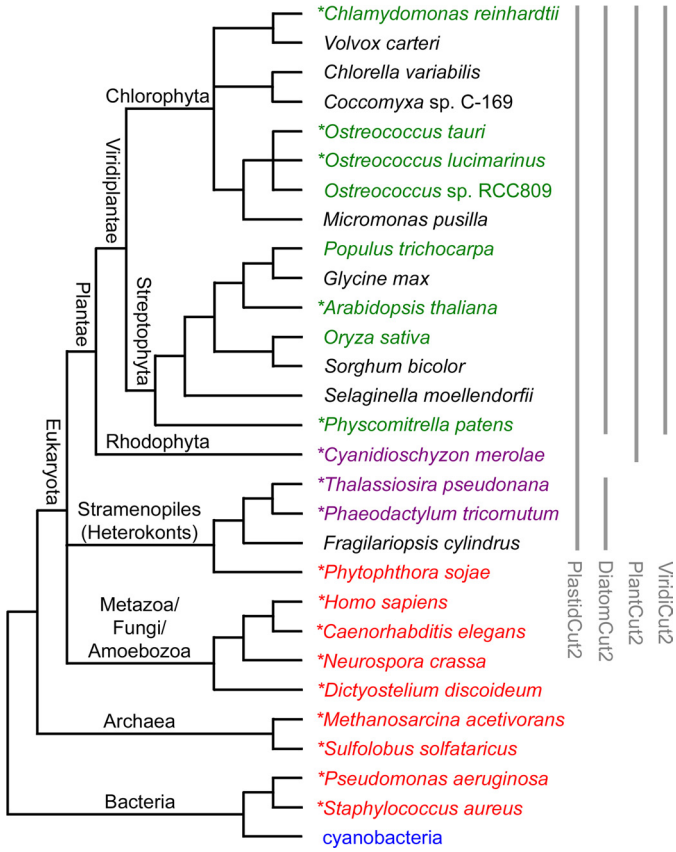

FIGURE 1. **Taxonomic tree of organisms used to build and test GreenCut2.** Eight photosynthetic organisms (*green*) were used in the construction of the GreenCut2. All eight organisms must encode orthologs of a protein for the protein to be included in the GreenCut2 except for the *Ostreococcus* species where an ortholog is only required to be in one of the three species. Conserved proteins encoded by any of the nine non-photosynthetic organisms (*red*) were excluded from the GreenCut2. A subset of GreenCut2 orthologs was identified in the three non-green, photosynthetic eukaryotes (*purple*). The genomes of other eukaryotes (*black*) were searched for orthologs of GreenCut2 proteins as validation of the inventory. Cyanobacteria (*blue*) were searched for homologs of GreenCut2 proteins. Organisms that contribute proteins to a subset of the GreenCut2 are bounded by a *gray line* to the *right* of the tree. The general taxonomic group to which an organism belongs is shown on the tree. The unrooted taxonomic tree represents evolutionary relationships between organisms but not evolutionary distance. *Asterisks* indicate those organisms whose genomes were used to determine GreenCut version 1.

Among the 597 GreenCut2 proteins, 124 are found in the genomes of green plants, *C. merolae*, and at least one diatom (supplemental File 1, Fig. S2). This set of 124 proteins has been designated the "PlastidCut2." The genes for PlastidCut2 proteins are conserved in the nuclear genomes of the diverse plastid-containing, photosynthetic eukaryotes investigated (within and outside the plant lineage). Therefore, the name PlastidCut2 is independent of the eukaryotes' evolutionary history. These proteins are likely to be critically important for plastid metabolism, including photosynthesis, which is suggested by an enrichment of functions associated with photosynthesis among the K category proteins and the greater fraction of PlastidCut2

TABLE 1
Proteins of known and unknown function
Numbers for version 1 of the GreenCut (22) were determined in October 2006, and numbers for version 2 (this study) were determined in August 2010.

Proteins	PlastidCut	DiatomCut-PlastidCut	PlantCut-PlastidCut	ViridiCut	Total	Percent
GreenCut2	124	96	65	312	597	
K	60	45	32	149	286	48
U	64	51	33	163	311	52
GreenCut v1	90	60	27	172	349	
K	29	18	9	79	135	39
U	61	42	18	93	214	61

proteins found in cyanobacteria (see below). Surprisingly, despite their high degree of conservation, the functions of 52% (64 of 124) of PlastidCut2 proteins are not known (Table 1).

The subset of GreenCut2 proteins found in the genome of at least one diatom is labeled "DiatomCut2" (supplemental File 1, Fig. S2). The proteins of this subgroup include the 124 proteins of the PlastidCut2 plus a set of 96 proteins that are not apparently conserved/encoded by the *C. merolae* genome. Similarly, the set of proteins found in green plants and *C. merolae* is labeled "PlantCut2," which includes PlastidCut2 proteins plus 65 additional proteins not apparently conserved/encoded by either of the diatom genomes analyzed in this study (supplemental File 1, Fig. S2). Green plants contain 312 proteins designated the "ViridiCut2." These proteins are not encoded by the genome of *C. merolae*, *P. tricornutum*, or *T. pseudonana* (supplemental File 1, Fig. S2). The ViridiCut2 is likely enriched in green lineage-specific functions, such as mechanisms of chlorophyll *a*/*b* protein regulation.

Validation of GreenCut2

For practical reasons, we used only a subset of genomes representing a divergent collection of reference organisms to generate the GreenCut2. To validate our choice of organisms, we tested the predicted proteomes of recently sequenced plants, algae, and diatoms.

Land Plants—To assess the conservation of GreenCut2 proteins in land plants, the genomes of *G. max* (soybean), *S. bicolor* (cereal grass), and *S. moellendorffii* (spike moss), which occupy phylogenetically distinct positions in the green plant tree of life relative to the plants used for generation of the GreenCut2 (Fig. 1), were searched for orthologs of GreenCut2 proteins. The analysis demonstrated that the genomes of *G. max*, *S. bicolor*, and *S. moellendorffii* may not encode one, one, and three GreenCut2 proteins, respectively (supplemental File 3, Table S3). The genes encoding these proteins may lie in genomic regions missing from the current genome assemblies, or the genes may have been selectively lost. Overall, the presence of genes encoding almost all (99%, or 592 of 597) of the GreenCut2 proteins in three additional plant genomes (a legume, a grass, and a fern), which are divergent from other green plants used in the construction of the GreenCut2, provides further evidence that the inventory of proteins in the GreenCut2 is especially relevant to and representative of all land plants of the green lineage and that the number of false positives is likely to be very low.

Algae—We queried the predicted proteomes of the chlorophyte lineage algae *V. carteri*, *C. variabilis* NC64A, *Coccomyxa* sp. C-169, and *M. pusilla* (56) (Fig. 1) for orthologs to the *Chlamydomonas* protein set. The *V. carteri* genome encodes 100% of the GreenCut2 proteins, the trebouxiophyte algae *C. variabilis* and *Coccomyxa* encode 96 and 89%, respectively, and *M. pusilla* encodes 89%. The GreenCut2 proteins that were not identified in these algae (supplemental File 3, Table S3) may be encoded by genes located in regions missing from the genome assembly, may be present on unsequenced chloroplast genomes, or may have been lost during genome reduction.

We note that of the 597 GreenCut2 proteins in *Chlamydomonas* 105 are missing in at least one of the other green algae (*V. carteri*, *C. variabilis*, *Coccomyxa*, *Ostreococcus* spp., and *M. pusilla*). With a few exceptions in the trebouxiophyte lineage (supplemental File 3, Table S3), there does not appear to be a consistent pattern of GreenCut2 protein loss among members of the Chlorophyta. However, we did observe a bias toward the loss of ViridiCut2 proteins ($p = 2e{-}4$). The above results suggest that the adaptation of algae to specific environmental niches could lead to genome specialization and/or reduction that is reflected in the loss of GreenCut2 proteins. Together with the results for land plants, we therefore suggest that the extent of conservation of the GreenCut2 inventory in a plant could serve as an indicator of a particular genome's specialization.

Diatoms—Interestingly, the diatoms *T. pseudonana* and *P. tricornutum* together appear to encode only a relatively small number (220 of 597) of GreenCut2 proteins. It was not clear whether this is attributable to reduced genome content due to specialization to their habitats or incomplete genome sequence assembly and gene prediction. To help address this question, the draft genome of the psychrophilic diatom *F. cylindrus* was queried for orthologs of GreenCut2 proteins. We identified 192 GreenCut2 proteins encoded in the *F. cylindrus* data set with 181 of these proteins representing 82% (181 of 220) of DiatomCut2 proteins. Because the inventory of GreenCut2 proteins is similar in the three diverse diatoms, the reduced number of GreenCut2 proteins in diatoms suggests that several core plastid functions in the green lineage are either not critical in diatoms or are performed by different pathways/processes, which makes a compelling case for studies of plastid biology in diatoms.

Interestingly, the *F. cylindrus* genome encodes 11 GreenCut2 proteins not found in *T. pseudonana* and *P. tricornutum* (supplemental File 3, Table S3). One of these is the copper-containing protein plastocyanin, which was previously shown also to be present in the oceanic diatom *Thalassiosira oceanica* (57). The demand by *F. cylindrus* for copper cofactor during plastocyanin production is presumably met by the

copper concentration in the Antarctic Ocean, which is similar to other oceanic waters (58). This is an excellent example of selective retention of a protein in an environment where it can be useful *versus* loss in organisms that occupy a different, perhaps copper-deficient niche.

Determination of False Positives/Negatives—We chose a moderate stringency criterion for determining orthologous relationships between GreenCut2 proteins (E-value $<1e-10$) to balance the capture of false positives *versus* the appearance of false negatives. Based on manual curation of all of the GreenCut2 proteins, the false positive frequency was estimated at 2.8% (see "Experimental Procedures" and supplemental File 4, Table S4). In an attempt to measure the exclusion of orthologous proteins from the GreenCut2, we examined the behavior of an inventory of previously characterized, nucleus-encoded proteins that are involved in photosynthesis and are known to be conserved in all green photosynthetic organisms. From this analysis, a false negative frequency (failure to detect legitimate orthologous pairs) of ~10% was estimated (see "Experimental Procedures"). There are a number of reasons why orthologs may be excluded from the GreenCut2. Some proteins with conserved functions among organisms may have diverged such that the identity criterion used for ortholog predictions is no longer adequate. This is particularly true for small proteins, such as PsbX and PsbR. Orthologous relationships can also be obscured by the expansion of protein families within a genome, such as with the light-harvesting chlorophyll-binding protein (LHC) family (discussed below), because mutual best hits cannot be identified. Furthermore, incomplete gene model predictions for any of the organisms used in our analysis might prevent identification of mutual best BLASTP hits. In *S. moellendorffii*, for example, 10% of the GreenCut2 proteins had to be identified by TBLASTN rather than BLASTP. Finally, we note that the reduced genomes of the *Ostreococcus* species are missing some proteins present in other algae and green plants, such as SQUAMOSA promoter-binding protein domain-containing transcription factors.

Several notable protein families associated with plants were not recovered in the GreenCut2 for various reasons. *Chlamydomonas* gene models for subunits of TIC are incomplete and, in some cases, highly diverged. For example, CrTic55 is not identified as an ortholog of AtTic55 in this work. A second family of proteins not fully captured in the GreenCut2 is that of the LHCs. Because the LHCs of *Chlamydomonas* are very similar to each other, co-orthologous relationships among these proteins interfere with identification of genuine one-to-one orthologous relationships between plants.

Plastid-encoded proteins were not considered in this analysis. Therefore, proteins encoded by the nuclear genome of one plant but by the plastid genome of another plant would not be recovered in the GreenCut2. However, manual curation suggests that this does not impact our results. For example, TufA is encoded by the *Chlamydomonas* chloroplast genome but by the nuclear genomes of other plants. Nonetheless, it does not belong in the GreenCut2 because TufA orthologs are also found in non-photosynthetic organisms. We did not focus on proteins encoded exclusively on plastid genomes because previous studies have elaborated on this subject (20, 59), and they have clear relevance to plastid biology.

TABLE 2
Subcellular localization

Experimental localizations were determined from the literature based on subproteomes of purified organelles or visualization of GFP fusion proteins. Predicted localizations were made using TargetP and Wolf-Psort.

GreenCut2	Sum	Experimental Predicted	Plastid	Mitochondria	Nucleus	Other
PlastidCut2	124	100 24	92 12	5 7	0 1	3 4
DiatomCut2 −PlastidCut2	96	66 30	49 15	2 2	0 6	15 7
PlantCut2 −PlastidCut2	65	38 27	33 8	1 5	0 7	4 7
ViridiCut2	312	175 137	142 63	10 17	6 30	17 27
Total	597	379 218	316 98	18 31	6 44	39 45
		597	414	49	50	84

Functional Meta-analysis of Localization

Plastid—Of the proteins in the PlastidCut2, 84% (104 of 124) were experimentally localized to or are predicted to be in the plastid (Table 2); of these, 50 are in the U/UP groups. In comparison, 52% (316 of 597) of all GreenCut2 proteins were experimentally localized to the chloroplast (supplemental File 2, Table S2). Because many GreenCut2 proteins are localized to plastids, they likely are involved in plastid-specific functions. However, it is very intriguing that 6 and 11% of the PlastidCut2 and GreenCut2 proteins, respectively, are experimentally located elsewhere than the plastid.

Nucleus or Mitochondria—Not all GreenCut2 proteins need to be localized to the plastid to be involved in plastid function. Proteins located elsewhere in the cell may be involved in regulating nuclear genes encoding chloroplast proteins (such as HY5; see below), participate in the biogenesis of the plastid and its components (such as CrANK22, a cytosolic chaperone for plastid membrane proteins), or have evolved independently in the plant lineage to function in plant-specific processes (such as PEX13; see below). Although numerous plastid proteins have been experimentally localized, the placement of proteins in the mitochondrion or nucleus is based mostly on prediction algorithms. A combination of experimental and informatic evidence suggests that 49 (8.2%) of the GreenCut2 proteins are localized to the mitochondrion with experimental evidence for 18 of the 49 proteins. Similarly, 50 (8.3%) GreenCut2 proteins are thought to be located in the nucleus, although experimental evidence supports the localization of only six of these. This result suggests that there has been much less experimental work to demonstrate the subcellular localization of green lineage proteins present in organelles other than the plastid and/or that TargetP and Wolf-Psort may overpredict the number of GreenCut2 proteins located in the mitochondrion and nucleus.

Five nuclear transcription factors of known function are present in the GreenCut2, including CrHY5 (AtHY5; At5g11260), which functions in chloroplast maturation in response to light signals (60, 61). Exposure of an *Arabidopsis hy5* mutant to UV-B irradiation causes reduced accumulation of the *AtFAO3* (At2g22650) transcript, which encodes an FAD-dependent oxidoreductase of the GreenCut2 (62). This protein is predicted to localize to mitochondria. This result suggests

that some GreenCut2 proteins, like HY5, may integrate activities associated with multiple cellular compartments.

Other Locations—There are 84 proteins in the GreenCut2, representing 14% of the total, that are not predicted to be localized to the chloroplast, mitochondrion, or nucleus. 34 of these proteins are predicted to be cytosolic, but only seven have been experimentally localized to the cytosol. Furthermore, 31 GreenCut2 proteins have been experimentally shown to be present in Golgi, endoplasmic reticulum, endosomes, peroxisomes, or plasma membranes. An additional 10 proteins are predicted to be in endosomes, peroxisomes, or plasma membranes, whereas two transmembrane proteins have not been localized to a specific cellular compartment.

Peroxisomes display significant diversity among organisms (63), and the peroxisomes of plants, although less studied than their animal and yeast counterparts, have divergent features in their matrix protein import machinery (64). An example is the GreenCut2 peroxisomal protein AtPEX13 (At3g07560) (65). AtPEX13 interacts with the peroxisomal targeting sequence receptor AtPEX7 and functions in docking proteins to the peroxisomal import complex, thus facilitating their transit into the organelle. Another GreenCut2 protein localized to the peroxisome is AtLACS7 (At5g27600) (66), a long-chain acyl-CoA synthetase. Because AtLACS7 contains both a type I and type II peroxisomal targeting sequence, it, like AtPEX13, may bind the type II peroxisomal targeting sequence receptor AtPEX7 and potentially interact with GreenCut2 protein AtPEX13, although this is highly speculative. In sum, there are various aspects of peroxisome metabolism, such as glycolate metabolism, which is associated with photorespiration, and the glyoxylate cycle, which is associated with fatty acid utilization (67), that have been tailored to meet the biological needs of plants, likely explaining the inclusion of peroxisomal proteins in the GreenCut2.

Functional Meta-analysis of Domains and Activities

To elucidate the diversity of functions performed by proteins of the GreenCut2, when possible, the proteins were assigned potential biochemical functions/activities based on both experimental and informatic data. U/UP proteins were sorted into broad functional groups based on gene ontology terms and the molecular functions of predicted domains (Fig. 2*A* and supplemental File 1, Fig. S3). We placed 63% of U/UP proteins into specific functional groups.

Photosynthesis, Redox, and Pigments—Among the proteins belonging to specific functional groups, those associated with photosynthetic processes have been most thoroughly characterized. Thus, most proteins in the "photosynthesis" category have known functions (59 of 62). Chloroplast localization is known or predicted for all proteins in this category. Proteins of unknown function in the photosynthesis category include CrCGL30 (At1g77090), which has sequence similarity to PsbP, and CrCGL160 (At2g31040), which has orthologs encoded in ATP synthase operons in cyanobacterial genomes and is related to ATP synthase subunit I. Recently, a peripheral membrane protein was visualized in a photosystem I crystal structure that was in physical proximity to PsaK and Lhca3 (68). Although the identity of this protein is not known, a candidate for this protein is the photosynthesis protein of unknown function CrCGL40 (At1g49975), a small polypeptide related to PsaN.

Redox proteins are critical for the acclimation of photosynthetic cells to changing intracellular redox conditions. The "redox" category consists of 16 K/KI and 16 UP proteins. Six proteins in this category are thioredoxins, three are ferredoxins, one is a rubredoxin, and one is a glutaredoxin. Ferredoxins are 4Fe-4S cluster proteins that accept electrons from photosystem I and deliver them to enzymes that require reductant to perform their catalytic functions. The ferredoxins CrFDX4 (At4g14890) and CrFDX6 (At1g32550) have been described in *Chlamydomonas* (69), but their substrate specificity remains unknown. The most well studied ferredoxin, CrPETF (At2g27510), corresponding to the photosystem I-affiliated leaf ferredoxin, is also recovered in the GreenCut2 but was placed in the photosynthesis category. In contrast, rubredoxins have an Fe-$(SCys)_4$ domain and appear to play a role in the protection of cells from oxidative damage. A mutant of *AtENH1* (At5g17170), which encodes a rubredoxin-like protein, exhibits elevated levels of reactive oxygen species in plastids and decreased tolerance to high salt conditions (70).

Chlorophyll and carotenoid metabolism are necessary for photosynthetic function and consequently are well studied processes. The "pigment" group comprises 24 K and four UP proteins. A high proportion of pigment category proteins (16 of 28) is conserved in the PlastidCut2 relative to the ViridiCut2 (Fig. 2, *B* and *C*; $p = 1e{-}5$), which suggests that the genes encoding pigment biosynthesis enzymes are highly conserved in all photosynthetic eukaryotes as has been shown for photosynthetic prokaryotes (11). One of the proteins of unknown function that is associated with pigment biosynthesis is CrVDR1 (At2g21860), which is related to violaxanthin de-epoxidase, although its biochemical activity has not been determined. The characterized *Arabidopsis* violaxanthin de-epoxidase gene *AtVDE1* (At1g08550) does not have a homolog encoded by the extant *Chlamydomonas* genome. We expect that CrVDR1 may participate in the regulation of pigment synthesis or have a novel catalytic activity in carotenoid/xanthophyll biosynthesis, and given that it is conserved in all plants whereas VDE1 is not, it must act in a more critical pathway than the xanthophyll cycle.

Macromolecular Metabolism and Signaling—The category designated "protein metabolism" includes 93 proteins with activities associated with the maturation and degradation of polypeptides. 26 proteins in this group are proteases and peptidases, and six of the 14 proteases that have a known function are components of the plastidic Clp protease (71). Although many of the predicted proteases may be involved in degradation of plastid proteins, some may be specific to the maturation of proteins incorporated into functional protein complexes or the processing of polypeptides as they are imported into plastids, similar to the plastidic type I signal peptidase (72, 73). Alternatively, some of these proteases may function to activate chloroplast signal transduction pathways. Chaperones and chaperonins also contribute heavily to this category with 23 members present. Although the functions for many chaperones are implicated by homology (74), the exact roles that some of these proteins play in protein assembly and repair are still a mystery.

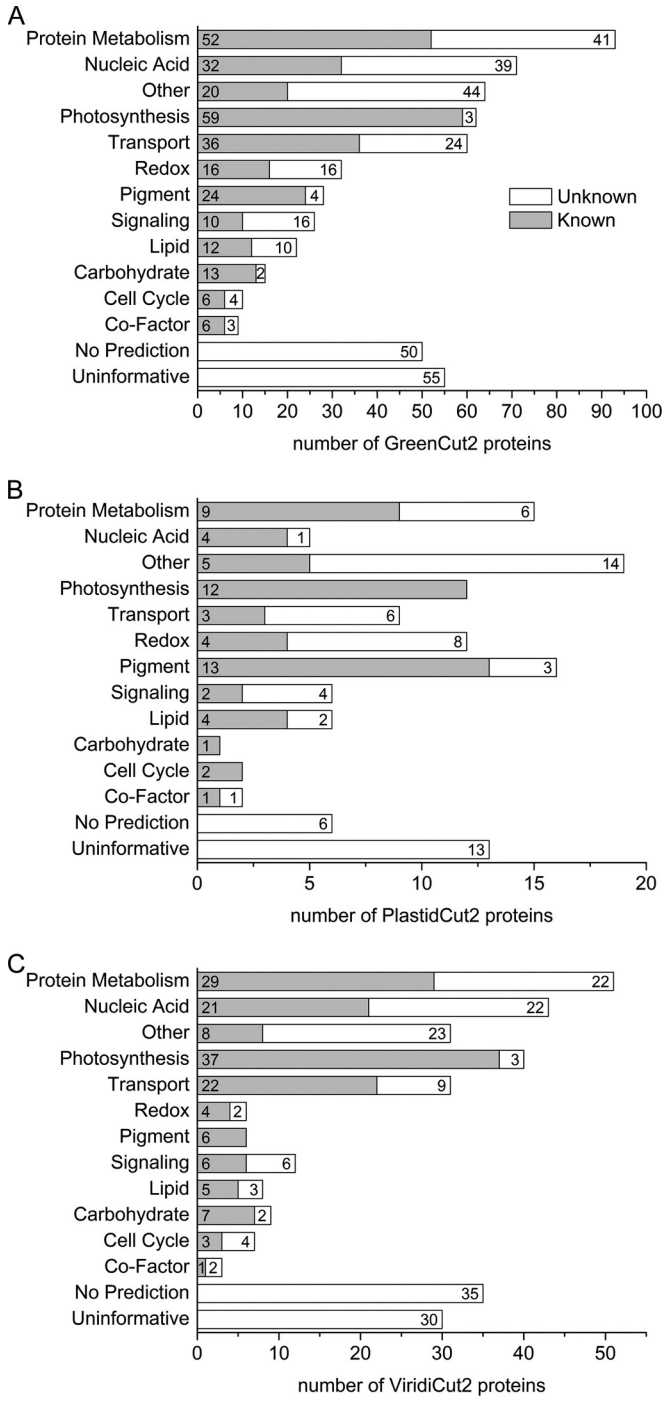

FIGURE 2. **Functional distribution of GreenCut2 proteins.** A stacked bar chart shows the numbers of proteins of known (*filled gray*) and unknown (*unfilled*) function assigned to each functional category for all GreenCut2 proteins (*A*), only the PlastidCut2 proteins (*B*), and only the ViridiCut2 proteins (*C*). Assignments to a functional category were made using the *Arabidopsis* MapMan ontology of known proteins or Pfam domain predictions for unknown proteins. The *number* of proteins in a category is shown in each *bar*. The *x axes* have been set so that the *length* of *bars* may be compared between *panels*. *Protein Metabolism*, protein maturation and degradation; *Nucleic Acid*, nucleic acid binding, modification, and transcription factors; *Other*, domain or motif to suggest a general function but not a specific functional category; *Photosynthesis*, photosynthetic apparatus and carbon fixation; *Transport*, protein and small molecule trafficking and transport; *Redox*, electron carriers and reduction/oxidation enzymes; *Pigment*, chlorophyll and carotenoid metabolism; *Signaling*, signal transduction; *Lipid*, lipid metabolism; *Carbohydrate*, starch and sugar metabolism; *Cell Cycle*, cell cycle and division; *Co-Factor*, cofactor metabolism; *No Prediction*, no informative motif or domain; *Uninformative*, domain of unknown function or structural motif that does not suggest a function.

The "nucleic acid" category contains 71 proteins that engage in nucleic acid transactions. Notably, there are 21 transcription factors, nine helicases, and 13 RNA modification enzymes (such as RNA methyltransferases). The roles for many of these plant-specific transcription factors and enzymes involved in post-transcriptional RNA maturation and modifications are largely unknown (75).

The "signal" category includes proteins involved in signal transduction in the chloroplast and cytosol. Within this category are nine protein kinases, five phosphatases, and two GTPases. Two potential conserved signaling pathways in plants in which these elements may participate involve communication between the nucleus and the plastid and the modulation of plastid physiology in response to stress conditions. Recently, the GreenCut2 signaling protein At2g48070 (CrCPLD33) was shown to mediate the chloroplast oxidative burst, which is part of the plant's immune response (76). Other types of chloroplast-nucleus signaling pathways mediated by GreenCut2 proteins might include coordination of plastid division during the cell cycle, modulation of the synthesis of thylakoid membranes (77, 78), and control of the stoichiometry of photosynthetic complexes within those membranes.

The "lipid" group includes 22 proteins. One member of this group, CrLPB1 (At3g56040), was originally identified as a protein important for the acclimation of *Chlamydomonas* to sulfur and phosphorus deprivation (79). The *Arabidopsis* ortholog of CrLPB1 was shown to be a UDP-glucose pyrophosphorylase, a chloroplast enzyme that is a component of the sulfolipid biosynthesis pathway (80). Sulfolipids are synthesized by photosynthetic organisms and are present in thylakoid membranes of plants, although they are not essential for cell growth. They can substitute for phosphatidylglycerol during phosphate deficiency (81) and serve as a sulfur reserve during sulfur deficiency (82). The enzyme(s) involved in recycling sulfate from sulfolipids is not known, but we suggest that two candidates are the GreenCut2 proteins CrCPL19 (At1g10040) and CrCPLD56 (At4g11570), which contain a putative esterase domain and a hydrolase domain, respectively. The levels of their mRNAs increase 2–3-fold in sulfur-deprived *Chlamydomonas* cells (83).

Uncategorized—Proteins in the category designated "other" include those with known functions that could not readily be placed in any of the above categories and also those with an unknown function but that contain a feature suggestive of a specific catalytic activity. An example of the former is AtAMI1 (At1g08980), a plant isoform of indole-3-acetamide amidohydrolase, which is a component of the tryptophan biosynthesis pathway (84). An example of the latter is CrCGL39 (At5g27710), a protein of unknown function with a possible hydrolase domain. The substrate hydrolyzed by CrCGL39, if any, is not known, hence its placement into the category other.

Proteins that have some conserved features but whose features do not suggest functionality were placed in the "uninformative" category. Of the 55 proteins in this category, 15 are described by a structural motif, and 29 are described by a "domain of unknown function." The remaining 11 proteins have protein interaction domains or poorly conserved catalytic sites. In contrast, proteins with no identifiable domain or indi-

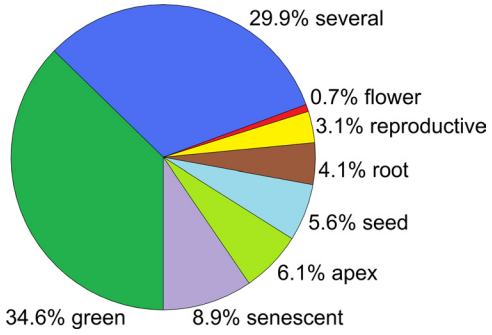

FIGURE 3. **Expression pattern of GreenCut2 genes in *Arabidopsis* organs.** Signal intensities from AtGenExpress developmental microarrays (46) were used to cluster *Arabidopsis* genes encoding GreenCut2 orthologs and co-orthologs into tissue expression categories based on high transcript abundance in one organ relative to other organs. The values do not add up to 100% because 50 of the 710 transcripts (7%) encoding the *Arabidopsis* GreenCut2 (co)-orthologs do not have associated probes on the Affymetrix ATH1 microarray chip.

cation of potential function based on information available in the literature were assigned to the "no prediction" category.

Functional Meta-analysis of RNA Abundance

Proteins in multicellular organisms are often specifically expressed in particular organ types. For example, transcripts encoding proteins involved in photosynthesis are abundant in green organs of land plants, whereas flowers often accumulate higher levels of transcripts encoding carotenoid biosynthesis enzymes. We thus inferred function of GreenCut2 proteins in the U and UP categories from their tissue/organ expression patterns and from co-expression with proteins of known function (in the K and KI categories). We used data from the AtGenExpress microarray project (46) to query the abundance of transcripts encoding the 710 *Arabidopsis* GreenCut2 orthologs and co-orthologs. The expression of a gene in a microarray experiment was defined as organ-specific if the intensity of the gene's microarray signal from a particular organ was greater than 25% of the intensity summed from all organs for that gene (see "Experimental Procedures"). Because RNA abundance is often correlated with protein abundance (85, 86), we assessed the relative abundance of RNAs encoding GreenCut2 proteins using two RNA-seq data sets. One data set was from photoautotrophically or photoheterotrophically grown *Chlamydomonas* cells (47), and the other was from a study of *Arabidopsis* shoots and roots (48). We reasoned that structural components of enzymes in photosynthesis and primary metabolism would be encoded by more abundant RNAs than proteins involved in regulation or assembly/biogenesis processes, and this is borne out by the analysis (see below).

Transcripts for 246 (35%) GreenCut2 orthologs and co-orthologs preferentially accumulate in green organs of *Arabidopsis* (Fig. 3) with ~95% of the corresponding proteins localized to plastids. As expected, transcripts for proteins in the photosynthesis category are enriched in green tissue ($p = 9e{-}11$) and are very abundant. We also observed a depletion of nucleic acid category transcripts in green tissues ($p = 4e{-}5$), suggesting that some of these transcription factors may be important for regulatory events in non-photosynthetic tissues.

Transcripts encoding 63 GreenCut2 proteins were most abundant in senescing tissue with 65% of the encoded proteins either demonstrated or predicted to be in the plastid. A number of these proteins may be involved in the turnover of plastid constituents, a process that is poorly understood but presumed to be enhanced in senescing tissues. Consistent with this possibility, a number of transcripts abundant in senescing tissue encode proteins that facilitate polypeptide degradation, alter the redox state of proteins, and modulate lipid metabolism. However, in non-senescing tissue, these proteins might participate in "housekeeping" regulation of the biosynthesis and turnover of protein complexes, pigment molecules, and lipid constituents.

Interestingly, 48 genes encoding GreenCut2 orthologs are most highly expressed in the shoot apex, and 31 are most highly expressed in the root. Both the shoot and root apical meristems contain rapidly dividing cells, which may account for the finding that 15 (19%) of the proteins encoded by these highly expressed transcripts are associated with cell proliferation, transcription, and DNA repair. It is possible that the root proteins are associated with the non-photosynthetic plastid and/or engage in processes that are independent of photosynthesis.

Distribution patterns for transcripts encoding K/KI and U/UP category GreenCut2 proteins in *Chlamydomonas* cells and *Arabidopsis* shoots are similar (Fig. 4, *A–C*). In photosynthetic cells, transcripts for GreenCut2 orthologs and co-orthologs of known function generally have higher abundances than those encoding proteins of unknown function. Many proteins engaged in high flux reactions accumulate to high levels in the cell, and because protein abundance often correlates with the amount of the corresponding transcript, transcripts encoding these proteins will be abundant relative to the "average" transcript. For example, transcripts encoding proteins of the photosynthetic apparatus, such as CrPSAD (At1g03130/At4g02770) and CrLHCA1 (At3g54890), are among the most abundant in *Chlamydomonas* (3200 and 4000 RPKM, respectively, in cells grown photoheterotrophically) (47). In contrast, regulatory proteins or assembly factors, such as AtMBB1 (At3g17040), an mRNA maturation factor for *psbB* (87), and CrCCS1 (At1g49380), which catalyzes the covalent attachment of heme to *c*-type cytochromes (88), are generally present in lower amounts, which is reflected by low mRNA abundances (~16 and ~15 RPKM, respectively, for *Chlamydomonas*). The fact that many mRNAs and proteins have been identified through molecular screens that more readily recover abundant targets explains why many characterized proteins in the K and KI categories are encoded by high abundance transcripts.

Interestingly, transcripts encoding *Chlamydomonas* GreenCut2 proteins and the *Arabidopsis* orthologs expressed in shoots have mean abundances (>100 and 860 RPKM, respectively) that are larger than the respective mean transcript abundances calculated for all nucleus-encoded transcripts from these organisms (~50 and 99 RPKM, respectively; $p = 1e{-}5$). A similar result was obtained for median abundances. These results may reflect a higher rate of transcription and/or increased half-lives of transcripts encoding GreenCut2 proteins compared with the average gene and transcript.

Of the 246 *Arabidopsis* transcripts encoding GreenCut2 orthologs and co-orthologs that accumulated in green organs to higher levels than in other tissues (Fig. 3), RNA-seq data demonstrated that 186 were more abundant in shoots than in

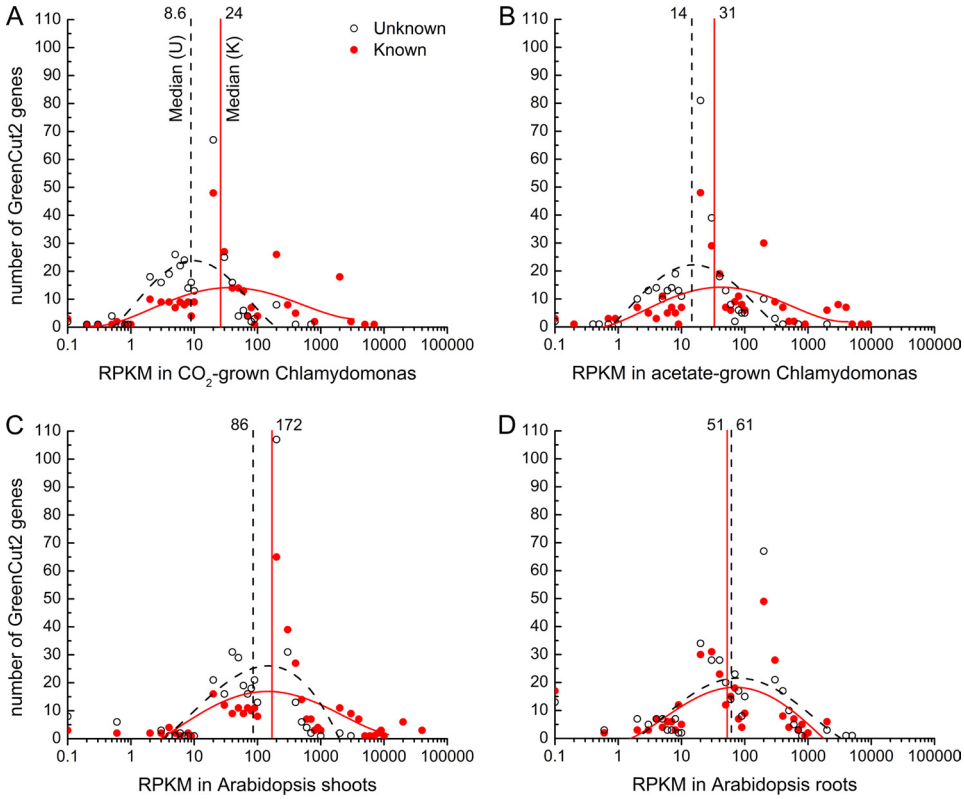

FIGURE 4. **GreenCut2 transcript abundance distribution in *Chlamydomonas* cells and *Arabidopsis* organs.** *A* and *B*, distribution of mRNA abundances from *Chlamydomonas* strain CC-1021 grown in Tris phosphate medium with CO_2 as a carbon source (*A*) or Tris acetate phosphate medium with acetate as a carbon source (*B*) (47). Transcripts from 597 genes encoding GreenCut2 proteins were binned by abundance, which is presented in RPKM values. *Closed red circles* represent encoded proteins of known function. *Open black circles* represent encoded proteins of unknown function. The medians of the known (*solid vertical red line*) and unknown (*dashed vertical black line*) transcripts are displayed with the corresponding median value. A polynomial best fit line to the distribution of transcript abundances is presented for known transcripts (*solid red*) and unknown transcripts (*dashed black*). *C* and *D*, distribution of mRNA abundances from *Arabidopsis* shoots (*C*) or roots (*D*) (48). Transcripts from 710 genes encoding GreenCut2 orthologs and co-orthologs were grouped into bins based on abundance.

roots by at least 3-fold with 48 of these 186 transcripts over 100-fold more abundant in shoot than in roots. Additionally, 60 mRNAs that were not green tissue-specific based on microarray information displayed at least 3-fold higher abundance in shoots than in roots when analyzed by RNA-seq. Finally, the abundance of transcripts for both K/KI and U/UP proteins is generally reduced in roots compared with shoots ($p = 4e{-}7$; Fig. 4*D*) likely because transcripts encoding proteins that function in photosynthesis are less abundant in roots. Only 12 transcripts are greater than 3-fold more abundant in roots than in shoots. One transcript (CGLD27; At5g67370) that is weakly expressed in roots, based on RNA-seq data, was demonstrated to be responsive to iron deficiency in *Arabidopsis* and *O. sativa* root tissue (89, 90), which suggests that some GreenCut2 proteins may perform functions important only under particular conditions. These results suggest that mutants of some GreenCut2 proteins may demonstrate an organ- or condition-dependent phenotype, which should be considered when investigating GreenCut2 proteins experimentally.

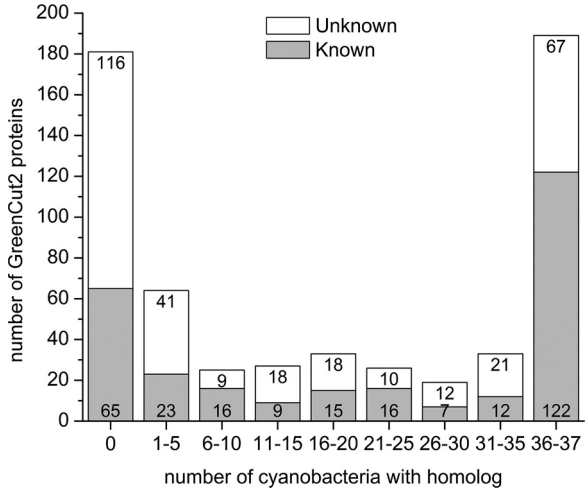

FIGURE 5. **Conservation of GreenCut2 proteins in cyanobacteria.** The amino acid sequences of the *Arabidopsis* GreenCut2 orthologs were used as queries in BLASTP searches against 37 cyanobacterial genomes. Best hit results with E-values $<1e{-}4$ were considered to be homologs. Proteins with known function are shown as *gray columns*, whereas proteins of unknown function are shown as *stacked white columns*. The *number* of proteins in each bin is shown.

Functional Meta-analysis of Prokaryote Versus Eukaryote

Many GreenCut2 proteins are localized to plastids and may have originated in the cyanobacterial endosymbiont that evolved into a plastid. Therefore, free-living cyanobacteria are likely to have homologs to many GreenCut2 proteins. To identify GreenCut2 proteins related to cyanobacterial proteins, the predicted proteomes of 37 fully sequenced cyanobacterial genomes were queried with *Arabidopsis* GreenCut2 proteins by BLASTP. The results of these comparisons reveal several interesting features that may relate to the evolution of GreenCut2 proteins. There is a bimodal distribution pattern with respect to the occurrence of

GreenCut2 homologs among the cyanobacteria (Fig. 5). Most proteins were either conserved in all genomes (189 of 597), suggesting a fundamental metabolic function, or in no genomes (181 of 597), suggesting a function related to eukaryotic-specific processes. For instance, significantly more GreenCut2 proteins assigned to the pigment and protein metabolism functional categories have homologs encoded by all or nearly all cyanobacterial genomes ($p < 7e{-}3$; supplemental File 1, Fig. S4). Although 64% of GreenCut2 proteins conserved in all or nearly all cyanobacteria have a known function, only 34% of GreenCut2 proteins without a cyanobacterial homolog are characterized. Furthermore, relative to other GreenCut2 subgroups, the ViridiCut2 is depleted for proteins with homologs in all cyanobacteria (82 of 312; $p = 4e{-}6$) and instead is enriched for proteins that do not have any cyanobacterial homologs (126 of 312; $p = 2e{-}12$). Finally, those GreenCut2 proteins placed in the no prediction category were either not present in any or associated with just a small subset of the cyanobacteria ($p = 2e{-}7$).

Proteins in the photosynthesis category were not enriched in any of the cyanobacterial genome bins. Only 34% (21 of 62) of GreenCut2 proteins involved in photosynthetic processes have homologs encoded by 36 or 37 of the cyanobacterial genomes, whereas 27% (17 of 62) do not have homologs in any of the cyanobacteria (supplemental File 1, Fig. S4). One protein in the latter category, Rubisco methyltransferase, modifies an N-terminal lysine residue of the Rubisco large subunit (33). The functional significance of this methylation event is not understood, although the similarity of the Rubisco methyltransferase SET domain to that of histone methyltransferases suggests that the protein has a eukaryotic origin.

Together, our results have a number of functional and evolutionary implications. Proteins that are well conserved in photosynthetic eukaryotes and in cyanobacteria are more likely to have been studied already in a photosynthetic reference organism (cyanobacteria, plants, and algae) and to have been attributed a function as exemplified by the chlorophyll biosynthetic pathway or proteins involved in redox metabolism. These proteins are well represented in the PlastidCut2. We suggest that these functions are defining characteristics of the majority of photosynthetic organisms. Conversely, proteins that are present in only some cyanobacteria are less well studied and may be associated with eukaryote-specific features of the plastid, such as protein import and nuclear signaling. Furthermore, the set of proteins without cyanobacterial homologs is depleted for expression in green *Arabidopsis* tissues ($p = 4e{-}7$), which suggests that these eukaryote-specific proteins are involved in processes that are not exclusively associated with photosynthetic function. From the above results, we suggest that analysis of plant-specific ViridiCut2 proteins is likely to illuminate nucleus-directed regulatory processes associated with plastid biochemistry and metabolism as well as with other green plant lineage-specific processes that are not associated with photosynthetic function.

Acknowledgments—We thank David Casero, Madeli Castruita, Maria Bernal, and Ute Krämer for preliminary access to RNA-seq data; Matteo Pellegrini and Huiying Li for bioinformatics advice; and M. Dudley Page and Davin Malasarn for suggestions on the manuscript.

REFERENCES

1. Knoll, A. H. (1992) *Science* **256,** 622–627
2. Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G., and Bhattacharya, D. (2004) *Mol. Biol. Evol.* **21,** 809–818
3. Gross, J., and Bhattacharya, D. (2009) *Nat. Rev. Genet.* **10,** 495–505
4. Jarvis, P. (2008) *New Phytol.* **179,** 257–285
5. Li, H. M., and Chiu, C. C. (2010) *Annu. Rev. Plant Biol.* **61,** 157–180
6. Abdallah, F., Salamini, F., and Leister, D. (2000) *Trends Plant Sci.* **5,** 141–142
7. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016
8. Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004) *Proteomics* **4,** 1581–1590
9. van Wijk, K. J. (2004) *Plant Physiol. Biochem.* **42,** 963–977
10. Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) *Science* **278,** 631–637
11. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., and Blankenship, R. E. (2002) *Science* **298,** 1616–1620
12. Comas, I., Moya, A., and González-Candelas, F. (2007) *BMC Evol. Biol.* **7,** Suppl. 1, S7
13. Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R., and Galperin, M. Y. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 13126–13131
14. Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., and Chisholm, S. W. (2007) *PLoS Genet.* **3,** e231
15. Gupta, R. S., and Mathews, D. W. (2010) *BMC Evol. Biol.* **10,** 24
16. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J. M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J. M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G. F., Straub, M. L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J. L. (2004) *Nature* **430,** 35–44
17. Souciet, J. L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., Sherman, D. J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Ségurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M. E., Fulton, R. S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M. L., Casarégola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., de Montigny, J., Neuvéglise, C., Thierry, A., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goëffon, A., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L., Ozanne, C., Richard, G. F., Sacerdot, C., Straub, M. L., and Talla, E. (2009) *Genome Res.* **19,** 1696–1709
18. Babenko, V. N., and Krylov, D. M. (2004) *Nucleic Acids Res.* **32,** 5029–5035
19. Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I., and Koonin, E. V. (2007) *Biol. Direct* **2,** 33
20. Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002) *Proc. Natl. Acad. Sci. U.S.A.* **99,** 12246–12251
21. Conte, M. G., Gaillard, S., Lanau, N., Rouard, M., and Périn, C. (2008) *Nucleic Acids Res.* **36,** D991–D998
22. 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., Shapiro, H., Lucas, S. M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chan-

freau, G., Chen, C. L., Cognat, V., Croft, M. T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P. A., Lemaire, S. D., Lobanov, A. V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J. V., Moseley, J., Napoli, C., Nedelcu, A. M., Niyogi, K., Novoselov, S. V., Paulsen, I. T., Pazour, G., Purton, S., Ral, J. P., Riaño-Pachón, D. M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S. L., Allmer, J., Balk, J., Bisova, K., Chen, C. J., Elias, M., Gendler, K., Hauser, C., Lamb, M. R., Ledford, H., Long, J. C., Minagawa, J., Page, M. D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A. M., Yang, P., Ball, S., Bowler, C., Dieckmann, C. L., Gladyshev, V. N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R. T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y. W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W. C., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I. V., Rokhsar, D. S., and Grossman, A. R. (2007) *Science* **318,** 245–250

23. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008) *Nucleic Acids Res.* **36,** D1009–D1014
24. Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P. F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W. B., Barker, E., Bennetzen, J. L., Blankenship, R., Cho, S. H., Dutcher, S. K., Estelle, M., Fawcett, J. A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K. A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D. R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P. J., Sanderfoot, A., Schween, G., Shiu, S. H., Stueber, K., Theodoulou, F. L., Tu, H., Van de Peer, Y., Verrier, P. J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A. C., Hasebe, M., Lucas, S., Mishler, B. D., Reski, R., Grigoriev, I. V., Quatrano, R. S., and Boore, J. L. (2008) *Science* **319,** 64–69
25. Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A. Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piégu, B., Ball, S. G., Ral, J. P., Bouget, F. Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., and Moreau, H. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 11647–11652
26. Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otillar, R., Merchant, S. S., Podell, S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., Tai, V., Vallon, O., Piganeau, G., Jancek, S., Heijde, M., Jabbari, K., Bowler, C., Lohr, M., Robbens, S., Werner, G., Dubchak, I., Pazour, G. J., Ren, Q., Paulsen, I., Delwiche, C., Schmutz, J., Rokhsar, D., Van de Peer, Y., Moreau, H., and Grigoriev, I. V. (2007) *Proc. Natl. Acad. Sci. U.S.A.* **104,** 7705–7710
27. Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006) *Science* **313,** 1596–1604
28. Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X. C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. A. (2010) *Nature* **463,** 178–183
29. Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L. K., Hellsten, U., Chapman, J., Simakov, O., Rensing, S. A., Terry, A., Pangilinan, J., Kapitonov, V., Jurka, J., Salamov, A., Shapiro, H., Schmutz, J., Grimwood, J., Lindquist, E., Lucas, S., Grigoriev, I. V., Schmitt, R., Kirk, D., and Rokhsar, D. S. (2010) *Science* **329,** 223–226
30. Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., Salamov, A., Terry, A., Yamada, T., Dunigan, D. D., Grigoriev, I. V., Claverie, J. M., and Van Etten, J. L. (2010) *Plant Cell* **22,** 2943–2955
31. Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002) *Science* **296,** 92–100
32. Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C. R. (2007) *Nucleic Acids Res.* **35,** D883–D887
33. Houtz, R. L., Royer, M., and Salvucci, M. E. (1991) *Plant Physiol.* **97,** 913–920
34. Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) *J. Mol. Biol.* **314,** 1041–1052
35. Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D., and van Wijk, K. J. (2009) *Nucleic Acids Res.* **37,** D969–D974
36. Kleffmann, T., Hirsch-Hoffmann, M., Gruissem, W., and Baginsky, S. (2006) *Plant Cell Physiol.* **47,** 432–436
37. Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., and Millar, A. H. (2007) *Nucleic Acids Res.* **35,** D213–D218
38. Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., and Rolland, N. (2010) *Mol. Cell. Proteomics* **9,** 1063–1084
39. Atteia, A., Adrait, A., Brugière, S., Tardif, M., van Lis, R., Deusch, O., Dagan, T., Kuhn, L., Gontero, B., Martin, W., Garin, J., Joyard, J., and Rolland, N. (2009) *Mol. Biol. Evol.* **26,** 1533–1548
40. Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., and Nakai, K. (2007) *Nucleic Acids Res.* **35,** W585–W587
41. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) *Nucleic Acids Res.* **38,** D211–D222
42. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005) *Nucleic Acids Res.* **33,** 5691–5702
43. Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M. (2004) *Plant J.* **37,** 914–939
44. Murata, N., and Suzuki, I. (2006) *J. Exp. Bot.* **57,** 235–247
45. Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y., and Tabata, S.

(2007) *DNA Res.* **14,** 207–216
46. Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J. U. (2005) *Nat. Genet.* **37,** 501–506
47. Castruita, M., Casero, D., Karpowicz, S. J., Kropat, J., Vieler, A., Hsieh, S. I., Cokus, S., Loo, J. A., Benning, C., Pellegrini, M., and Merchant, S. S. (2011) *Plant Cell*, 10.1105/tcp.111.084400
48. Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P. Y., Stroud, H., Yu, Y., Hetzel, J. A., Kuo, F., Kim, J., Cokus, S. J., Casero, D., Bernal, M., Huijser, P., Clark, A. T., Krämer, U., Merchant, S. S., Zhang, X., Jacobsen, S. E., and Pellegrini, M. (2010) *Nature* **466,** 388–392
49. Nakao, M., Okamoto, S., Kohara, M., Fujishiro, T., Fujisawa, T., Sato, S., Tabata, S., Kaneko, T., and Nakamura, Y. (2010) *Nucleic Acids Res.* **38,** D379–D381
50. Jancek, S., Gourbière, S., Moreau, H., and Piganeau, G. (2008) *Mol. Biol. Evol.* **25,** 2293–2300
51. Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S. Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., Yoshida, Y., Nishimura, Y., Nakao, S., Kobayashi, T., Momoyama, Y., Higashiyama, T., Minoda, A., Sano, M., Nomoto, H., Oishi, K., Hayashi, H., Ohta, F., Nishizaka, S., Haga, S., Miura, S., Morishita, T., Kabeya, Y., Terasawa, K., Suzuki, Y., Ishii, Y., Asakawa, S., Takano, H., Ohta, N., Kuroiwa, H., Tanaka, K., Shimizu, N., Sugano, S., Sato, N., Nozaki, H., Ogasawara, N., Kohara, Y., and Kuroiwa, T. (2004) *Nature* **428,** 653–657
52. Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. W., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P., and Rokhsar, D. S. (2004) *Science* **306,** 79–86
53. Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J. P., Chiovitti, A., Choi, C. J., Coesel, S., De Martino, A., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P. G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Oudot-Le Secq, M. P., Napoli, C., Obornik, M., Parker, M. S., Petit, J. L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Van de Peer, Y., and Grigoriev, I. V. (2008) *Nature* **456,** 239–244
54. Archibald, J. M., and Keeling, P. J. (2002) *Trends Genet.* **18,** 577–584
55. Gould, S. B., Waller, R. F., and McFadden, G. I. (2008) *Annu. Rev. Plant Biol.* **59,** 491–517
56. Worden, A. Z., Lee, J. H., Mock, T., Rouzé, P., Simmons, M. P., Aerts, A. L., Allen, A. E., Cuvelier, M. L., Derelle, E., Everett, M. V., Foulon, E., Grimwood, J., Gundlach, H., Henrissat, B., Napoli, C., McDonald, S. M., Parker, M. S., Rombauts, S., Salamov, A., Von Dassow, P., Badger, J. H., Coutinho, P. M., Demir, E., Dubchak, I., Gentemann, C., Eikrem, W., Gready, J. E., John, U., Lanier, W., Lindquist, E. A., Lucas, S., Mayer, K. F., Moreau, H., Not, F., Otillar, R., Panaud, O., Pangilinan, J., Paulsen, I., Piegu, B., Poliakov, A., Robbens, S., Schmutz, J., Toulza, E., Wyss, T., Zelensky, A., Zhou, K., Armbrust, E. V., Bhattacharya, D., Goodenough, U. W., Van de Peer, Y., and Grigoriev, I. V. (2009) *Science* **324,** 268–272
57. Peers, G., and Price, N. M. (2006) *Nature* **441,** 341–344
58. Corami, F., Capodaglio, G., Turetta, C., Soggia, F., Magi, E., and Grotti, M. (2005) *J. Environ. Monit.* **7,** 1256–1264
59. Sasaki, N. V., and Sato, N. (2010) *Database* **2010,** bap025
60. Oyama, T., Shimura, Y., and Okada, K. (1997) *Genes Dev.* **11,** 2983–2995
61. Chattopadhyay, S., Ang, L. H., Puente, P., Deng, X. W., and Wei, N. (1998) *Plant Cell* **10,** 673–683
62. Oravecz, A., Baumann, A., Máté, Z., Brzezinska, A., Molinier, J., Oakeley, E. J., Adám, E., Schäfer, E., Nagy, F., and Ulm, R. (2006) *Plant Cell* **18,** 1975–1990
63. Gabaldón, T. (2010) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365,** 765–773
64. Galland, N., and Michels, P. A. (2010) *Eur. J. Cell Biol.* **89,** 621–637
65. Mano, S., Nakamori, C., Nito, K., Kondo, M., and Nishimura, M. (2006) *Plant J.* **47,** 604–618
66. Bonsegna, S., Slocombe, S. P., De Bellis, L., and Baker, A. (2005) *Arch. Biochem. Biophys.* **443,** 74–81
67. Gietl, C. (1992) *Biochim. Biophys. Acta* **1100,** 217–234
68. Amunts, A., Toporik, H., Borovikova, A., and Nelson, N. (2010) *J. Biol. Chem.* **285,** 3478–3486
69. Terauchi, A. M., Lu, S. F., Zaffagnini, M., Tappa, S., Hirasawa, M., Tripathy, J. N., Knaff, D. B., Farmer, P. J., Lemaire, S. D., Hase, T., and Merchant, S. S. (2009) *J. Biol. Chem.* **284,** 25867–25878
70. Zhu, J., Fu, X., Koo, Y. D., Zhu, J. K., Jenney, F. E., Jr., Adams, M. W., Zhu, Y., Shi, H., Yun, D. J., Hasegawa, P. M., and Bressan, R. A. (2007) *Mol. Cell. Biol.* **27,** 5214–5224
71. Adam, Z., and Clarke, A. K. (2002) *Trends Plant Sci.* **7,** 451–456
72. Inoue, K., Baldwin, A. J., Shipman, R. L., Matsui, K., Theg, S. M., and Ohme-Takagi, M. (2005) *J. Cell Biol.* **171,** 425–430
73. Shipman, R. L., and Inoue, K. (2009) *FEBS Lett.* **583,** 938–942
74. Schroda, M. (2004) *Photosynth. Res.* **82,** 221–240
75. Stern, D. B., Goldschmidt-Clermont, M., and Hanson, M. R. (2010) *Annu. Rev. Plant Biol.* **61,** 125–155
76. Belhaj, K., Lin, B., and Mauch, F. (2009) *Plant J.* **58,** 287–298
77. Sakamoto, W., Zaltsman, A., Adam, Z., and Takahashi, Y. (2003) *Plant Cell* **15,** 2843–2855
78. Park, S., and Rodermel, S. R. (2004) *Proc. Natl. Acad. Sci. U.S.A.* **101,** 12765–12770
79. Chang, C. W., Moseley, J. L., Wykoff, D., and Grossman, A. R. (2005) *Plant Physiol.* **138,** 319–329
80. Okazaki, Y., Shimojima, M., Sawada, Y., Toyooka, K., Narisawa, T., Mochida, K., Tanaka, H., Matsuda, F., Hirai, A., Hirai, M. Y., Ohta, H., and Saito, K. (2009) *Plant Cell* **21,** 892–909
81. Yu, B., and Benning, C. (2003) *Plant J.* **36,** 762–770
82. Sugimoto, K., Sato, N., and Tsuzuki, M. (2007) *FEBS Lett.* **581,** 4519–4522
83. González-Ballester, D., Casero, D., Cokus, S., Pellegrini, M., Merchant, S. S., and Grossman, A. R. (2010) *Plant Cell* **22,** 2058–2084
84. Neu, D., Lehmann, T., Elleuche, S., and Pollmann, S. (2007) *FEBS J.* **274,** 3440–3451
85. Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008) *Science* **320,** 938–941
86. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009) *Mol. Biosyst.* **5,** 1512–1526
87. Vaistij, F. E., Boudreau, E., Lemaire, S. D., Goldschmidt-Clermont, M., and Rochaix, J. D. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97,** 14813–14818
88. Inoue, K., Dreyfuss, B. W., Kindle, K. L., Stern, D. B., Merchant, S., and Sodeinde, O. A. (1997) *J. Biol. Chem.* **272,** 31747–31754
89. Dinneny, J. R., Long, T. A., Wang, J. Y., Jung, J. W., Mace, D., Pointer, S., Barron, C., Brady, S. M., Schiefelbein, J., and Benfey, P. N. (2008) *Science* **320,** 942–945
90. Zheng, L., Huang, F., Narsai, R., Wu, J., Giraud, E., He, F., Cheng, L., Wang, F., Wu, P., Whelan, J., and Shou, H. (2009) *Plant Physiol.* **151,** 262–274