PharaohFUN: PHylogenomic Analysis foR plAnt prOtein History and FUNction elucidation

Ramos-González, Marcos[1,2]; Ramos-González, Víctor[2]; Arvanitidou, Christina[1,2]; Hernández-García, Jorge[3]; García-González, Mercedes[1,4] and Romero-Campero, Francisco José[1,2*]

[1] Instituto de Bioquímica Vegetal y Fotosíntesis, Universidad de Sevilla – Consejo Superior de Investigaciones Científicas; Av. Américo Vespucio 49, 41092 Seville (Spain).

[2] Department of Computer Science and Artificial Intelligence, Universidad de Sevilla; Av. Reina Mercedes s/n, 41012 Seville (Spain).

[3] Laboratory of Biochemistry, Wageningen University, Stippeneng 4, 6708WE Wageningen (The Netherlands).

[4] Department of Plant Biochemistry and Molecular Biology, Universidad de Sevilla; Av. Reina Mercedes s/n, 41012 Seville (Spain).

[*] Corresponding author email: fran@us.es

Abstract

Motivation: Since DNA sequencing has turned commonplace, the development of efficient methods and tools to explore gene sequences has become indispensable. In particular, despite photosynthetic eukaryotes constituting the largest percentage of terrestrial biomass, computational functional characterization of gene sequences in these organisms still predominantly relies on comparisons with *Arabidopsis thaliana* and other angiosperms. This paper introduces PharaohFUN, a web application designed for the evolutionary and functional analysis of protein sequences in photosynthetic eukaryotes, leveraging orthology relationships between them.

Results: PharaohFUN incorporates a homogeneous representative sampling of key species in this group, bridging clades that have traditionally been studied separately, thus establishing a comprehensive evolutionary framework to draw conclusions about sequence evolution and function. For this purpose, it incorporates modules for exploring gene tree evolutionary history, domain identification, multiple sequence alignments, and functional annotation. The study of the CCA1 protein exemplifies how PharaohFUN unifies results for both land plants and chlorophyte microalgae, accurately tracing the evolutionary history of this protein.

Availability and Implementation: PharaohFUN web application is freely available at https://greennetwork.us.es/AlgaeFUN/ and its source code can be freely accessed at https://github.com/fran-romero-campero/AlgaeFUN.

Contact: fran@us.es

Introduction

With the falling cost of massive sequencing technologies, the availability of genomes of new organisms is growing exponentially, allowing for the inclusion of representatives in clades that enable the reconstruction of the evolutionary histories of lineages on a large scale (Rokas and Carroll 2005). However, the functional characterization of the gene sequences identified in these new genomes is not progressing at the same rate as their sequencing, giving rise to a bottleneck in the application of these organisms as models of their corresponding clades (Cantalapiedra et al. 2021).

One clade in particular where this phenomenon has a marked impact is Viridiplantae. In 2020, the INSDC (Arita, Karsch-Mizrachi and Cochrane 2021) contained genomes for more than 800 species of this group, but their distribution remains heavily unbalanced: while 543 correspond to angiosperms, taxa such as gymnosperms, bryophytes or most algae are poorly represented (Kress et al. 2022). The imbalance in the functional annotation of these genomes is even more pronounced. Although organisms such as *Arabidopsis thaliana* and *Oryza sativa* have highly refined annotations, other taxa do not receive as much attention (Van Bel et al. 2018). Annotation based on these species establishes an important bias, especially since most new species do not receive neither computational nor experimental reviews or manual curation (Van Bel et al. 2018; Kress et al. 2022). In addition, most functional annotation databases neglect entire clades of this group, as is the case for streptophyte algae, key models for understanding plant colonization of land (Huerta-Cepas

et al. 2019; Leebens-Mack et al. 2019; Hanschen and Starkenburg 2020). The unequal availability of resources among species translates into a clear predominance of *A. thaliana* and other angiosperms in terms of the number of published papers (Marks et al. 2023).

In recent years, it has been demonstrated how working in an evolutionary framework facilitates the resolution of classical problems in genomics. One such example is the assignment of relationships between genes, where software such as SHOOT, based on phylogenetic methods, has been found to be more accurate than those based on local pairwise alignments, such as BLAST (Emms and Kelly 2022). Another instance is the annotation of genomic regions, where the TOGA tool incorporates information from orthologous genes to delineate coding and non-coding regions, outperforming previous methods in terms of annotation completeness (Kirilenko et al. 2023).

In this work, we introduce PharaohFUN, a web application aiming to overcome functional annotation limitations in photosynthetic organisms by facilitating the inter-species transfer of functional information based on the orthology relationships between genes. To achieve this, PharaohFUN incorporates genomic information from 46 species spanning Viridiplantae, Rhodophyta, Glaucophyta, Cryptophyta and Stramenopiles, establishing an evolutionary framework among them. By integrating phylogenetic and evolutionary contexts of each gene, along with sequence and domain studies and leveraging species with refined annotation, PharaohFUN allows users to build reliable hypotheses about the function of a gene of interest in a few minutes.

Methods and data

PharaohFUN is a web-based tool written in R, using the practical framework of the Shiny package. This tool has been included in the AlgaeFUN platform (Romero-Losada et al. 2022), which integrates transcriptomic, genomic and phylogenomic analyses oriented towards microalgae in a single suite. However, PharaohFUN extends its scope including up to 18 land plants models for study.

PharaohFUN is powered by Orthofinder 2 (Emms and Kelly 2019) and CAFE5 (Mendes et al. 2020) results, utilizing a set of curated proteomes encompassing different clades crucial for assessing the evolution in photosynthetic eukaryotes. Our platform also incorporates modules for exploring of functional domains and sequence conservation (PFAM domains and Multiple Sequence Alignment), annotation (GO terms, KEGG orthology and literature-based data), as well as sequence identification.

PharaohFUN incorporates two models with different numbers of species on which to perform the analyses (**Table S1**). The Viridiplantae model includes the proteomes of 36 species representing the main clades of the green lineage, namely Chlorophyta and Streptophyta. On the other hand, the Global model, with 46 species, also incorporates representatives of the Glaucophyta and Rhodophyta groups, completing the Archaeplastida clade, as well as several Cryptophytes and Stramenopiles.

These models were generated using Orthofinder 2 on the proteomes of the aforementioned species, extracted from Phytozome (Goodstein et al. 2012), PhycoCosm (Grigoriev et al. 2021), FernBase (www.fernbase.org), MarpolBase (Kawamura et al. 2022), ORCAE (Sterck et al. 2012) and some other specific sources (Luo et al. 2019; Li et al. 2020; Wang et al. 2021; Liu et al. 2022). The proteomes were previously filtered to include

only the proteic products of the primary transcript, ensuring that the clustering remained unaltered by the appearance of multiple proteins with high similarity derived from the same gene. Therefore, we will hereafter refer to genes instead of proteins, since each protein sequence is representative of a single gene. The Multiple Sequence Alignments (MSA) function of the software was used to generate the final orthogroups and gene trees, reconciling with the species trees generated using DendroBLAST and bibliographically validated (Strassert et al. 2021) (**Fig. 1**).

The Viridiplantae model contains 803,613 genes assigned to 49,811 orthogroups, representing 87.7% of the total genome size. As for the Global model, 920,612 genes (85.8%) are organized into 57,308 orthogroups. Although the Global model includes Viridiplantae, it should not be considered as a subset of the former, since the assignment of genes to orthogroups is essentially different. This discrepancy arises from the core concept of orthogroup: genes derived from a common gene in the ancestor of the species under study. Consequently, the difference in the age of the common ancestors between the two models, along with gene duplications and losses experienced through their connecting branches, influences the distribution of genes in each orthogroup. This distinction is illustrated by the G50 statistic, which measures the number of genes in an orthogroup such that 50% of the genes are in orthogroups of that size or larger. The G50 value is 63 for the Global model and 53 for Viridiplantae, indicating a larger orthogroup size in the Global model. A deformation in the distribution of species in the orthogroups is also observed. While the Global model has 566 orthogroups containing genes from all species, this number rises to 1,163 for Viridiplantae, pointing to a greater conservation of genes in the root of the tree for this model. Certain species exhibit a significantly higher number of identified duplications compared to other species within their respective clades. Notable cases are those of *Mesostigma viride*, *Spirogloea muscicola* and *Triticum aestivum*, attributed to whole-genome duplication events and polyploidy (El Baidouri et al. 2017; Cheng et al. 2019; Liang et al. 2020).

Due to the variation in the number of genes per species in the orthogroups, an analysis of their contractions and expansions was performed using CAFE5 software with species trees dated using the R ape package. To ensure accurate parameter estimation for birth-death processes, a subset of the orthogroups was selected to avoid distortion caused by high variation in the number of copies between species and the occurrence of orthogroups with no representative element in their root. The selected orthogroups fulfill the following criteria: presence of more than twelve species, less than 100 copies within the same species, and a maximum variation of 30 copies between species. A 3-category discrete gamma model, with an alpha parameter estimated by maximum likelihood, was employed to characterize Among Family Rate Variation, a parameter which considers the possibility that different orthogroups belong to different evolutionary rate categories. Furthermore, we considered the possibility that different branches of the species tree were subject to different evolutionary rates. Accordingly, we applied a model with 4 different birth-death parameters in the case of Viridiplantae, and 6 in the case of the global model, distributed as follows: Stramenopiles and Cryptophytes, Rhodophytes and Glaucophytes, Chlorophytes, Streptophyte algae, Non-Spermatophyte Embryophytes and Spermatophytes.

Based on these results, PharaohFUN is structured in 4 main modules, each one rendering publication-ready plots and standard format files to ensure compatibility with other specific software (**Fig. 2**).

1) Input and Sequence Identification

The tool receives as input the ID of a gene corresponding to one of the genomes of the species included in the analysis, typically following the notation used in Phytozome and Phycocosm. To overcome potential ambiguities, the "Identify Sequences" tab allows users to input an amino acid sequence and search for the most similar sequences in the proteome of a specific organism. This task is accomplished using the DIAMOND aligner (Buchfink, Xie and Huson 2015). Moreover, this functionality facilitates the mapping of protein sequences from related species to identify highly similar sequences in the species included in the model.

2) Gene Trees and Orthogroup Size Evolution

The Gene Trees tab serves as the initial step in PharaohFUN's analyses, creating and representing gene trees corresponding to specific orthogroups. Trees corresponding to orthogroups containing more than four genes are generated using distance matrices derived from the MSA and are subsequently modified by PharaohFUN to meet user's requirements. The tool takes three main inputs to generate the tree: the query gene ID, the species to be included and the optional reconciliation with the species tree. This enables users to restrict analyses to specific clades of interest. In addition, the reconciliation feature facilitates the mapping of orthogroup history relative to species evolution to graphically identify duplication and speciation events. Alternatively, an analysis based solely on sequence similarity can be performed. The Gene Trees tab provides information about the genes included in the tree filtered according to the user's guidelines, along with their corresponding plotting. Results can be exported in various formats: TSV (IDs), FASTA (protein sequences), NEWICK (tree) and PNG (tree image).

The "Orthogroup Size Expansion/Contraction" tab complements the previous information when understanding the evolution of an orthogroup throughout the species tree is the objective. This feature uses the output of CAFE5 to explore changes in the orthogroup size mapped onto the species tree. It reconstructs the number of genes belonging to the orthogroup in the common ancestors of each clade and identifies the number of genes gained or lost along each branch. Additionally, it incorporates a statistical analysis of the orthogroups that have experienced expansions or contractions in any of their branches. In this tab, the species tree is displayed with information on the number of genes within the orthogroup in each node, along with gene significant gains or losses of genes in each branch. For orthogroups with no significant changes detected or that were absent at the root of the species tree, non-significance is reported. Images are generated using cafeplotter software (https://github.com/moshi4/CafePlotter) and can be downloaded in PNG format. The raw trees are also available for download as NEWICK files.

3) Domains and Sequence Similarities

The focus of this module is to compare different sequences originating from the same gene tree, assessing their potential functional relationships through the analyses of PFAM domains they contain and the conserved and divergent regions according to an MSA. The tool implements a connection with the PFAM server to send a query for desired protein sequences. The results of the query are presented as a structured table displaying the positions of the different domains in each sequence, along with information about each domain. Additionally, an image of the proteins representing domain locations is provided.

In the MSA tab, selected protein sequences undergo alignment using the MUSCLE algorithm (Edgar 2004). This feature is useful when studying a specific clade of the gene tree, enabling the examination of sequence conservation within that group rather than in the entire tree. The output of this section consists of the plain text alignment and two downloadable files: the MSA in FASTA format and the alignment with logos and colors indicating the chemical nature of each amino acid.

4) Functional Annotation

One of the main objectives of PharaohFUN is to facilitate the formulation of hypotheses about the function of proteins with deficient functional annotation leveraging their evolutionary relationship with proteins of known function. To achieve this, PharaohFUN implements three different analyses: GO enrichment, KEGG orthology and literature-based annotation.

In the "GO Terms" tab, a gene ontology enrichment analysis is performed on a subset of genes in the tree for Biological Processes, Molecular Functions or Cellular Components categories. This analysis functionally characterizes genes based on existing functional annotations of phylogenetically close genes. The program displays a list of existing GO terms for each gene and presents two plots. The first plot, an emapplot, visualizes relationships between different groups of terms based on connections by genes involved in different processes. The second plot, the GO terms association plot, classifies GO terms clustered according to shared parental terms, providing a summarized view.

In the "KEGG Orthology (KO)" tab, once genes are selected, they are compared with proteome annotations generated using the KAAS tool (Moriya et al. 2007) with 17 photosynthetic eukaryotic organisms as reference. The annotation stores the inferred KO terms for each protein sequence that are subsequently used as input for a KEGG pathway enrichment analysis. This analysis reveals the various metabolic or signaling pathways in which the selected genes are involved. The output is presented in a tabular format, and users can also choose to plot and visualize a specific pathway to and the steps in which the selected genes act.

Finally, the "Literature Annotation" tab provides a connection to the PlantConnectome web tool (https://connectome.plant.tools/) (Fo and Mutwil 2023). This feature allows users to retrieve extensive information from plant science literature, obtained through the application of Large Language Models to process articles related to this field. Users can associate the results from the two previous tabs with other bibliographically supported biological information. Additionally, a link to the paper from which each piece of information has been inferred are provided to expand or verify the accuracy of the results.

Results

As a case study to illustrate PharaohFUN's functionalities, CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene from *A. thaliana* was chosen as a central regulator of circadian rhythms, which is largely conserved throughout Viridiplantae. *Arabidopsis thaliana*, *Marchantia polymorpha*, *Ostreococcus tauri* and *Klebsormidium nitens* were selected as species of interest along with the Viridiplantae model. The application of the "Identify Sequences" function on the amino acid sequence of the protein extracted from Uniprot (ID P92973), resulted in the identifier AT2G46830, with 100% identity. The reconciled gene tree for these species using the found identifier showed a division into two clades (**Fig. 3A**).

The first clade contained the CCA1 and LHY genes from *A. thaliana*, along with a subclade of *A. thaliana* genes, including members from the REVEILLE family such as RVE1, RVE2 and RVE7. A gene from *O. tauri* and another one from *K. nitens* were also identified in this clade, both of which have been previously reported to encode CCA1 orthologs in these organisms (Linde et al. 2017). An *M. polymorpha* gene was also found in this clade; however, existing literature suggests the absence of a CCA1 homolog in this organism (Linde et al. 2017). Subsequent application of other functionalities implemented in PharaohFUN led us to exclude this *M. polymorpha* gene as a CCA1 ortholog, confirming previous analysis, as discussed below.

The second clade was constituted by proteins traditionally referred to as the LHY-CCA1-Like (LCL) family, which included RVE3, RVE5, RVE6 and RVE8. Within this clade, a gene from *M. polymorpha* known as MpRVE, and one from *K. nitens* were identified supporting predictions from previous analysis (Linde et al. 2017). Interestingly, two genes from *O. tauri* were also identified, although only one of them had been previously described. Therefore, our analysis unveiled a new gene in *O. tauri* that may share its function with RVEs.

To further explore the outcomes derived from the gene tree, we used the rest of functionalities in PharaohFUN. Notably, the analysis of orthogroup contraction / expansion revealed (**Fig. 3D**) the acquisition of an additional gene in *O. tauri* relative to the rest of the Mamiellophyceae species clade, lending support to the observation of two RVE genes for this microalga in the gene tree. To gather more evidence and refine previous results, a PFAM domain analysis was conducted. The proteins from the gene tree, with the exception of the *M. polymorpha* CCA1 potential ortholog, were found to contain the Myb-like DNA-binding domain (PF00249.34) or a closely related variant (PF13921.9) that overlapped with it at the same position and with the same length (45 amino acids). This domain is known to be highly prevalent in transcription factors that constitute the central circadian clock (Carré and Kim 2002). Notably, both *O. tauri* genes contained this domain, providing further confirmation that they are indeed RVE genes. In contrast, the *M. polymorpha* CCA1 potential ortholog was much shorter than the other proteins and contained a different Myb DNA-binding domain in a distinct position (**Fig. 3B**). Next, we performed a MSA using the sequences of all the proteins from the gene tree. The alignment revealed the fully conserved SHAQKYF motif in all proteins, except for the *M. polymorpha* CCA1 potential ortholog (**Fig. 3C**). This motif, found in Myb DNA binding domains, is typically associated with RVE and CCA1 orthologs (Linde et al. 2017). The lack of this motif and the presence of the variant Myb DNA-binding domain in the *M. polymorpha* CCA1 potential ortholog suggests a strong divergence that may have led to loss of function or neofuctionalization. Therefore, we disregard this gene in agreement with the current existing literature. Nonetheless, further experimentation would be required to validate these predictions.

We applied the remaining functionalities in PharaohFUN to leverage the evolutionary results obtained thus to elucidate potential functions for the uncharacterized genes in this example. A GO terms enrichment analysis on the orthogroup revealed processes related to circadian rhythm transcription, auxin biosynthesis, seed germination and flower development, catabolic processes, response to blue light, response to temperature and temperature compensation of the circadian clock (**Fig. 4A-B**). However, it is important to consider that the species whose genes are used to transfer functional information can greatly impact the accuracy of the

results. For instance, the inclusion of *A. thaliana* in the analysis led to the appearance of seed germination and flower development processes not applicable to a chlorophyte microalga such as *O. tauri*. Therefore, the study of clades in the tree is crucial for selecting genes that may carry information of interest for a particular study gene. In addition, the KO annotation confirmed the presence of homologs of CCA1 and LHY in the orthogroup, highlighting their involvement in pathways related to circadian rhythm (**Fig. 4C**). Finally, the literature annotation provided valuable biological information, including descriptions of relationships such as complex formation with LHY (which belongs to the same orthogroup), repression of TOC1 through binding, control over circadian rhythms, and more than 200 further relationships. These annotations can be used to experimentally test hypotheses regarding conservation of function among predicted orthologs (**Fig. 4D**).

This example showcases the strengths of PharaohFUN, which arise from its ability to combine multiple techniques for phylogenetic analysis, gene trees, PFAM domains and MSA to identify potential orthologs and study their evolutionary history. Based on these analyses, functional annotations can be associated with unknown proteins (Gil et al. 2023).

Discussion

Currently, there is a limited collection of tools available for quickly identifying orthologs of plant genes, and few of them include a diverse range of plant species in their databases. Notable exceptions are SHOOT (Emms and Kelly 2022), FunTree (Furnham et al. 2012) or, exclusively for plant species, PLAZA (Van Bel et al. 2022) which cater to specific objectives and capabilities differing from those of PharaohFUN, as summarized in **Table 1**.

SHOOT primarily focuses on the phylogenetic location of new genes by incorporating them into a database through MSA. While it provides links to gene information stored in Phytozome, it lacks modules for domain analysis, orthogroup evolution along species trees, and functional annotation of complete orthogroups. The limited representation of algae among its 30 species Plant database (only 4 chlorophytes and 1 streptophyte) is evident when conducting searches, as demonstrated by the example of searching for the CCA1 gene, which only returns results in terrestrial plants (**Fig. S1A**) despite its widespread presence across Viridiplantae (Linde et al. 2017).

FunTree is designed to study the evolution of enzyme superfamilies by integrating functional, structural and phylogenetic information. Unlike the previous tool, it reconstructs common ancestors of gene groups. Its focus solely on specific enzyme superfamilies and their activities restricts its applicability and genes that do not code for enzymes in the registered superfamilies are not analyzed. Moreover, reconstructions are performed at the level of the most likely ancestral enzyme activity (in terms of their EC number), not based on orthogroup status. Therefore, features to explore gene evolutionary history are not provided. In the case of the CCA1 example, neither the Uniprot ID search nor the beta sequence-based search function yielded any results.

To some extent, PLAZA pursues a similar goal to PharaohFUN's, but it is limited to studying angiosperms, with separate instances for PLAZA Dicots 5.0 and PLAZA Monocots 5.0. In contrast, PharaohFUN aims to represent all photosynthetic organism groups uniformly, including clades outside Viridiplantae or

Archaeplastida. In this sense, while pico-PLAZA (Van Bel et al. 2018) targets 11 chlorophyte genera, it lacks representation for streptophyte algae, which serves as a liaison to the 4 land plants considered in its database, and only has one non-angiosperms embryophyte species (*Physcomitrium patens*). The lack of evolutionary linking elements between groups produces separation into isolated instances. While this may lead to the creation of refined gene groups, it hinders an integrated analysis of orthogroups. Bias can be incorporated in the clustering process towards the creation of small high similar gene groups corresponding to very closely related species excluding related but distant genes from orthogroups.

In contrast, PharaohFUN employs uniform species sampling to eliminate bias towards any specific clade. This approach allows for the implementation of an analysis of the evolution of different orthogroups along the species tree, a feature not supported by PLAZA. Furthermore, in terms of annotation, PharaohFUN incorporates the new literature annotation utility provided by PlantConnectome, facilitating the exploration of biological data associated with specific terms of interest and enabling annotation and representation of KEGG pathways. While both SHOOT and pico-PLAZA lack the functionalities to directly perform functional analysis for all the genes in the tree and redirects users to other external tools PharaohFUN overcomes this limitation by providing specific analysis features for these tasks within the same platform. Moreover, PharaohFUN leverages DIAMOND to enable the rapid mapping of proteins from related species to their respective representatives in the provided dataset. This approach expands PharaohFUN's utility compared to the ID-based search implemented in PLAZA and the basic form of PharaohFUN usage, thereby addressing possible ambiguity issues with gene IDs across different annotation versions. As a comparison with PharaohFUN, the search for the gene encoding CCA1 in pico-PLAZA found an orthologous gene family (ORTHO03P024982) with only four angiosperm clade-specific genes (**Fig. S1B**), despite its presence in chlorophyte species such as *O. tauri*, reflecting the aforementioned bias due to lack of evolutionary cohesion.

In summary, PharaohFUN significantly increases and unifies the evolutionary range for studies outside the flowering plants. It addresses the growing demand for research on photosynthetic eukaryotes outside this dominant group, particularly highlighting its utility in analyzing aquatic species of biotechnological or evolutionary interest. In addition, PharaohFUN stands out among other evolutionary analysis tools due to its unique features integrating domain, sequence and functional annotation information into a single platform. Thus providing a comprehensive approach that enables researchers to conduct in-depth and holistic analysis of gene evolution and function, making PharaohFUN a valuable tool for the research community on photosynthetic organisms.

As future directions, we expect PharaohFUN to be widely embraced by users in the field of plant sciences due to its user-friendly interface and its ability to facilitate evolutionary and functional analysis across numerous species of interest. The tool will continue to evolve based on feedback and interactions with users being regularly updated to incorporate the latest genome versions and include new representative species form different clades dealing with the rapid expansion in the number of genomes available. PharaohFUN's modular design allows for seamless incorporation of new species to the models for analysis, without requiring major changes to its underlying architecture. This adaptability ensures that PharaohFUN will

remain a relevant and effective tool meeting the evolving needs of the scientific community working in photosynthetic organisms.

References

1. Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol. 2005;22(5):1337–44.
2. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38(12):5825–9.
3. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021 Jan;49(D1):D121–4.
4. Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, et al. Green plant genomes: What we know in an era of rapidly expanding opportunities. Proc Natl Acad Sci U S A. 2022;119(4):1–9.
5. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. 2018 Jan;46(D1):D1190–6.
6. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309–14.
7. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574(7780):679–85.
8. Hanschen ER, Starkenburg SR. The state of algal genome quality and diversity. Algal Res. 2020;50(May):101968.
9. Marks RA, Amézquita EJ, Percival S, Rougon-Cardoso A, Chibici-Revneanu C, Tebele SM, et al. A critical analysis of plant science literature reveals ongoing inequities. Proc Natl Acad Sci. 2023 Mar 7;120(10):e2217564120.
10. Emms DM, Kelly S. SHOOT: phylogenetic gene search and ortholog inference. Genome Biol. 2022;23(1):1–13.
11. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene annotation with orthology inference at scale. Science. 2023;380(6643):eabn3107.
12. Romero-Losada AB, Arvanitidou C, de los Reyes P, García-González M, Romero-Campero FJ. ALGAEFUN with MARACAS, microALGAE FUNctional enrichment tool for MicroAlgae RnA-seq and Chip-seq AnalysiS. BMC Bioinformatics. 2022;23(1):1–15.
13. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 Nov 14;20(1).

14. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2020;36(22–23):5516–8.
15. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40:D1178-86.
16. Grigoriev I V, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, et al. PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. 2021 Jan;49(D1):D1004–11.
17. Kawamura S, Romani F, Yagura M, Mochizuki T, Sakamoto M, Yamaoka S, et al. MarpolBase Expression: A Web-Based, Comprehensive Platform for Visualization and Analysis of Transcriptomes in the Liverwort *Marchantia polymorpha*. Plant Cell Physiol. 2022 Nov;63(11):1745–55.
18. Sterck L, Billiau K, Abeel T, Rouzé P, Van de Peer Y. ORCAE: online resource for community annotation of eukaryotes. Nat Methods. 2012;9(11):1041.
19. Liu Y, Wang S, Li L, Yang T, Dong S, Wei T, et al. The *Cycas* genome and the early evolution of seed plants. Nat Plants. 2022;8(4):389–401.
20. Wang L, Zhu T, Rodriguez JC, Deal KR, Dubcovsky J, McGuire PE, et al. *Aegilops tauschii* genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. G3 Genes|Genomes|Genetics. 2021 Dec 1;11(12):jkab325.
21. Luo Q, Bian C, Tao M, Huang Y, Zheng Y, Lv Y, et al. Genome and Transcriptome Sequencing of the Astaxanthin-Producing Green Microalga, *Haematococcus pluvialis*. Genome Biol Evol. 2019 Jan;11(1):166–73.
22. Li F-W, Nishiyama T, Waller M, Frangedakis E, Keller J, Li Z, et al. *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. Nat Plants. 2020;6(3):259–72.
23. Strassert JFH, Irisarri I, Williams TA, Burki F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat Commun. 2021;12(1):1879.
24. Liang Z, Geng Y, Ji C, Du H, Wong CE, Zhang Q, et al. *Mesostigma viride* Genome and Transcriptome Provide Insights into the Origin and Evolution of Streptophyta. Adv Sci. 2020 Jan 1;7(1):1901850.
25. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. Cell. 2019;179(5):1057-1067.e14.
26. El Baidouri M, Murat F, Veyssiere M, Molinier M, Flores R, Burlot L, et al. Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). New Phytol. 2017 Feb 1;213(3):1477–86.
27. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
29. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007 Jul;35:W182-5.
30. Fo K, Mutwil M. PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts. bioRxiv. 2023 Jan 1;2023.07.11.548541.

31. Linde AM, Eklund DM, Kubota A, Pederson ERA, Holm K, Gyllenstrand N, et al. Early evolution of the land plant circadian clock. New Phytol. 2017;216(2):576–90.

32. Carré IA, Kim J. MYB transcription factors in the *Arabidopsis* circadian clock. J Exp Bot. 2002 Jul 1;53(374):1551–7.

33. Gil S, Feord HK, van Ooijen G. Homologs of ancestral CNNM proteins affect magnesium homeostasis and circadian rhythmicity in a model eukaryotic cell. Int J Mol Sci. 2023;24(3):2273.

34. Furnham N, Sillitoe I, Holliday GL, Cuff AL, Laskowski RA, Orengo CA, et al. Exploring the Evolution of Novel Enzyme Functions within Structurally Defined Protein Superfamilies. PLOS Comput Biol. 2012 Mar 1;8(3):e1002403.

35. Van Bel M, Silvestri F, Weitz EM, Kreft L, Botzki A, Coppens F, et al. PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. Nucleic Acids Res. 2022 Jan 7;50(D1):D1468–74.

Figures

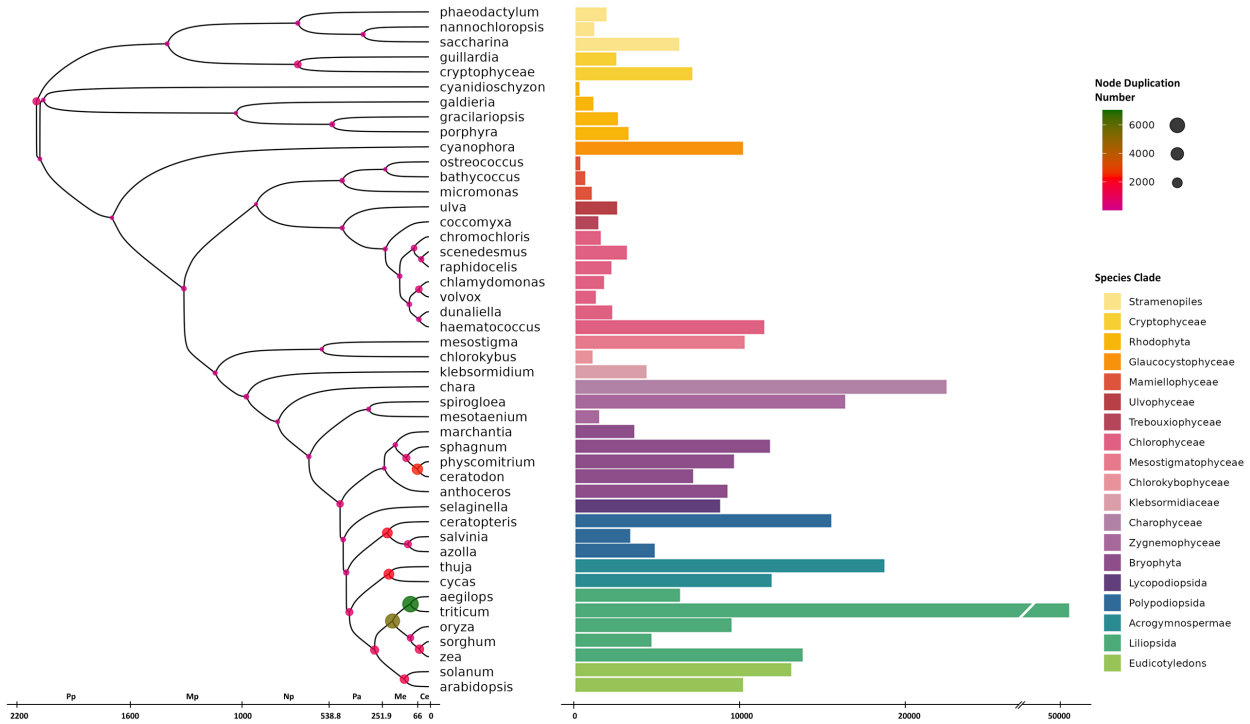


Figure 1. Dated species tree used for model construction. **Left**: dated species tree measured in millions of years. Eras terminology: **Pp** Paleoproterozoic, **Mp** Mesoproterozoic, **Np** Neoproterozoic, **Pa** Paleozoic, **Me** Mesozoic and **Ce** Cenozoic. Dating times were extracted from (Strassert et al. 2021). The inferred gene duplication number of internal nodes is mapped by circle color and size. **Right**: gene duplication number of extant species. Clades are distinguished by different bar color according to legend.

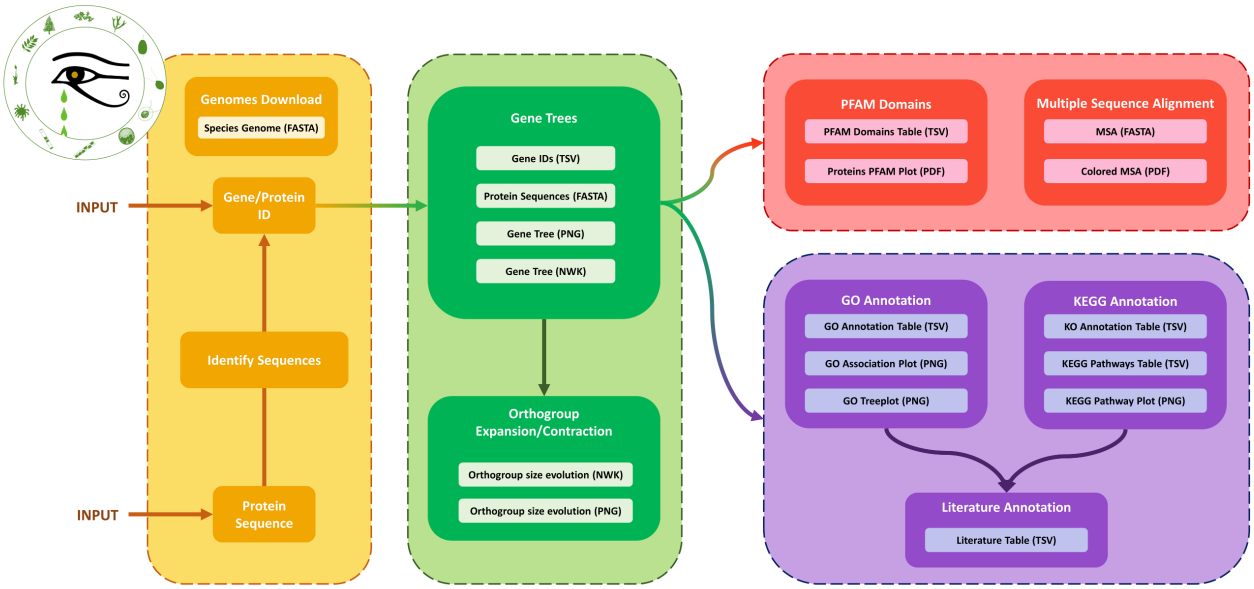


Figure 2. General architecture of PharaohFUN. Dashed boxes represent modules, continuous ones represent input types or analysis functionalities. Boxes in each functionality represent supported output. Arrows show the standard workflow through the different modules. Yellow: Input and Sequence Identification module. Green: Gene Trees and Orthogroup Size Evolution module. Red: Domains and Sequence Similarities module. Purple: Functional Annotation module.

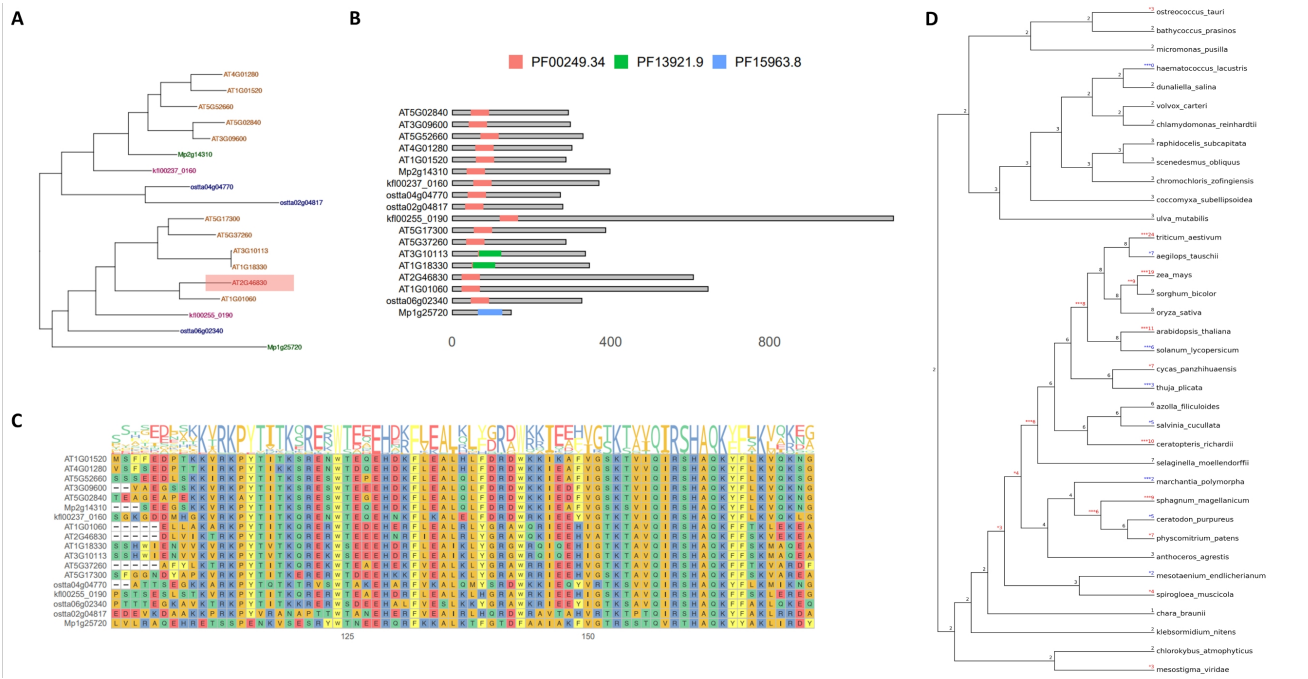
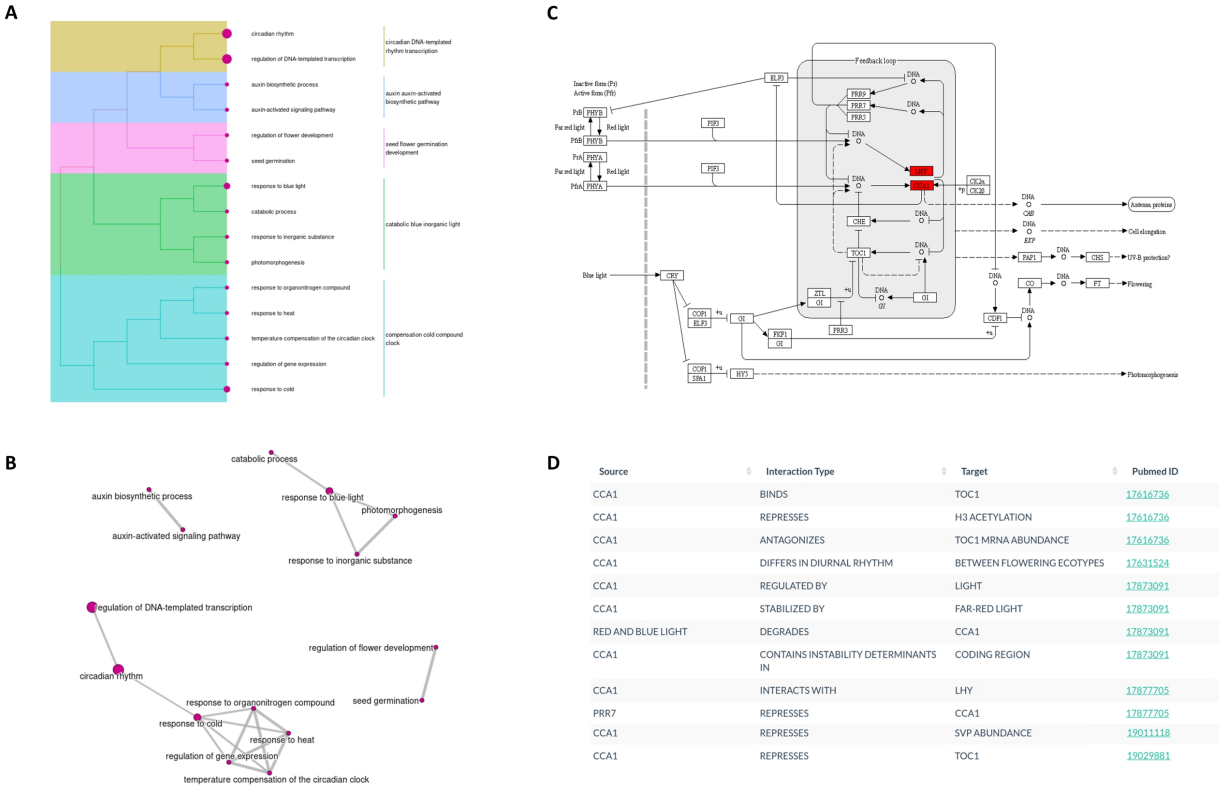


Figure 3. Overview of results for CCA1 evolution using PharaohFUN. **A** Gene Tree for CCA1 orthogroup in *A. thaliana* (yellow), *M. polymorpha* (green), *K. nitens* (pink) and *O. tauri* (blue) species. Query gene is highlighted in red. **B** PFAM Domains location in proteins corresponding to this orthogroup. **C** MSA for the orthogroup proteic sequences. **D** Size evolution of orthogroup through species tree, showing significant gene gains in red and significant gene losses in blue. P-value: * < 0.05, ** < 0.01, *** < 0.001.



Source	Interaction Type	Target	Pubmed ID
CCA1	BINDS	TOC1	17616736
CCA1	REPRESSES	H3 ACETYLATION	17616736
CCA1	ANTAGONIZES	TOC1 MRNA ABUNDANCE	17616736
CCA1	DIFFERS IN DIURNAL RHYTHM	BETWEEN FLOWERING ECOTYPES	17631524
CCA1	REGULATED BY	LIGHT	17873091
CCA1	STABILIZED BY	FAR-RED LIGHT	17873091
RED AND BLUE LIGHT	DEGRADES	CCA1	17873091
CCA1	CONTAINS INSTABILITY DETERMINANTS IN	CODING REGION	17873091
CCA1	INTERACTS WITH	LHY	17877705
PRR7	REPRESSES	CCA1	17877705
CCA1	REPRESSES	SVP ABUNDANCE	19011118
CCA1	REPRESSES	TOC1	19029881

Figure 4. Overview of results for CCA1 orthogroup functional annotation using PharaohFUN. **A** Treeplot showing clustered GO terms for biological processes. **B** Association plot showing biological processes performed by the same genes. **C** KEGG pathway for Circadian Rhythms in Plants, highlighting genes in the orthogroup (CCA1 and LHY) in red. **D** Example of some rows of literature annotation result for CCA1.

Tables

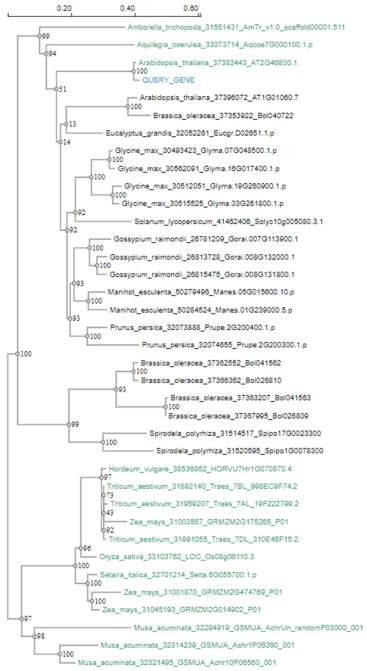
Table 1. Comparison between phylogenetic web-based tools. Number of species is measured in number of different genera per clade. * FunTree uses taxonomy data to guide reconciliation of trees, but species tree is not explicitly shown.

	PharaohFUN	SHOOT	FunTree	PLAZA 5.0 (dicots + monocots)	pico-PLAZA 3.0
Angiosperms	7	17	*	~ 80 + 30	3
Non-angiosperms land plants	11	8	*	5	1
Streptophyte algae	6	1	*	1	0
Chlorophyte algae	12	4	*	3	11
Non-viridiplantae archaeplastida	6	3 (only Rhodophyta)	*	0	3 (only Rhodophyta)
Stramenopiles algae	3	0	*	0	7
Cryptophyte algae	2	0	*	0	0
Gene trees	Yes	Yes	No	Yes	Yes

Integrated domains information	Yes	Only as individual links to gene databases	Yes	Only for query gene, and as individual links for the rest	Only for query gene, and as individual links for the rest
Ancestral reconstruction	Yes (for number of genes in orthogroup)	No	Yes (for EC function)	No	No
Direct functional annotation for all genes	Yes	No	Complete record but assignment is left to manual exploration	Complete record but assignment is left to manual exploration	Complete record but assignment is left to manual exploration
Include non-enzyme genes	Yes	Yes	No	Yes	Yes
Sequence-based search	Yes (via ID conversion)	Yes (via MSA)	Beta	No	No
Direct functional annotation	GO, KEGG (including EC)	No	GO, EC, MACiE	GO, Mapman (including EC)	GO, Mapman (including EC)
Literature annotation	Yes	No	No	No	No
Support evolutionary studies	Yes	Yes	Only for enzymes	Yes	Yes
Support functional studies	Yes	No	Yes	Yes	Yes

Supplementary data

A



B

Page 1 of 1, showing 4 records out of 4 total, starting on record 1, ending on 4

gene_id	species	strand	chr	outlier
AT1G01060	ath	-	1	No
AT2G46830	ath	+	2	No
ATR_00001G05020	atr	+	scaffold00001	No
OS08G06110	osa	+	8	No

Supplementary Figure S1. Results for CCA1 rendered by **A** SHOOT and **B** pico-PLAZA 3.0.

Supplementary Table 1. Genomic data used in PharaohFUN, including species names, genome version, species clade and genome source. Sources of genomes not downloaded from major databases are indicated as their corresponding reference.

Organism	GENOME VERSION	CLADE	GENOME SOURCE
Aegilops tauschii	Aet_v4.0	Streptophyta (Liliopsida)	[20]
Anthoceros agrestis	Bonn	Streptophyta (Bryophyta)	[22]
Arabidopsis thaliana	Araport11	Streptophyta (Eudicotyledons)	Phytozome
Azolla filiculoides	V1.1	Streptophyta (Polypodiopsida)	FernBase

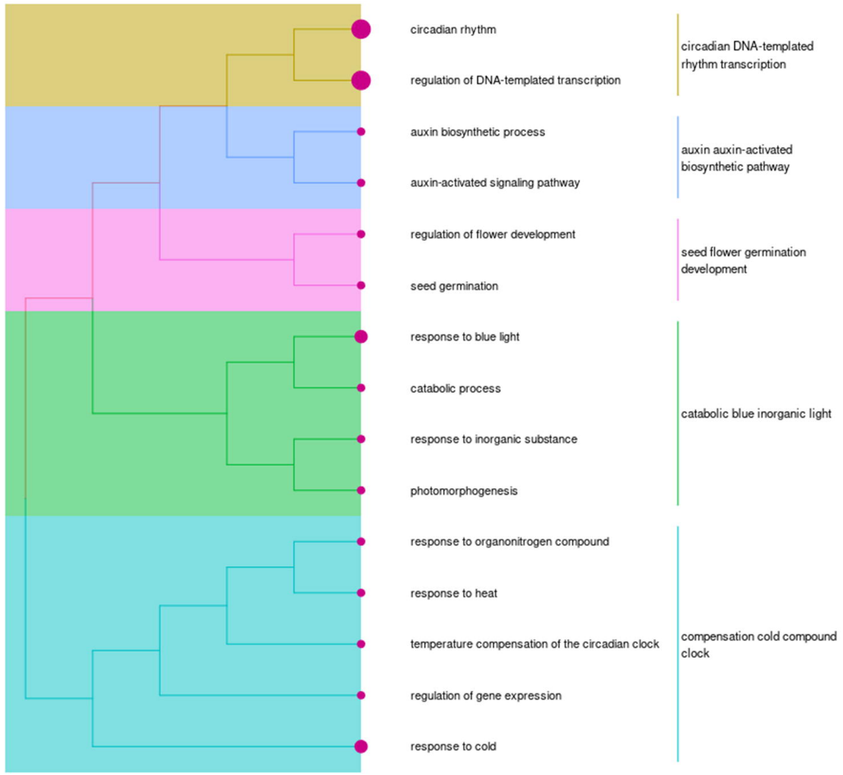
Bathycoccus prasinos	RCC1105	Chlorophyta (Mamiellophyceae)	ORCAE
Ceratodon purpureus	R40	Streptophyta (Bryophyta)	Phytozome
Ceratopteris richardii	V2.1	Streptophyta (Polypodiopsida)	Phytozome
Chara braunii	S276	Streptophyta (Charophyceae)	Phycocosm
Chlamydomonas reinhardtii	v.5.5	Chlorophyta (Chlorophyceae)	Phytozome
Chlorokybus atmophyticus	CCAC 0220	Streptophyta (Chlorokybophyceae)	Phycocosm
Chromochloris zofingiensis	v.5.2.3.2	Chlorophyta (Chlorophyceae)	Phytozome
Coccomyxa subellipsoidea	C-169, v3	Chlorophyta (Trebouxiophyceae)	Phycocosm
Cryptophyceae sp.	CCMP2293 v1.0	Cryptophyta	Phycocosm
Cyanidioschyzon merolae	10D	Rhodophyta	Phycocosm
Cyanophora paradoxa	CCMP329	Glaucophyta	Phycocosm
Cycas panzhihuaensis	V1.0	Streptophyta (Acrogymnospermae)	[19]
Dunaliella salina	v.1.0	Chlorophyta (Chlorophyceae)	Phytozome
Galdieria sulphuraria	SAG 21.92	Rhodophyta	Phycocosm

Gracilariopsis chorda	SKKU-2015	Rhodophyta	Phycocosm
Guillardia theta	CCMP2712 v1.0	Cryptophyta	Phycocosm
Haematococcus lacustris	GCA_011766145.1_Lacustris_1.0	Chlorophyta (Chlorophyceae)	[21]
Klebsormidium nitens	V1.1	Streptophyta (Klebsormidiaceae)	Phycocosm
Marchantia polymorpha	V5.1	Streptophyta (Bryophyta)	MarpolBase
Mesostigma viride	NIES-296	Streptophyta (Mesostigmatophyceae)	Phycocosm
Mesotaenium endlicherianum	SAG 12.97	Streptophyta (Zygnemophyceae)	Phycocosm
Micromonas pusilla	CCMP1545 v3.0	Chlorophyta (Mamiellophyceae)	Phytozome
Nannochloropsis gaditana	B-31	Stramenopiles	Phycocosm
Oryza sativa	v.7.0	Streptophyta (Liliopsida)	Phytozome
Ostreococcus tauri	RCC 4221 v3.0	Chlorophyta (Mamiellophyceae)	Phycocosm
Phaeodactylum tricornutum	CCAP 1055/1 v2.0	Stramenopiles	Phycocosm
Physcomitrium patens	V3.3	Streptophyta (Bryophyta)	Phytozome
Porphyra umbilicalis	V1.5	Rhodophyta	Phytozome
Raphidocelis subcapitata	NIES-35	Chlorophyta (Chlorophyceae)	Phycocosm
Saccharina japonica	str.Ja	Stramenopiles	Phycocosm

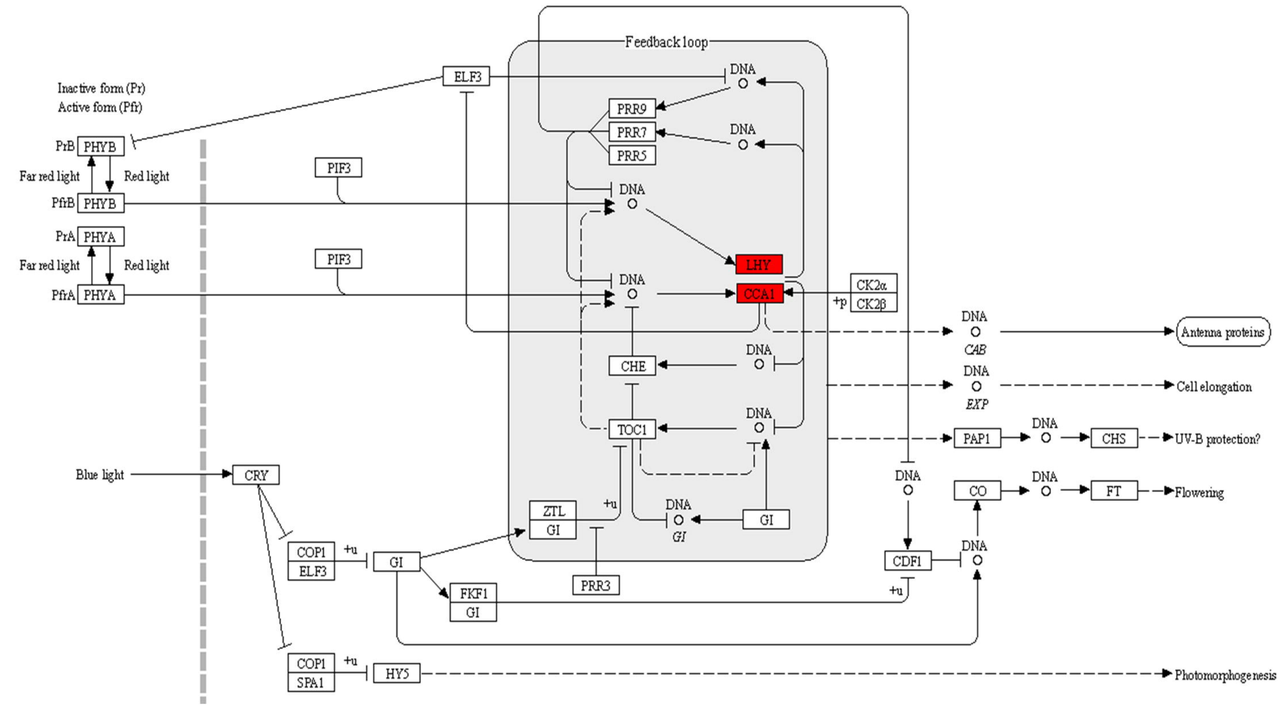
Salvinia cucullata	V1.2	Streptophyta (Polypodiopsida)	FernBase
Scenedesmus obliquus	UTEX 393	Chlorophyta (Chlorophyceae)	Phycocosm
Selaginella moellendorffii	V1.0	Streptophyta (Lycopodiopsida)	Phytozome
Solanum lycopersicum	ITAG4.0	Streptophyta (Eudicotyledons)	Phytozome
Sorghum bicolor	454 v3.1.1	Streptophyta (Liliopsida)	Phytozome
Sphagnum magellanicum	V1.1	Streptophyta (Bryophyta)	Phytozome
Spirogloea muscicola	CCAC 0214	Streptophyta (Zygnemophyceae)	Phycocosm
Thuja plicata	v.3.1	Streptophyta (Acrogymnospermae)	Phytozome
Triticum aestivum	V2.2	Streptophyta (Liliopsida)	Phytozome
Ulva mutabilis	Foyn	Chlorophyta (Ulvophyceae)	Phycocosm
Volvox carteri	V2.1	Chlorophyta (Chlorophyceae)	Phytozome
Zea mays	RefGenv4	Streptophyta (Liliopsida)	Phytozome

499

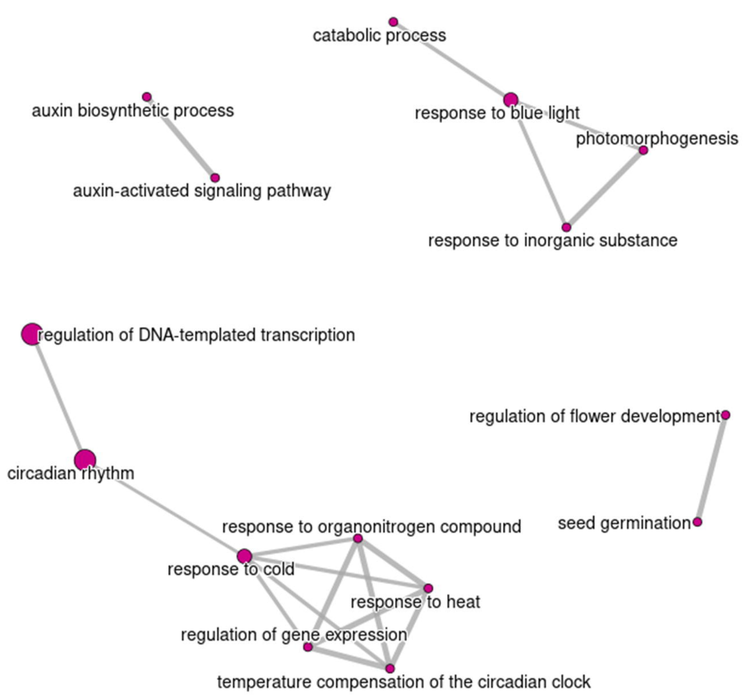
A



C

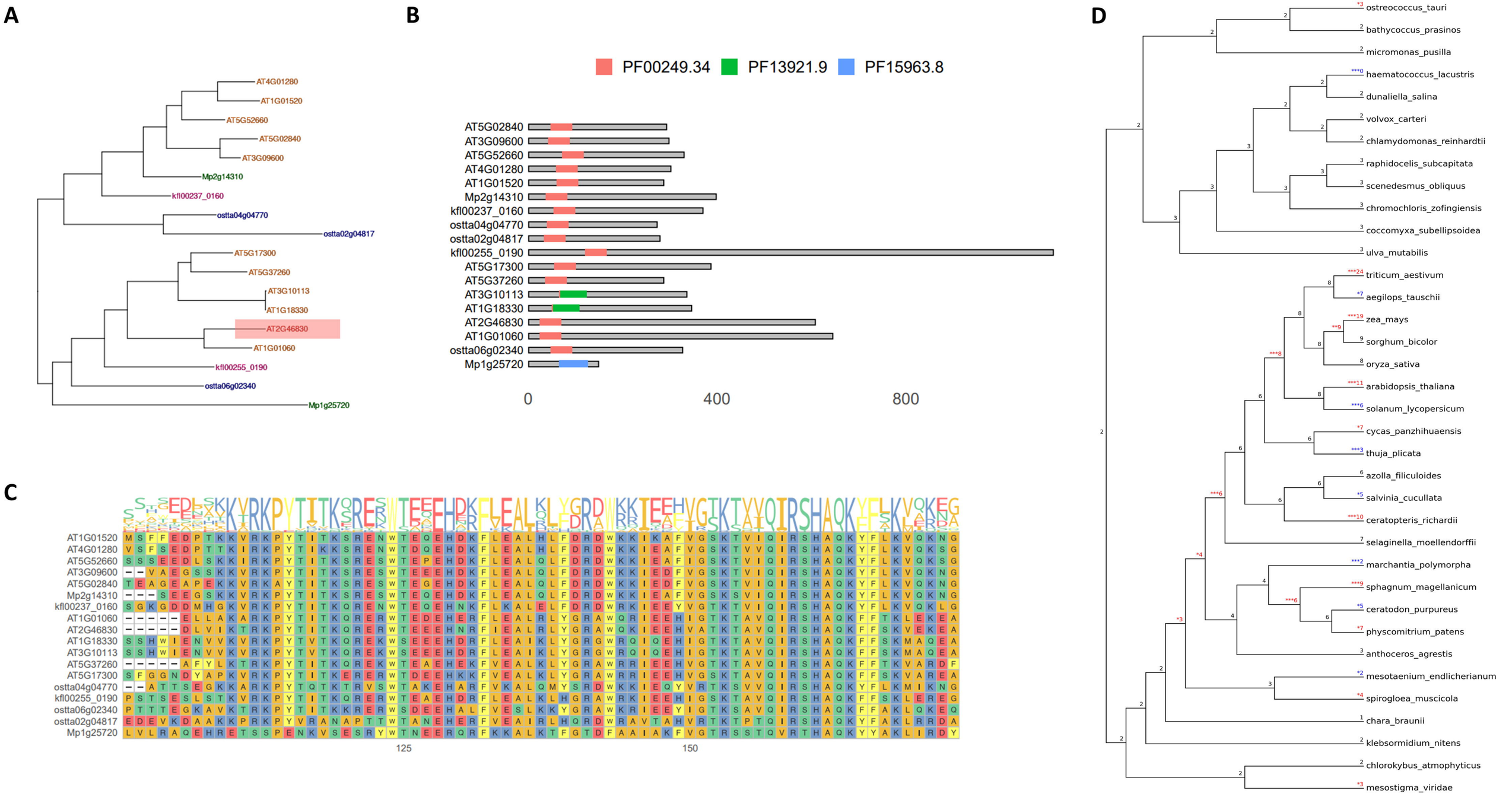


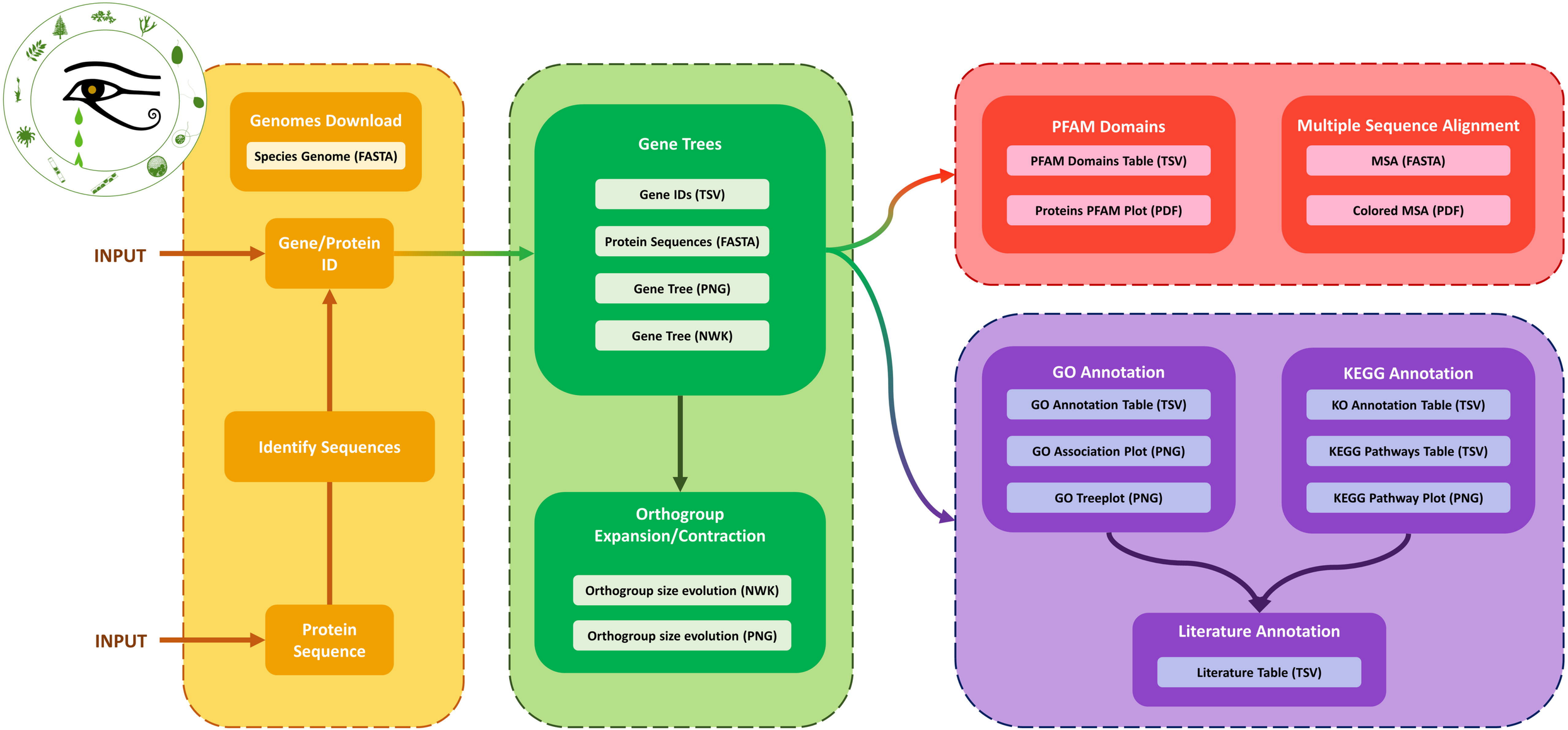
B

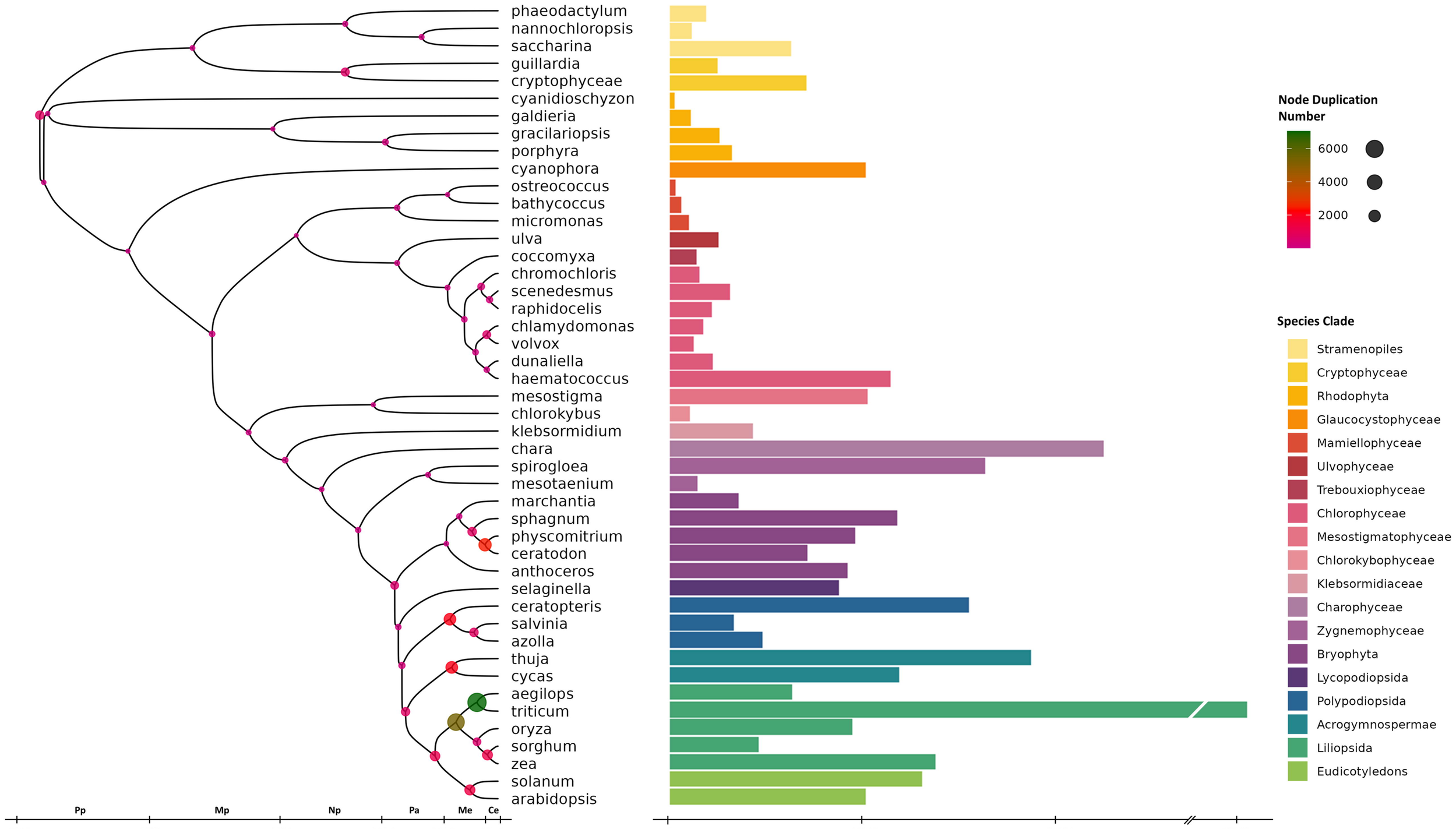


D

Source	Interaction Type	Target	Pubmed ID
CCA1	BINDS	TOC1	17616736
CCA1	REPRESSES	H3 ACETYLATION	17616736
CCA1	ANTAGONIZES	TOC1 MRNA ABUNDANCE	17616736
CCA1	DIFFERS IN DIURNAL RHYTHM	BETWEEN FLOWERING ECOTYPES	17631524
CCA1	REGULATED BY	LIGHT	17873091
CCA1	STABILIZED BY	FAR-RED LIGHT	17873091
RED AND BLUE LIGHT	DEGRADES	CCA1	17873091
CCA1	CONTAINS INSTABILITY DETERMINANTS IN	CODING REGION	17873091
CCA1	INTERACTS WITH	LHY	17877705
PRR7	REPRESSES	CCA1	17877705
CCA1	REPRESSES	SVP ABUNDANCE	19011118
CCA1	REPRESSES	TOC1	19029881


A
AT4G01280
AT1G01520
AT5G52660
AT5G02840
AT3G09600
Mp2g14310
kfl00237_0160
ostta04g04770
ostta02g04817
AT5G17300
AT5G37260
AT3G10113
AT1G18330
AT2G46830
AT1G01060
kfl00255_0190
ostta06g02340
Mp1g25720
B
PF00249.34
PF13921.9
PF15963.8
0
400
800
C
125
150
D
ostreococcus_tauri
bathycoccus_prasinos
micromonas_pusilla
haematococcus_lacustris
dunaliella_salina
volvox_carteri
chlamydomonas_reinhardtii
raphidocelis_subcapitata
scenedesmus_obliquus
chromochloris_zofingiensis
coccomyxa_subellipsoidea
ulva_mutabilis
triticum_aestivum
aegilops_tauschii
zea_mays
sorghum_bicolor
oryza_sativa
arabidopsis_thaliana
solanum_lycopersicum
cycas_panzhihuaensis
thuja_plicata
azolla_filiculoides
salvinia_cucullata
ceratopteris_richardii
selaginella_moellendorffii
marchantia_polymorpha
sphagnum_magellanicum
ceratodon_purpureus
physcomitrium_patens
anthoceros_agrestis
mesotaenium_endlicherianum
spirogloea_muscicola
chara_braunii
klebsormidium_nitens
chlorokybus_atmophyticus
mesostigma_viridae


Genomes Download
Species Genome (FASTA)
INPUT
Gene/Protein ID
Identify Sequences
INPUT
Protein Sequence
Gene Trees
Gene IDs (TSV)
Protein Sequences (FASTA)
Gene Tree (PNG)
Gene Tree (NWK)
Orthogroup Expansion/Contraction
Orthogroup size evolution (NWK)
Orthogroup size evolution (PNG)
PFAM Domains
PFAM Domains Table (TSV)
Proteins PFAM Plot (PDF)
Multiple Sequence Alignment
MSA (FASTA)
Colored MSA (PDF)
GO Annotation
GO Annotation Table (TSV)
GO Association Plot (PNG)
GO Treeplot (PNG)
KEGG Annotation
KO Annotation Table (TSV)
KEGG Pathways Table (TSV)
KEGG Pathway Plot (PNG)
Literature Annotation
Literature Table (TSV)


phaeodactylum
nannochloropsis
saccharina
guillardia
cryptophyceae
cyanidioschyzon
galdieria
gracilariopsis
porphyra
cyanophora
ostreococcus
bathycoccus
micromonas
ulva
coccomyxa
chromochloris
scenedesmus
raphidocelis
chlamydomonas
volvox
dunaliella
haematococcus
mesostigma
chlorokybus
klebsormidium
chara
spirogloea
mesotaenium
marchantia
sphagnum
physcomitrium
ceratodon
anthoceros
selaginella
ceratopteris
salvinia
azolla
thuja
cycas
aegilops
triticum
oryza
sorghum
zea
solanum
arabidopsis
Pp
Mp
Np
Pa
Me
Ce
2200
1600
1000
538.8
251.9
66
0
0
10000
20000
50000
Node Duplication Number
6000
4000
2000
Species Clade
Stramenopiles
Cryptophyceae
Rhodophyta
Glaucocystophyceae
Mamiellophyceae
Ulvophyceae
Trebouxiophyceae
Chlorophyceae
Mesostigmatophyceae
Chlorokybophyceae
Klebsormidiaceae
Charophyceae
Zygnemophyceae
Bryophyta
Lycopodiopsida
Polypodiopsida
Acrogymnospermae
Liliopsida
Eudicotyledons