PharaohFUN: PHylogenomic Analysis foR plAnt prOtein History and FUNction elucidation

Authors: Marcos Ramos-González, Víctor Ramos-González, Emma Serrano-Pérez, Christina Arvanitidou, Jorge Hernández-García, Mercedes García-González, Francisco José Romero-Campero

Abstract

Since DNA sequencing has become 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.

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, expansion and contraction events, ancestral states, domain identification, multiple sequence alignments, and diverse functional annotation. It also incorporates different search modes to facilitate its use and increase its reach within the community. Tests were performed on the whole transcription factor toolbox of *Arabidopsis thaliana* and on CCA1 protein to assess its utility for both large-scale and fine-grained phylogenetic studies. These exemplify how PharaohFUN accurately traces the corresponding evolutionary histories of these proteins by unifying results for land plants, streptophyte and chlorophyte microalgae. Thus, PharaohFUN democratizes access to these kind of analyses in photosynthetic organisms for every user, independently of their prior training in bioinformatics.

Introduction

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

One clade where this phenomenon has a marked impact is Viridiplantae. In 2020, the INSDC (International Nucleotide Sequence Database Collaboration) (Arita et al. 2021) contained genomes for more than 800 species of this group, but their distribution remains heavily unbalanced: while 543 correspond to angiosperms, other 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 computational or 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 classic problems in genomics. One such example is the assignment of relationships between genes. In this context, software based on phylogenetic methods, such as SHOOT, has been found to be more accurate than those based on pairwise alignments, such as BLAST (Emms and Kelly 2022). Another instance is the annotation of genomic regions. Here, TOGA incorporates

information from orthologous genes to delineate coding and non-coding regions, outperforming previous methods in terms of annotation completeness (Kirilenko et al. 2024). Finally, the use of ancestral reconstructed sequences to train Protein Language Models has been shown to improve their accuracy. In particular, this has led to the development of more meaningful epistasis representations than those produced by larger models trained without incorporating evolutionary data (Matthews et al. 2023).

In this work, we introduce PharaohFUN, a web-based tool aiming to facilitate the exploration of the evolutionary landscape of photosynthetic eukaryotes. PharaohFUN also seeks to facilitate inter-species transfer of functional information based on orthology relationships between genes. To achieve this, PharaohFUN incorporates genomic information from 167 species spanning Prasinodermophyta, Chlorophyta and Streptophyta plus 4 rhodophytes as outgroups. Thus, PharaohFUN currently supports one of the largest numbers of plant species among the software dedicated to orthology and functional transfer studies, especially for neglected non-angiosperm plant species. The phylogenetic placement of these major clades establishes a practical framework for drawing hypotheses about the evolutionary relationships between genes.

PharaohFUN integrates phylogenetic and evolutionary contexts of each gene, along with sequence and domain studies and leverages species with refined annotation. In turn, this allows users to build reliable hypotheses about the evolutionary history and function of a gene of interest in a few minutes. Unlike most existing phylogenetic software, PharaohFUN not only accelerates studies of individual genes. Through its Batch Mode, it allows simultaneous analysis of long lists of genes and sequences, facilitating large-scale evolutionary analyses. At the same time, it does not rely solely on its database species but allows the inclusion of sequences from external organisms. In addition to traditional tools for domain, sequence and protein function exploration, PharaohFUN incorporates features absent in other phylogenetic software. These include ancestral state reconstruction of gene families, protein-protein interactions and literature annotation.

We tested the utility and reliability of PharaohFUN by applying it to two case studies. These demonstrate its ability to infer the evolutionary trajectories and function of individual and entire sets of genes. The results support previous conclusions about evolutionary events related to terrestralization and central circadian clock proteins. For all these reasons, PharaohFUN is positioned as a tool with great potential for end-to-end phylogenetic studies in plant organisms, saving users large amounts of time and resources without results precision reduction.

1. Analysis pipeline

PharaohFUN is a web-based tool written in R, using the practical framework of the Shiny package. The complete analysis pipeline is summarized in Fig. 1A. PharaohFUN includes the proteomes of 167 species representing the main clades of the green lineage, namely Prasinodermophyta, Chlorophyta and Streptophyta (Sup. Table S1). As for the distribution of species by clades, PharaohFUN aims for a balanced representation avoiding clustering biases.

This model has been generated using *Orthofinder* v3.0 (Emms and Kelly 2019; Emms et al. 2025) on the proteomes of the species mentioned in Sup. Table S1. The proteomes were previously filtered to include only the protein products of the primary transcript, as recommended in *Orthofinder* best practices. It ensures that the clustering remains 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 final model was generated in two steps. Firstly, a subset of 58 species homogeneously distributed throughout the complete Viridiplantae clade was selected, prioritizing species from different groups with refined annotation (Fig. 1B). The corresponding orthogroups (OGs) were calculated using the Multiple Sequence Alignment (MSA) function of the software, incorporating 4 rhodophyte species as outgroup to Viridiplantae within Archaeplastida. We will refer to this first model with a balanced sampling of species as the phylodiverse model. Then, we assigned the genes from the remaining

proteomes to those precomputed OGs, adding them to MSAs and gene trees, and reconciling the latter with the bibliographically validated species tree (Strassert et al. 2021; Timilsena et al. 2022) (Sup. Fig. 1).

The final model contains 4,529,870 genes assigned to 97,610 OGs, representing 92.8% of the total genomes size. G50 statistic measures the number of genes in an OG such that 50% of the genes are in OGs of that size or larger. It corresponds to is 437, and 56 OGs contain genes from all species. Certain species exhibit a significantly higher number of identified gene duplications compared to other species within their respective clades (Fig. 1B). Notable cases are those of *Mesostigma viride*, *Spirogloea muscicola*, *Abies alba*, *Alsophila spinulosa* and *Cymbomonas tetramitiformis* attributed to whole-genome duplication (WGD) or triplication events and polyploidy (Cheng et al. 2019; Liang et al. 2020; Stull et al. 2021; Huang et al. 2022; Qiao et al. 2022; Cui et al. 2023; Gyaltshen et al. 2023).

Due to the variation in the number of genes per species in the OGs, an analysis of their contractions and expansions was performed using *count* (Csűös 2010) software with a fossil-calibrated species tree. The resulting tree from *Orthofinder* workflow was dated using the relaxed molecular clock model implemented in the R package *ape*. The selected model accounts for different rates of evolution among nodes drawn from a gamma distribution, with constraints derived from fossil record and TimeTree (Kumar et al. 2022) (Sup. Table S2). To allow for the reconstruction of ancestral states in all OGs, an asymmetric Wagner parsimony model was used. The software's default gain penalty was relaxed to 4, to account for horizontal gene transfer (HGT) and whole genome duplications (WGDs) events known in this clade. (Wickell and Li 2020; Ma et al. 2022; Clark 2023).

Finally, existing Gene Ontology (GO) annotations for each proteome were downloaded and KEGG annotations were generated using KAAS (Moriya et al. 2007) with 20 photosynthetic eukaryotic organisms as reference. In the case of STRING data (Szklarczyk et al. 2019), the proteomes were mapped against the database sequences, using a 95% identity threshold, to associate the

PharaohFUN IDs with those of STRING and generate the table of physical interactions. The type of evidence for the interaction was also associated: direct evidence or interolog.

2. Tool architecture

The OGs, gene trees, expansion/contraction/gain/loss models and annotations derived from the described workflow served as the basis of the results provided by PharaohFUN for each search. The complete architecture of the application is shown in Fig. 2.

The first step of any analysis with PharaohFUN consists of the identification of the OG of interest and its position in the tree of the input. The tool is designed to be flexible with respect to the input type to solve the ambiguity problems that exist in many databases. There are 5 search options: by gene ID (typically following the notation used in Phytozome (Goodstein et al. 2012) and Phycocosm (Grigoriev et al. 2021)), by sequence, by OG ID, in batch mode and by sequence of an organism not present in the database. In this last case, the DIAMOND aligner (Buchfink et al. 2015) is used to search for the most similar sequence in the supported proteomes when the input sequence actually belongs to a supported organism or to one evolutionarily close to some of the supported species. When the sequence belongs to a highly divergent organism from the ones supported, PharaohFUN incorporates a search based on the novel software tool SHOOT (Emms and Kelly 2022). This search mode identifies the most likely OG and includes the new sequence in the correct position of the gene tree. Finally, PharaohFUN incorporates a batch mode for the analysis of complete gene lists, allowing their inclusion as gene IDs or sequences from the same organism.

Once the OG to which the query belongs has been identified, PharaohFUN incorporates a series of features to extract information about its evolutionary context and function. They are organized in three main modules, as follows.

2.1. Gene Trees and Orthogroup Size Evolution

This module creates and represents gene trees corresponding to specific OGs, allowing for the selection of the species to be included and enabling users to restrict analyses to specific clades of interest. The tool incorporates four alternative methods for its construction: Maximum-Likelihood, Bayesian Inference, Neighbour Joining and UPGMA. While Maximum-Likelihood trees are directly derived from the analysis pipeline, the other methods rely on the MSAs calculated for each OG. Maximum-Likelihood corresponds to resolved gene trees derived from *Orthofinder* workflow, i.e. gene trees computed using *FastTree* and reconciled with the species tree (Emms and Kelly 2019). Neighbour Joining and UPGMA options directly compute the gene trees for the selected sequences from the corresponding MSAs, employing a bootstrap of 100, using the R package *phangorn* (Schliep 2011). Neighbour Joining resulting trees are then rooted by the midpoint of their longest branch. Trees determined by Bayesian Inference were precomputed using the Markov Chain Monte Carlo algorithm implemented in *ExaBayes* (Aberer et al. 2014). The generation of these trees was conducted using two independent runs and two coupled chains for each one, using a convergence criterion to stop the exploration of the probability landscape. The complete set of parameters employed is described in the file Sup. Table S3. Trees derived from all these building options are visualized using the *ggtree* package (Yu 2020). Their interactive exploration is made possible through *phylowidget* (https://github.com/sdwfrost/phylowidget). In addition, it is possible to collapse subclades to restrict the search, removing clades and not entire species.

The “Orthogroup Size Expansion/Contraction” tab complements the previous information exploring expansion, contraction, gain and loss events undergone by an OG throughout its evolution. This feature uses the output of *count* to explore changes in OG sizes mapped onto the species tree. It reconstructs the number of genes belonging to an OG in the common ancestors of each clade and the expansions, contractions, gains and losses suffered by the OG along each branch. In this tab, an interactive species tree is displayed with information on the number of genes within the OG in each node, along with the evolutionary events determined in each branch. Its MRCA is also highlighted.

2.2. Domains and Sequence Similarities

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

In the MSA tab, two methods are provided: ClustalOmega and MAFFT. The former uses the ClustalOmega algorithm (Sievers et al. 2011) to align selected sequences. The latter uses the complete OG MSA generated using MAFFT and filters out non-selected proteins. Columns that correspond only to gaps in the sequence subset are removed. 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 an interactive alignment viewer that can also be used to find motifs using regex notation.

2.3. Functional Annotation

One of the 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 four different analyses: GO enrichment, KEGG orthology, STRING interactions and literature-based annotation.

In the “GO Terms” tab, the gene ontology terms of a subset of genes in the tree are listed. This analysis functionally characterizes genes based on existing functional annotations of phylogenetically close genes A list of existing GO terms is displayed for each gene summarized in two different graphical representations. The first one is an emapplot visualizing relationships

between different groups of terms based on connections by genes involved in different processes. The second representation is a treeplot clustering GO terms according to their semantic similarity, providing a summarized view.

In the “KEGG Orthology” (KO) tab, once genes are selected, their corresponding KO terms are identified. Subsequently, KEGG pathways are determined based on KO terms present in the selection. This analysis reveals the different molecular pathways in which the selected genes could be involved. The output is presented in a tabular format, and users can also choose to visualize a specific pathway to explore in which steps the selected genes act.

The “STRING Interactions” tab incorporates information on protein-protein interactions contained in the STRING database (Szklarczyk et al. 2019). The tool displays all the proteins in the organism that interact with each of the chosen ones, as well as information on the evidence supporting this interaction. In turn, the tool analyzes OG co-evolution by identifying to which OG each of the target proteins belongs to. Moreover, it also generates a graphical representation of the interaction network of a subset of proteins of interest and a link to the corresponding interactive network. For the proteins in the network, 3D structure and SMART domains information can be accessed.

Finally, the “Literature Annotation” tab provides a connection to the PlantConnectome web tool (Lim et al. 2024). 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 previous tabs with other bibliographically supported biological information. Additionally, a link to the paper from which each piece of information has been inferred is provided to expand or verify the accuracy of the results.

3. Case studies

PharaohFUN has been developed for two distinct types of studies. On the one hand, it allows the fine-grained study of the evolution of a gene of interest and the annotation of associated genes. On

the other hand, it can simplify and accelerate phylogenomic studies for large numbers of genes. Its use in these scenarios is presented by two representative case studies.

3.1. Case study 1: Characterization of a putative CCA1 in *Helianthus annuus*

To demonstrate PharaohFUN's utility for fine-grained studies, the sequence of a putative CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) protein from *Helianthus annuus* (entry A0A251TYD1) was downloaded from UniProt and analyzed using the "Sequence from new organism" function. CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) is a central regulator of circadian rhythms largely conserved throughout Viridiplantae. *A. thaliana, B. rapa, A. hypocondriacus, M. polymorpha, O. tauri* and *P. patens* were selected as species of interest. The resulting gene tree showed a division into three main clades (Sup. Fig. 2). The query protein was located in that shown in Fig. 3A, split in two subclades.

The first subclade contained *A. thaliana* CCA1 and LHY, along with the REVEILLE family members RVE1, 2 and 7/7-like. It also included the previously identified CCA1 orthologs in *P. patens* (CCA1a and CCA1b) and *O. tauri* (Okada et al. 2009; A.-M. Linde et al. 2017). In agreement with this clade representing CCA1/LHY orthologs, no *M. polymorpha* sequences were included (A.-M. Linde et al. 2017). We can observe that the protein of *H. annus* groups with that of *A. hypocondriacus*, also from superasterids clade. As for *B. rapa*, it presents the gene distribution indicated in (Lou et al. 2012) as a consequence of a whole genome triplication and gene loss after its divergence from *A. thaliana*. Thus, two LHY copies and one CCA1 are present, together with a higher number of RVEs. However, it presents an additional protein not described in the previous study. This protein is Brara.I05586, present in the LHY/CCA1 clade with lower sequence similarity to these two than the aforementioned copies.

The second clade was constituted by proteins traditionally referred to as the LHY-CCA1-Like (LCL) family. Within this clade, a gene from *M. polymorpha* known as MpRVE, and one from *O.*

tauri were identified, supporting predictions from previous analysis (A.-M. Linde et al. 2017). Similarly, the known increase in the number of copies of RVE 4/8/6 and the retention of a single copy of RVE3 and RVE5 in *B. rapa* (Lou et al. 2012), as well as the existence of three PpRVE copies (A.M. Linde et al. 2017) were detected.

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. 3B) a contraction in the branch leading to *M. polymorpha* and expansions in that of *B. rapa*. To gather more evidence and refine previous results, a PFAM domain analysis was conducted with PharaohFUN. The proteins from the gene tree were found to contain the Myb-like DNA-binding domain (PF00249) and often a REVEILLE 6-like domain (PF24904). This domain is known to be highly prevalent in transcription factors that constitute the central circadian clock (Carré and Kim 2002). Accordingly, *Helianthus annus* putative CCA1 ortholog showed PF00249 domain. The rest of the orthogroup proteins present this domain in different locations and combinations (Sup. Fig S3). Next, we performed MSA using the sequences of these proteins plus two different sequences from an external clade (AT1G75250 and AT3G10595). Using the PharaohFUN function “Find motifs”, the conserved SHAQKYF motif was searched for. This motif is typically associated with RVE and CCA1 orthologs (A.-M. Linde et al. 2017). It was identified in all proteins (or its variant SHAQKFF) except for the two Myb-like DNA-binding proteins not clustered in the CCA1-LCL clade (Fig. 3D). Interestingly, in the *B. rapa* Brara.I05586 protein, this domain is truncated, which may have led to loss of function and explain its distance with respect to both CCA1 and LHY.

We applied the remaining functionalities in PharaohFUN to leverage the evolutionary results obtained thus to elucidate potential functions for the uncharacterized gene in this example. It is important to consider that the species whose genes are used to transfer functional information can greatly impact the accuracy of the results. Therefore, the study of clades in the tree is crucial for selecting genes that may carry information of interest for a particular study gene. Exploration of GO terms from genes belonging to the CCA1 clade revealed processes mostly related to circadian

rhythm transcription and response to cold and, with lower abundance, auxin biosynthesis, seed germination, flower development, catabolic processes, response to blue light and temperature compensation of the circadian clock (Fig. 4A). In addition, the KO annotation confirmed the presence of homologs of CCA1 and LHY in the orthogroup, highlighting their role in pathways related to circadian rhythm (Fig. 4B). The literature annotation feature provided valuable biological information, including descriptions of relationships such as complex formation with LHY (which belongs to the same OG), repression of TOC1 through binding, control over circadian rhythms, and more than 200 further relationships (Fig. 4C). These annotations can be used to experimentally test hypotheses regarding conservation of function among predicted orthologs.

Finally, to study coevolution between OGs with interacting proteins, the STRING function was employed, selecting as input the CCA1 orthologs identified in *O. tauri*, *P. patens* and *A. thaliana*. The OG whose proteins interacted with this subset a higher number of times was found to correspond to ID OG0001357 (Sup. Table S4), which was entered into the OG search mode developed for this purpose. This revealed an orthogroup comprising PHYTOCHOME A-E. The role of the interaction between CCA1/LHY and PHYB in the response to far-red light has been well described, regulating processes such as shade avoidance, circadian clock entrainment and photomorphogenesis (Yeom et al. 2014; S. Wang et al. 2022; Bian et al. 2025). These proteins also appeared when plotting the interaction network of CCA1 in PharaohFUN (Fig. 4D). The second one, OG0000363, was found to contain a cluster of genes related to Lysine-Specific Demethylase 1 (LSD1)-like histone demethylases LDL1/2. In addition to appearing in CCA1 interaction network (Fig. 4D), there is evidence in the literature that LDL1/2 can interact with CCA1 and LHY to repress TOC1 expression (Hung et al. 2018) and that these proteins are conserved in the green lineage (Martignago et al. 2019), validating our results.

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, even when studying species not included in

the dataset. Based on these analyses, functional annotations can be associated with unknown proteins (Gil et al. 2023).

3.2. Case study 2: Analysis of the evolutionary history of the complete endowment of TFs of *A. thaliana* across Viridiplantae

PharaohFUN can also be used as a tool for large-scale phylogenomic studies of any gene set of interest. To demonstrate this, an analysis was performed to explore the evolutionary history throughout Viridiplantae of the complete set of transcription factors (TFs) of *Arabidopsis thaliana*. Traditionally, this history has been generated individually for each TF family or even their subclades. Here we show how PharaohFUN can be used to perform a global study in a quick and easy way. Our results are shown to be consistent with the literature as demonstrated in terms of tree topologies and ancestral state of OGs. The orthologs recovery by PharaohFUN as a function of evolutionary distance has been also tested.

3.2.1. PharaohFUN reconstructs known orthology and gene tree topologies

The complete set of transcription factors sequences from *A. thaliana* was inputted to PharaohFUN Batch Mode selecting 54 species in the green lineage. Using the classification of PlantTFDB (Jin et al. 2017), the 51,892 sequences resulting from this search were classified into 58 families of TFs. PharaohFUN Batch Mode results were used to analyze the OGs that composed each family, their MRCA, their associated number of genes per species and their PFAM domains (Sup. Table S5-S6). Individual OGs were identified for 37 of the 58 families, while 10 were split into more than 5 OGs (Sup. Table S5). 8 OGs belonged to more than one family. This division into several OGs for specific TF families followed two main reasons: different MRCAs for subgroups of the same family and modification of functional domains in subgroups. Accordingly, many of them reflect accurately

the evolutionary history of TF families and the true topologies of their gene trees. Examples for this are BES1, WRKY and C2H2 families.

The BES1 TF family is an example of a domain-based division into different OGs. PharaohFUN split the BES1 family into two OG (Fig. 4A). The first one of them was present throughout the entire Viridiplantae from its MRCA and presented a Glycosyl hydrolase family 14 domain as well as some N-terminal BES1/BZR1 domains (Fig. 4B). The MRCA of the second OG arose in the ancestor of the clade constituted by Klebsormidiophyceae+Phragmoplastophyta, and only possessed the BES1/BZR1 domain. Genes from chlorophyte algae in the first OG revealed only the hydrolase domain, whereas BES1/BZR1 did not appear. Therefore, one OG of this family reflects the evolution of BES1 TF and the other represents the evolution of beta amylases that at a particular point acquired BES1 domains. This result is in agreement with previously published work indicating that streptophyte algae and land plants possess this TF family, unlike chlorophytes (Wilhelmsson et al. 2017; Mecchia et al. 2021).

The PharaohFUN results regarding the WRKY family, one of the largest TF families in *A. thaliana*, were analyzed to test its performance on families with complex evolutionary histories. The WRKY family was divided into 3 OGs. The first one corresponded to WRKY TFs (combining groups I, II and III) and the second, to the TIR-NB-LRR group (Kalde et al. 2003; Wang et al. 2015; Villacastin et al. 2021). As for the third OG, it was found to correspond to MAPK/ERK kinase kinases, with only one protein (WRKY19/MEKK4) exhibiting a WRKY domain in *A. thaliana*, in addition to its kinase domain. In turn, the topology of the WRKY tree reflected the phylogenetic studies performed on this group. Group III maintained the organization in four clades described in (Wang et al. 2015), while group I appeared independently and II showed the separation of the clades II-a, II-b, II-c, II-d and II-e (Huang et al. 2012) (Sup. Fig. S4A). Group II did not constitute a single entity, but subgroups IIa/b, IId/e y IIc were confined to phylogenetically distinct groups. This organization was already pointed out in (Mohanta et al. 2016) as a consequence of performing a phylogeny based on a larger number of plant species instead of just dicots. Thus, our results are consistent with their

revisions on this family phylogeny. In the same study, the authors pointed out the existence of chimeric proteins with a WRKY and a TIR, LRR or kinase domain occurring mostly in flowering plants and not present in all plant species, fully explaining the division into orthogroups returned by PharaohFUN. Finally, as for the TIR-NB-LRR OG, the analysis of its evolutionary history showed a complex scenario, with an origin prior to the division between chlorophytes and streptophytes, but abundant losses in multiple lineages (Sup. Fig. S4B). Among them, monocots stand out, while some chlorophyte algae such as *Chromochloris zofingiensis* and streptophytes such as *Klebsormidium nitens* exhibited representatives of these proteins, in agreement to previous findings (Tarr and Alexander 2009; Andolfo et al. 2019; Shao et al. 2019). Traditionally, this family was believed to have appeared in streptophyte algae due to the low number of representatives of the different clades of Chlorophyta included in the corresponding analysis (Shao et al. 2019), which exemplifies the importance of the inclusion of these organisms in evolutionary studies and justifies the high number of them included in our tool.

Finally, C2H2 is the family that was split into the largest number of orthogroups, 18. In this case, this division was not artifactual either, but faithfully represented the discordance presented in (Li et al. 2022) between the phylogeny of C2H2 zinc-finger proteins (C2H2 ZFPs) constructed for multiple plant species and the sequence similarity of the ZFPs in the individual species. Thus, the mapping of the different OGs onto the AtZFPs tree reported in that study allowed us to observe how the three main OGs corresponded to the three largest contiguous clades with proteins of the same global group (groups III, IVb and IVc) in the AtZFPs tree, whereas the appearance of proteins of different groups in the subclades promoted the split of the corresponding proteins into different OGs (Sup. Fig. S5).

This demonstrates that PharaohFUN's OGs reflect the evolutionary trajectories of the gene families and that they coincide with reported tree topologies. Thus, it can be reliably used for orthology and phylogenetic studies.

3.2.2. PharaohFUN correctly traces the evolutionary history of OGs

Our tool goes beyond gene trees and allows users to explore the evolutionary history of the different OGs, in terms of the distribution of the number of genes per species identifying expansion/contraction/gain/loss events. In addition to the individualized study of the aforementioned TF families, we used PharaohFUN to perform a global and integrated analysis of all these genes. Accordingly, the number of orthologs of *A. thaliana* genes was determined for each TF family over Viridiplantae. Fig. 6A shows the distribution of genes in each family, consisting of the sum of genes from all OGs associated with each family. For the same representation separated into OGs instead of TF families (direct PharaohFUN output), refer to Sup. Fig. S6. Overall, there was a substantial number of TF families without representatives in chlorophyte microalgae, whereas those that were present exhibited a lower number of genes compared to vascular plants. Likewise, streptophyte algae were postulated as a middle ground between both groups. Although most TF families are present in all or some organisms of this paraphyletic group, their gene abundance was low, similar to those of chlorophytes.

These results suggest that an important number of TF families exhibited today by *A. thaliana* arose prior to land colonization by the green lineage possibly promoting this key event in life evolution. To further explore this, we used the expansion, contraction, gain and loss events in OGs through the entire species tree identified using the Batch mode in PharaohFUN. This analysis also identified the position of the MRCA associated with each OG (Sup. Fig. S7). In addition, gains at the level of each internal node were calculated (Fig. 6B and Sup. Fig. S8). Our results indicated that many OGs corresponding to *A. thaliana* TF families are not of recent evolutionary origin, instead a large number (83) was already present in Viridiplantae MRCA. Likewise, the number of OGs that appeared in the MRCAs of streptophyte algae, Embryophyta and Magnoliopsida is particularly relevant, pointing to a progressive increase in the repertoire of TFs. Also interesting is the high number of Arabidopsis-specific gains (18), associated to OGs present only in this genus. OGs originated between the emergence of Streptophyta and Embryophyta were related to the TF families

shown in Table 1. It should be noted that no OG present in *A. thaliana* could have emerged in a clade that does not contain this species. This explains the absence of new OGs in the common ancestor of other groups. MRCA of TF families was compared to that determined by (Wilhelmsson et al. 2017) for families that allowed direct comparison and to (Jin et al. 2017) for the rest. To allow comparison, the MRCA of the oldest OG for each TF was manually curated, performing a domain analysis to discard OGs that did not correspond to the TF family, as we exemplified for BES1.

Thus, our inference matched theirs for 55 out of 58 families. As for the remaining three, we placed CAMTA MRCA in Chlorophyta+Streptophyta, consistent with (Jin et al. 2017) but prior to the inference in (Wilhelmsson et al. 2017), which may be due to the inclusion of more chlorophyte algal species in our dataset. In the case of VOZ, we detected its origin in Anydrophyta (Rensing 2020), because of its presence in *Zygnema circumcarinatum*, *Zygnema cf. cylindricum* and *Mesotaenium kramstae* genomes, which were not included in their datasets. Finally, we inferred the emergence of LSD in Phragmoplastophyta+Klebsormidiophyceae instead of Viridiplantae. Unlike the other studies, the aim of this one is to show the recoverability of orthogroups based only on *A. thaliana* proteins, so there may be other OGs that contain proteins with the domains corresponding to LSD TFs in chlorophytes but are the result of other acquisitions without conservation of the rest of the sequence, resulting in the observed results. Emergence and gain of the total OG set depicted in Fig. 6B, including those that did not constitute the main group of the corresponding TF family are listed in Sup. Table S7.

In addition to the emergence of new OGs and TF families, the number of genes associated to them has changed significantly during the evolution in Viridiplantae. By studying the number of expanded OGs in the MRCA of the different clades, we identified four key events in the development of *A. thaliana* TF families. Namely, the common ancestors of Anydrophyta, Embryophyta, Tracheophyta and Magnoliopsida (Fig. 6C and Sup. Fig. S9). In the case of Embryophyta, it correlates with new TF-associated OGs emergence events as described above.

However, the MRCA of Anydrophyta exhibited one of the most notable duplication events without the emergence of many OGs, second only to Tracheophyta.

OGs contractions in TFs families were also characterized (Sup. Fig. S10). Ancestors internal to Bryophyta, Lycopodiopsida and Polypodiopsida clades exhibited important contractions. Such events were more frequent at tips, probably reflecting a current evolutionary exploration, not necessarily successful adaptations that would be transmitted in the future. As a special case of these, losses (contractions leading to gene absence in the corresponding species) were especially relevant in chlorophytes (Sup. Fig. S11).

Many studies have been conducted to shed light on the molecular adaptations that took place in the aquatic ancestors of land plants that promoted their terrestrialization. With this case study, we show the utility of PharaohFUN to conduct this type of global analyses, saving time and making them available for every user. We have shown that the emergence of new OGs comprising TFs in *A. thaliana* has played an important role since the emergence of Streptophyta, accumulating new families that plants would later exploit for terrestralization. These results are in agreement with published work that links these innovations, for example, with high light irradiance tolerance in streptophyte algae (Rensing 2020). Some of the species from this group proliferate in shallow water or seasonal pools, requiring pre-adaptations to terrestrial conditions and, consequently, regulators of gene expression (Wilhelmsson et al. 2017; Lai et al. 2020). These results are also in agreement with the conclusions of other works which identified these OG gains to have taken place before the conquest of land (Wilhelmsson et al. 2017; Bowles et al. 2024). The specific OG expansions identified by PharaohFUN showed that this phenomenon is prevalent in the ancestors of Anydrophyta and the different Embryophyta clades, enabling the co-option of the existing families for other purposes (Table 1 and Sup. Table S7). These expansions have been thoroughly studied, with studies supporting that they allowed the preassembly of signaling and regulatory pathways that would be later used by land plants (Wilhelmsson et al. 2017; de Vries et al. 2018; Rensing 2020; Serrano-Pérez et al. 2022; Hernández-García et al. 2024). Originally, their function was different

from that exhibited by this group, for example, regarding phytohormone signaling (de Vries et al. 2018; Serrano-Pérez et al. 2022). Regarding contractions, although they were prevalent in certain clades of Viridiplantae, their influence on shaping the *A. thaliana* TFs toolbox was lower than that of expansions and new OGs emergence. This is further supported by the literature on contractions of TFs families being determinant in clades such as Bryophyta, but exhibiting a reduced number for the branches leading to *A. thaliana* (Wilhelmsson et al. 2017).

Taken together, these results illustrate how PharaohFUN can accurately trace globally the evolutionary history of the different OGs, assisting the formulation of hypotheses about the evolution of protein families of interest based on the changes in OG size. Ancestral State Reconstruction and expansion/contraction studies are widely used to hypothesize how gene gains and losses influenced the evolutionary trajectories of extant organisms (Gàlvez-Morante et al. 2024). However, their performance requires highly specific training. PharaohFUN democratizes access to these studies for photosynthetic organisms, with the accuracy shown in these case studies. It highlights the ability of PharaohFUN to perform not only individual fine-grained studies but also global phylogenetic ones.

3.2.3. PharaohFUN enables the recovery of orthologs at high evolutionary distances

In this case study, only the evolutionary history of OGs associated with *A. thaliana* TFs was traced. As a result, OGs not present in this species were not included in the analysis. We have determined the abundance of these to measure PharaohFUN's ability to retrieve genes with a shared evolutionary history. Accordingly, we downloaded the complete TFs sets in 7 different species from PlantTFDB and assessed how many appeared in our results to evaluate the sensitivity of our tool as a function of divergence time between the source species and the target species (Sup. Table S8). The result was a mean sensitivity of 0.84 across all species, with lowest value corresponding to *K. nitens* (0.74) and highest to *O. sativa* and *S. lycopersicum* (0.93). At the level of individual TF families

(Fig. 6E), the median of their sensitivities was 1 for all species (i.e., most families of each species were fully recovered by orthology with *A. thaliana*). The family with the lowest recovery rate in the different species was C2H2, with a median of 0.44. The overall high percentages point to a low abundance in other species of genes in OGs associated with TF families not present in *A. thaliana*. Furthermore, the distribution of sensitivities along the species tree suggests high effectiveness in identification of orthologs even at large evolutionary distances, probably due to the homogeneous species sampling employed in PharaohFUN.

Subsequently, TFs from *B. prasinos*, *P. patens*, *S. bicolor* and *V. carteri* were used as additional input to that of *A. thaliana* and the recovery rates of the previous seven species were recalculated (Fig. 6F and Sup. Table S9). By introducing representatives of clades in which *A. thaliana* is not present, OGs whose origin can be traced back to the common ancestor of those clades could be identified, improving sensitivity despite the very small number of new species added. The overall mean sensitivity was 0.90 in this case. The species that now presented the highest percentage of TFs recovered was *O. sativa* (0.97), probably due to the influence of *S. bicolor* in the identification of TFs originated at Liliopsida MRCA. On the other hand, the species with the lowest value continued to be *K. nitens*, unsurprising because species encompassing common ancestors of this organism not shared with *A. thaliana* were not added. The same phenomenon is detected in *S. lycopersicum*, which shows a minimal recovery rate improvement. The median recovery rate per TF family for each species remained 1 without exception, and the family with the lowest value was LFY (0.75). C2H2 family now showed a median of 0.88, compared to 0.44 without the input of the new species. This demonstrated that new OGs were recovered by improving the sampling in clades, not related to a poor performance of the tool, but to true evolutionary histories of different OGs.

The previous results demonstrate how PharaohFUN is able to recover a large percentage of the proteins associated with TFs in *A. thaliana* from other plant species, while maintaining a correct topology of the resulting gene trees and an overall correct MRCA inference of each group. We have demonstrated its adequacy with results from well-established tools for TFs analysis such as

TAPscan (Petroll et al. 2025). TAPscan focuses on transcription-associated proteins and excels in this particular case study, due to being domain-oriented software and leveraging a balanced species sampling, using a similar two-step model creation as that of PharaohFUN. This domain orientation is especially advantageous for TF evolution studies, in which proteins are characterized by extensive functional domains that can be used to classify each family. It avoids the fragmentation into OGs that can be observed for some large families with widespread domains across the plant genomes. However, PharaohFUN does not use this paradigm since it may not be valid in the case of proteins that do not present functional domains and exhibits important limitations when applied to proteins containing domains of reduced extension. In these instances, the phylogenetic signal is not sufficient to create a reliable phylogeny of the proteins under study (Sjölander et al. 2011). PharaohFUN, being a general tool for the study of any protein set, is OGs-oriented like other tools of this nature, such as eggNOG-DB (Hernández-Plaza et al. 2023) and pico-PLAZA (Vandepoele et al. 2013). Therefore, the presented case study is one of the most potentially challenging analysis that can be carried out by PharaohFUN when compared against domain-oriented software, due to the large family sizes being evaluated. However, we have been able to verify how a high number of families (37) were contained in a single OG, how the expected split into OGs is usually motivated by the evolutionary history of the family as corroborated by specific literature, and how the evolutionary events at the species level were generally in agreement with the results from state-of-the-art specialized software such as TAPscan.

All these results demonstrate the robustness of PharaohFUN as a tool to detect orthologs of genes of interest at large scale. The selected species set allow for orthology relationships to be detected at high evolutionary distances, by combining uniform sampling of clades and the inclusion of streptophyte algae. This case study showcases the great amount of evolutionary data that PharaohFUN outputs by using a simple gene or sequence list as input.

4. Comparison with other web-tools and databases

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 (Sillitoe and Furnham 2016), eggNOG-DB (Hernández-Plaza et al. 2023), PhylomeDB (Fuentes et al. 2022) or, exclusively for plant species, pico-PLAZA (Vandepoele et al. 2013), which cater to specific objectives and capabilities differing from those of PharaohFUN, as summarized in Table 2.

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 at the default tree level (Sup. Fig. S12A) despite its widespread presence across Viridiplantae (A.-M. 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 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 OG 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 result.

To some extent, pico-PLAZA pursues a similar goal to PharaohFUN’s but, while PharaohFUN aims to uniformly represent all Viridiplantae groups, pico-PLAZA focuses on chlorophyte and stramenopiles algae. In this sense, although pico-PLAZA targets 11 chlorophyte genera, it lacks representation for streptophyte algae, which would serve as a liaison to the 4 land plants considered

in its database, and only supports one non-angiosperm embryophyte species (*P. patens*). The lack of linking elements between these evolutionarily distant groups produces separation into isolated instances. Although this may lead to the creation of refined gene groups, it hinders an integrated analysis of OGs. Bias can be incorporated in the clustering process towards the creation of small highly similar gene groups corresponding to very closely related species excluding related but distant genes from OGs. 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 OGs along the species tree, a feature not supported by pico-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. It also enables annotation and representation of KEGG pathways, STRING networks, structural models for STRING-supported genes and coevolution of OGs associated with interacting proteins. While SHOOT lacks the functionalities to directly perform functional analysis for a whole OG and pico-PLAZA indirectly allows it (through individual gene search, selection and custom gene set analysis restricted for registered users in Workbench), PharaohFUN overcomes this limitation by providing specific analysis features for these tasks within the same platform, always centered in whole OGs or custom genes and clades. As a comparison with PharaohFUN, the search for the gene encoding CCA1 in *A. thaliana* in pico-PLAZA found an orthologous gene family (ORTHO03P024982) with only four angiosperm clade-specific genes (Sup. Fig. S12B), despite its presence in chlorophyte species such as *O. tauri*, reflecting the aforementioned bias due to lack of evolutionary cohesion (A.-M. Linde et al. 2017). There exist other PLAZA instances focused on specific groups, such as dicots and monocots (Van Bel et al. 2022), but all suffer from the limitation in generalization for evolutionary studies over long distances due to the lack of representatives outside those clades.

This limitation arises from the trade-off between exploring the full evolutionary range of plant evolution and the need to use genomes with sufficient completeness and contiguity for

chromosomal position studies (Vandepoele et al. 2013; Van Bel et al. 2022). Thus, while PharaohFUN focuses on the study of proteomes across clades without providing this feature, pico-PLAZA sacrifices this continuity in species sampling in favor of incorporating data from genomic regions, at the cost of not including species that serve as a link between land plants and chlorophyte algae. The impact of this design is accentuated by the fact that land plant species supported by pico-PLAZA have undergone several Whole Genome Duplication events, making it more likely to create OGs more focused on that particular clade, without recovering their algal counterparts (Wang et al. 2022). While consistent with all other instances of PLAZA, this decision effectively leaves unsupported a group of organisms currently heavily represented in evolutionary studies, where omics and molecular studies are already routine (Bierenbroodspot et al. 2024).

eggNOG-DB (Hernández-Plaza et al. 2023) is a popular tool for performing comparative genomic analysis on a wide range of organisms, including Viridiplantae. It is focused on complete OG information, with individualized data available in the tree representation. It includes detailed functional annotation, incorporating GO, KEGG, CAZy and BiGG data. However, it incorporates a very low proportion of non-angiosperm land plants and green algae within its Viridiplantae set, making it difficult to use in plant evolution studies outside of flowering plants. It does not offer a batch mode, but fully automatic large-scale functional annotation of the sequences can be performed via eggNOG-mapper v2 (Cantalapiedra et al. 2021), making it a powerful tool for transferring functional data. However, this analysis does not report gene trees, MSAs or information on the evolutionary history of the sequences. Currently, eggNOG-mapper v2 supports up to version 5.0 of eggNOG-DB. In this case, eggNOG-DB 6.0 did recover the CCA1/LHY ortholog in *O. tauri* (Sup. Fig. S12C).

Finally, PhylomeDB (Fuentes et al. 2022) is another popular tool for phylogenetic studies spanning the tree of life. It contains several collections or phylomes focused on different species and provides information on them such as gene trees, domains and MSAs. However, in general it supports a low number of plant species, as can be seen in Table 2 for its largest plant phylome. PhylomeDB also

exhibits the aforementioned limitations in terms of representatives outside the angiosperm group. Thus, despite incorporating *O. tauri* into the collection, it does not recover its CCA1/LHY ortholog, nor any outside the angiosperms clade (Sup. Fig. S12D). Furthermore, this tool is focused on phylogenetic analysis of many species but does not incorporate functional information or evolutionary history reconstruction for them.

With this comparison, it becomes clear that PharaohFUN supports a significantly higher number of plant species than the other tools. The difference is dramatic in the case of non-angiosperm land plants and streptophyte and chlorophyte algae, while maintaining an intensive sampling of angiosperms without compromising the balance of clades, thanks to the creation of a phylodiverse model to which subsequently assign the rest of the genes. PharaohFUN also stands out for its integration of abundant functional and domain information with interactive exploration of MSAs and its flexibility regarding search modes, allowing large-scale evolutionary studies or the inclusion of sequences from species absent in the dataset, features not offered by most tools. Finally, it is the only one that offers information on physical interactions for studies of coevolution between different proteins, literature-based annotation and reconstructions of the evolutionary history of orthogroups along the species tree, in terms of ancestral states at internal nodes and mapping of expansion/contraction/gain/loss events on its branches.

Conclusions and future directions

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 and non-seed land plants. 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, it

provides 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 include new representative species from different clades dealing with the rapid expansion in the number of genomes available. PharaohFUN's modular design allows for seamless incorporation and analysis of new species, 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.

Acknowledgments

We would like to acknowledge the revision and testing of the tool carried out by several members of the "Spanish National Research Network on Evolution of Regulatory Mechanisms in Development and Plant Signalling" which include but are not limited to Pilar Cubas, Miguel Ángel Blázquez, Javier Fuentes and Lluís García.

Author Contributions

FJRC and MGG conceived and conceptualized the project. FJRC, MGG, MRG and JHG designed the structure of the tool. MRG, VRG, ESP and FJRC developed the web app. MRG, VRG, ESP, CA, FJRC and JHG carried out the collection of genomes and annotation of photosynthetic eukaryotes included in the tool. MRG, FJRC and MGG performed all the analysis and interpreted the results. MRG, FJRC and MGG wrote the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by grants PID2021-123984OB-I00 (ELECTRA) and TED2021-129651B-I00 (RESILIENCE) from the Spanish Ministry of Science and Innovation. MRG was supported by an FPU predoctoral grant FPU22/00511 from the Spanish Ministry of Science and Innovation. VRG was supported by a predoctoral grant Fundación Ramón Areces 2022. ESP was supported by project MOMENTUM MMT24-IBVF-01. CA was supported by Junta de Andalucia predoctoral grant PREDOC_00999.

Data Availability

The complete source code for PharaohFUN is freely available at the following GitHub repository: https://github.com/ramosgonzmarc/PharaohFUN

Conflict of interest

The authors declare no conflict of interest.

References

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Mol Biol Evol* 31:2553–2556. https://doi.org/10.1093/molbev/msu236

Andolfo G, Di Donato A, Chiaiese P, De Natale A, Pollio A, Jones JDG, Frusciante L, Ercolano MR. 2019. Alien Domains Shaped the Modular Structure of Plant NLR Proteins. *Genome Biol Evol* 11:3466–3477. https://doi.org/10.1093/gbe/evz248

Arita M, Karsch-Mizrachi I, Cochrane G, Collaboration on behalf of the INSD. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Res* 49:D121–D124. https://doi.org/10.1093/nar/gkaa967

Bian Y, Song Zhuolong, Liu C, Song Zhaoqing, Dong J, Xu D. 2025. The BBX7/8-CCA1/LHY transcription factor cascade promotes shade avoidance by activating PIF4. *New Phytologist* 245:637–652. https://doi.org/10.1111/nph.20256

Bierenbroodspot MJ, Pröschold T, Fürst-Jansen JMR, de Vries S, Irisarri I, Darienko T, de Vries J. 2024. Phylogeny and evolution of streptophyte algae. *Ann Bot* 134:385–400. https://doi.org/10.1093/aob/mcae091

Bowles AMC, Williams TA, Donoghue PCJ, Campbell DA, Williamson CJ. 2024. Metagenome-assembled genome of the glacier alga *Ancylonema* yields insights into the evolution of streptophyte life on ice and land. *New Phytologist*. https://doi.org/10.1111/nph.19860

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60. https://doi.org/10.1038/nmeth.3176

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 38:5825–5829. https://doi.org/10.1093/molbev/msab293

Carré IA, Kim J. 2002. MYB transcription factors in the *Arabidopsis* circadian clock. *J Exp Bot* 53:1551–1557. https://doi.org/10.1093/jxb/erf027

Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, et al. 2019. Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* 179:1057-1067.e14. https://www.sciencedirect.com/science/article/pii/S0092867419311699

Clark JW. 2023. Genome evolution in plants and the origins of innovation. *New Phytologist* 240:2204–2209. https://doi.org/10.1111/nph.19242

Csűös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26:1910–1912. https://doi.org/10.1093/bioinformatics/btq315

Cui J, Zhu Y, Du H, Liu Z, Shen S, Wang T, Cui W, Zhang R, Jiang S, Wu Y, et al. 2023. Chromosome-level reference genome of tetraploid *Isoetes sinensis* provides insights into evolution and adaption of lycophytes. *Gigascience* 12:giad079. https://doi.org/10.1093/gigascience/giad079

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238. https://doi.org/10.1186/s13059-019-1832-y

Emms DM, Kelly S. 2022. SHOOT: phylogenetic gene search and ortholog inference. *Genome Biol* 23:85. https://doi.org/10.1186/s13059-022-02652-8

Emms DM, Liu Y, Belcher L, Holmes J, Kelly S. 2025. OrthoFinder: scalable phylogenetic orthology inference for comparative genomics. *bioRxiv* 2025.07.15.664860. http://biorxiv.org/content/early/2025/07/16/2025.07.15.664860.abstract

Fuentes D, Molina M, Chorostecki U, Capella-Gutiérrez S, Marcet-Houben M, Gabaldón T. 2022. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res* 50:D1062–D1068. https://doi.org/10.1093/nar/gkab966

Gàlvez-Morante A, Guéguen L, Natsidis P, Telford MJ, Richter DJ. 2024. Dollo Parsimony Overestimates Ancestral Gene Content Reconstructions. *Genome Biol Evol* 16:evae062. https://doi.org/10.1093/gbe/evae062

Gil S, Feord HK, van Ooijen G. 2023. Homologs of Ancestral CNNM Proteins Affect Magnesium Homeostasis and Circadian Rhythmicity in a Model Eukaryotic Cell. *Int J Mol Sci* 24. https://www.mdpi.com/1422-0067/24/3/2273

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grigoriev I V, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, Dusheyko S, Nikitin R, Mondo SJ, Salamov A, et al. 2021. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res* 49:D1004–D1011. https://doi.org/10.1093/nar/gkaa898

Gyaltshen Y, Rozenberg A, Paasch A, Burns JA, Warring S, Larson RT, Maurer-Alcalá XX, Dacks J, Narechania A, Kim E. 2023. Long-Read–Based Genome Assembly Reveals Numerous Endogenous Viral Elements in the Green Algal Bacterivore *Cymbomonas tetramitiformis*. *Genome Biol Evol* 15:evad194. https://doi.org/10.1093/gbe/evad194

Hanschen ER, Starkenburg SR. 2020. The state of algal genome quality and diversity. *Algal Res* 50:101968. https://www.sciencedirect.com/science/article/pii/S2211926420301016

Hernández-García J, Carrillo-Carrasco VP, Rienstra J, Tanaka K, de Roij M, Dipp-Álvarez M, Freire-Ríos A, Crespo I, Boer R, van den Berg WAM, et al. 2024. Evolutionary origins and functional diversification of Auxin Response Factors. *Nat Commun* 15:10909. https://doi.org/10.1038/s41467-024-55278-8

Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, Kirsch R, Rattei T, Letunic I, Jensen LJ, et al. 2023. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 51:D389–D394. https://doi.org/10.1093/nar/gkac1022

Huang S, Gao Y, Liu J, Peng X, Niu X, Fei Z, Cao S, Liu Y. 2012. Genome-wide analysis of WRKY transcription factors in *Solanum lycopersicum*. *Molecular Genetics and Genomics* 287:495–513. https://doi.org/10.1007/s00438-012-0696-6

Huang X, Wang W, Gong T, Wickell D, Kuo L-Y, Zhang X, Wen J, Kim H, Lu F, Zhao H, et al. 2022. The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat Plants* 8:500–512. https://doi.org/10.1038/s41477-022-01146-6

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically

annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47:D309–D314. https://doi.org/10.1093/nar/gky1085

Hung F-Y, Chen F-F, Li C, Chen C, Lai Y-C, Chen J-H, Cui Y, Wu K. 2018. The Arabidopsis LDL1/2-HDA6 histone modification complex is functionally associated with CCA1/LHY in regulation of circadian clock genes. *Nucleic Acids Res* 46:10669–10681. https://doi.org/10.1093/nar/gky749

Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, Gao G. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45:D1040–D1045. https://doi.org/10.1093/nar/gkw982

Kalde M, Barth M, Somssich IE, Lippok B. 2003. Members of the *Arabidopsis* WRKY Group III Transcription Factors Are Part of Different Plant Defense Signaling Pathways. *Molecular Plant-Microbe Interactions* 16:295–305. https://doi.org/10.1094/MPMI.2003.16.4.295

Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, Morales AE, Ahmed A-W, Kontopoulos D-G, Hilgers L, et al. 2024. Integrating gene annotation with orthology inference at scale. *Science* 380:eabn3107. https://doi.org/10.1126/science.abn3107

Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, Liu X, Soltis PS. 2022. Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc Natl Acad Sci U S A* 119:1–9.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* 39:msac174. https://doi.org/10.1093/molbev/msac174

Lai X, Chahtane H, Martin-Arevalillo R, Zubieta C, Parcy F. 2020. Contrasted evolutionary trajectories of plant transcription factors. *Curr Opin Plant Biol* 54:101–107. https://www.sciencedirect.com/science/article/pii/S1369526620300327

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–685. https://doi.org/10.1038/s41586-019-1693-2

Li X, Cao X, Li J, Niu Q, Mo Y, Xiao L. 2022. Genome-wide characterization of C2H2 zinc-finger gene family provides insight into the mechanisms and evolution of the dehydration–rehydration responses in *Physcomitrium* and *Arabidopsis*. *Front Plant Sci* Volume 13-2022. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.953459

Liang Z, Geng Y, Ji C, Du H, Wong CE, Zhang Q, Zhang Y, Zhang Pingxian, Riaz A, Chachar S, et al. 2020. *Mesostigma viride* Genome and Transcriptome Provide Insights into the Origin and Evolution of Streptophyta. *Advanced Science* 7:1901850. https://doi.org/10.1002/advs.201901850

Lim SC, Fo K, Sunil RS, Itharajula M, Chuah YS, Foo H, Davey EE, Fullwood M, Thibault G, Mutwil M. 2024. PlantConnectome: knowledge graph encompassing >70,000 plant articles. *bioRxiv* 2023.07.11.548541. http://biorxiv.org/content/early/2024/09/26/2023.07.11.548541.abstract

Linde A-M, Eklund DM, Kubota A, Pederson ERA, Holm K, Gyllenstrand N, Nishihama R, Cronberg N, Muranaka T, Oyama T, et al. 2017. Early evolution of the land plant circadian clock. *New Phytologist* 216:576–590. https://doi.org/10.1111/nph.14487

Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR. 2012. Preferential Retention of Circadian Clock Genes during Diploidization following Whole Genome Triplication in *Brassica rapa*. *Plant Cell* 24:2415–2426. https://doi.org/10.1105/tpc.112.099499

Ma J, Wang S, Zhu X, Sun G, Chang G, Li L, Hu X, Zhang S, Zhou Y, Song C-P, et al. 2022. Major episodes of horizontal gene transfer drove the evolution of land plants. *Mol Plant* 15:857–871. https://www.sciencedirect.com/science/article/pii/S1674205222000491

Marks RA, Amézquita EJ, Percival S, Rougon-Cardoso A, Chibici-Revneanu C, Tebele SM, Farrant JM, Chitwood DH, VanBuren R. 2023. A critical analysis of plant science literature reveals ongoing inequities. *Proceedings of the National Academy of Sciences* 120:e2217564120. https://doi.org/10.1073/pnas.2217564120

Martignago D, Bernardini B, Polticelli F, Salvi D, Cona A, Angelini R, Tavladoraki P. 2019. The Four FAD-Dependent Histone Demethylases of *Arabidopsis* Are Differently Involved in the Control of Flowering Time. *Front Plant Sci* 10. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.00669

Matthews DS, Spence MA, Mater AC, Nichols J, Pulsford SB, Sandhu M, Kaczmarski JA, Miton CM, Tokuriki N, Jackson CJ. 2023. Leveraging ancestral sequence reconstruction for protein representation learning. *bioRxiv* 2023.12.20.572683. http://biorxiv.org/content/early/2023/12/21/2023.12.20.572683.abstract

Mecchia MA, García-Hourquet M, Lozano-Elena F, Planas-Riverola A, Blasco-Escamez D, Marquès-Bueno M, Mora-García S, Caño-Delgado AI. 2021. The BES1/BZR1-family transcription factor MpBES1 regulates cell division and differentiation in *Marchantia polymorpha*. *Current Biology* 31:4860-4869.e8. https://www.sciencedirect.com/science/article/pii/S0960982221011866

Mohanta TK, Park Y-H, Bae H. 2016. Novel Genomic and Evolutionary Insight of WRKY Transcription Factors in Plant Lineage. *Sci Rep* 6:37309. https://doi.org/10.1038/srep37309

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35:W182–W185. https://doi.org/10.1093/nar/gkm321

Okada R, Kondo S, Satbhai SB, Yamaguchi N, Tsukuda M, Aoki S. 2009. Functional characterization of CCA1/LHY homolog genes, PpCCA1a and PpCCA1b, in the moss *Physcomitrella patens*. *The Plant Journal* 60:551–563. https://doi.org/10.1111/j.1365-313X.2009.03979.x

Petroll R, Varshney D, Hiltemann S, Finke H, Schreiber M, de Vries J, Rensing SA. 2025. Enhanced sensitivity of TAPscan v4 enables comprehensive analysis of streptophyte transcription factor evolution. *The Plant Journal* 121:e17184. https://doi.org/10.1111/tpj.17184

Qiao X, Zhang S, Paterson AH. 2022. Pervasive genome duplications across the plant tree of life and their links to major evolutionary innovations and transitions. *Comput Struct Biotechnol J* 20:3248–3256. https://www.sciencedirect.com/science/article/pii/S2001037022002410

Rensing SA. 2020. How Plants Conquered Land. *Cell* 181:964–966. https://www.sciencedirect.com/science/article/pii/S0092867420305730

Rokas A, Carroll SB. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. *Mol Biol Evol* 22:1337–1344. https://doi.org/10.1093/molbev/msi121

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593. https://doi.org/10.1093/bioinformatics/btq706

Serrano-Pérez E, Romero-Losada AB, Morales-Pineda M, García-Gómez ME, Couso I, García-González M, Romero-Campero FJ. 2022. Transcriptomic and Metabolomic Response to High Light in the Charophyte Alga *Klebsormidium nitens*. *Front Plant Sci* 13. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.855243

Shao Z-Q, Xue J-Y, Wang Q, Wang B, Chen J-Q. 2019. Revisiting the Origin of Plant NBS-LRR Genes. *Trends Plant Sci* 24:9–12. https://doi.org/10.1016/j.tplants.2018.10.015

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. https://doi.org/10.1038/msb.2011.75

Sillitoe I, Furnham N. 2016. FunTree: advances in a resource for exploring and contextualising protein function evolution. *Nucleic Acids Res* 44:D317–D323. https://doi.org/10.1093/nar/gkv1274

Sjölander K, Datta RS, Shen Y, Shoffner GM. 2011. Ortholog identification in the presence of domain architecture rearrangement. *Brief Bioinform* 12:413–422. https://doi.org/10.1093/bib/bbr036

Strassert JFH, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat Commun* 12:1879. https://doi.org/10.1038/s41467-021-22044-z

Stull GW, Qu X-J, Parins-Fukuchi C, Yang Y-Y, Yang J-B, Yang Z-Y, Hu Y, Ma H, Soltis PS, Soltis DE, et al. 2021. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat Plants* 7:1015–1025. https://doi.org/10.1038/s41477-021-00964-4

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47:D607–D613. https://doi.org/10.1093/nar/gky1131

Tarr DEK, Alexander HM. 2009. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res Notes* 2:197. https://doi.org/10.1186/1756-0500-2-197

Tamilsena PR, Wafula EK, Barrett CF, Ayyampalayam S, McNeal JR, Rentsch JD, McKain MR, Heyduk K, Harkess A, Villegente M, et al. 2022. Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. *Front Plant Sci* 13. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.876779

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K. 2018. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res* 46:D1190–D1196. https://doi.org/10.1093/nar/gkx1002

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

Vandepoele K, Van Bel M, Richard G, Van Landeghem S, Verhelst B, Moreau H, Van de Peer Y, Grimsley N, Piganeau G. 2013. pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ Microbiol* 15:2147–2153. https://doi.org/10.1111/1462-2920.12174

Villacastin AJ, Adams KS, Boonjue R, Rushton PJ, Han M, Shen JQ. 2021. Dynamic differential evolution schemes of WRKY transcription factors in domesticated and wild rice. *Sci Rep* 11:14887. https://doi.org/10.1038/s41598-021-94109-4

de Vries J, Curtis BA, Gould SB, Archibald JM. 2018. Embryophyte stress signaling evolved in the algal progenitors of land plants. *Proceedings of the National Academy of Sciences* 115:E3471–E3480. https://doi.org/10.1073/pnas.1719230115

Wang J-L, Wang H-W, Cao Y-N, Kan S-L, Liu Y-Y. 2022. Comprehensive evolutionary analysis of the TCP gene family: Further insights for its origin, expansion, and diversification. *Front Plant Sci* 13. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.994567

Wang S, Sun Q, Zhang M, Yin C, Ni M. 2022. WRKY2 and WRKY10 regulate the circadian expression of PIF4 during the day through interactions with CCA1/LHY and phyB. *Plant Commun* 3:100265. https://www.sciencedirect.com/science/article/pii/S2590346221001796

Wang Y, Feng L, Zhu Y, Li Y, Yan H, Xiang Y. 2015. Comparative genomic analysis of the WRKY III gene family in populus, grape, arabidopsis and rice. *Biol Direct* 10:48. https://doi.org/10.1186/s13062-015-0076-3

Wickell DA, Li F-W. 2020. On the evolutionary significance of horizontal gene transfers in plants. *New Phytologist* 225:113–117. https://doi.org/10.1111/nph.16022

Wilhelmsson PKI, Mühlich C, Ullrich KK, Rensing SA. 2017. Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. *Genome Biol Evol* 9:3384–3397. https://doi.org/10.1093/gbe/evx258

Yeom M, Kim H, Lim J, Shin A-Y, Hong S, Kim J-I, Nam HG. 2014. How Do Phytochromes Transmit the Light Quality Information to the Circadian Clock in *Arabidopsis*? *Mol Plant* 7:1701–1704. https://doi.org/10.1093/mp/ssu086

Yu G. 2020. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics* 69:e96. https://doi.org/10.1002/cpbi.96

Figures

Figure 1. PharaohFUN's model creation and composition. **A.** Analysis pipeline used for the generation of all input data used by PharaohFUN. Files derived from different software workflows are indicated through colored arrows. **B.** Phylodiverse species tree. **Left**: ultrameric species tree. 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. PharaohFUN architecture. The tool is divided into three main modules, with five different search modes. Each search mode is associated with a different colored arrow, and the common steps are represented in black. Color boxes stand for different types of objects, as determined in the legend. The asterisk marks that the elements are equivalent, and all outputs derived from "Gene tree" are also produced from "Gene tree with new sequence".

Figure 3. Overview of results for *Helianthus annus* putative CCA1 evolution using PharaohFUN. **A.** Gene tree for *Helianthus annus* putative CCA1 orthogroup using *A. thaliana, B. rapa, A. hypocondriacus, M. polymorpha, O. tauri* and *P. patens* as reference species. The query gene is highlighted in red. **B.** Size evolution of orthogroup through species tree. **C.** PFAM Domains location in proteins corresponding to CCA1/LCL clade. **D.** MSA in the vicinity of the conserved SHAQKYF domain, highlighted by a black box, for the protein sequences in CCA1/LCL clade plus two additional *A. thaliana* proteins from the same orthogroup.

Figure 4. Overview of results for CCA1-clade functional annotation using PharaohFUN. **A.** Treeplot showing clustered GO terms for biological processes. **B.** KEGG pathway for Circadian Rhythms in Plants, highlighting genes in the clade (CCA1 and LHY) in red. **C.** Example of some rows of literature annotation result for "CCA1" using exact search. **D.** STRING interaction network of AT2G46830 gene. Red arrows indicate genes from orthogroups that appear in the first rows of conserved interactions results.

Figure 5. Results for the different OGs associated to BES1 family. **A.** Gene trees. **B**. PFAM domains location. Panels: **Up** OG corresponding to BES1/BZR1 TFs, **Down** OG corresponding to beta amylases.

Figure 6. Evolution of the TF families from *A. thaliana* in Viridiplantae. **A.** Heatmap showing the normalized gene number per TFs family in 54 species supported by PharaohFUN. TF families are ordered based on their abundance pattern similarity through the different organisms. **B.** Species tree with the number of gene acquisition events mapped onto its internal nodes. Up: OG gains. Down: OG expansions. **D.** Raincloud plot showing the distribution of sensitivity (recovery rate) in the identification of genes of each TF family in each test species using *A. thaliana* genes as a source. **E.** Raincloud plot showing the distribution of sensitivity in the identification of genes of each TF family in each test species using *A. thaliana*, *B. prasinos*, *V. carteri*, *P. patens* and *S. bicolor* genes as source. The raincloud plot depicts a typical boxplot along with the individual jittered measure points and their density is represented at the side of the boxplot.

Sup. Figure S1. PharaohFUN's complete species tree. Dated species tree used for model construction. **Left**: dated species tree measured in millions of years. 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.

Sup. Figure S2. Gene tree corresponding to the complete CCA1 OG. Three differentiated clades are highlighted.

Sup. Figure S3. PFAM domains location in sequences from the complete CCA1 OG.

Sup. Figure S4. Results for the WRKY family. **A.** WRKY OG tree. Notation of groups I and II followed that of (Huang et al. 2012) and notation of group III that of (Wang et al. 2015). **B.** Evolutionary events in Viridiplantae for the OG comprising TIR-NB-LRR proteins.

Sup. Figure S5. OGs containing C2H2 genes in *A. thaliana* mapped onto the AtZFPs tree adapted from (Li et al. 2022).

Sup. Figure S6. Heatmap showing the row-normalized gene number per OG associated to TF families in 54 Viridiplantae species.

Sup. Figure S7. Species tree with the number of OGs associated to TF families in *A. thaliana* exhibiting its MRCA in each of its internal nodes. Color and point size determine total number.

Sup. Figure S8. Species tree with the number of gained OGs associated to TF families in *A. thaliana* in each of its internal and terminal nodes. Color and point size determine total number.

Sup. Figure S9. Species tree with the number of expanded OGs associated to TF families in *A. thaliana* in each of its internal and terminal nodes. Color and point size determine total number.

Sup. Figure S10. Species tree with the number of contracted OGs associated to TF families in *A. thaliana* in each of its internal and terminal nodes. Color and point size determine total number.

Sup. Figure S11. Species tree with the number of lossed OGs associated to TF families in *A. thaliana* in each of its internal and terminal nodes. Color and point size determine total number.

Sup. Figure S12. Results for CCA1 rendered by **A.** SHOOT, **B.** pico-PLAZA 3.0, **C.** eggNOG-DB 6.0 and **D.** PhylomeDB.

Tables

Table 1. TF families associated with OGs that emerged or underwent expansion in different ancestors of *A. thaliana.*

Table 2. Comparison between phylogenetic web-based tools. *****FunTree uses taxonomy data to guide reconciliation of trees, but species tree is not explicitly shown. ** Numbers correspond to the largest plant phylome. *** eggNOG-DB 6.0 does not incorporate a batch mode, but fully automatic functional annotation of sequences can be performed using eggNOG-mapper v2 on eggNOG-DB 5.0.

Sup. Table S1. Genomic data used in PharaohFUN, including species names, genome version and genome source.

Sup. Table S2. Constraints used for dating the species tree in Mya. Restrictions are indicated by minimum and maximum time points for different ancestors of the clades, using fossil data from (Strassert et al. 2021) and data from (Kumar et al. 2022).

Sup. Table S3. ExaBayes parameter file.

Sup. Table S4. Number of genes associated to each OG of those interacting with CCA1 orthologs in *A. thaliana*, *P. patens* and *O. tauri*.

Sup. Table S5. Assignment of OGs by TF family, their common ancestor and the number of genes in each species according to PharaohFUN results.

Sup. Table S6. Number and type of PFAM domains of *A. thaliana* genes in each TF-related OG according to PharaohFUN results.

Sup. Table S7. Evolutionary events occurred in each node from the species tree for TF-related OGs.

Sup. Table S8. Sensitivities per TF family in each test species using *A. thaliana* genes as source.

Sup. Table S9. Sensitivities per TF family in each test species using *A. thaliana*, *B. prasinos*, *V. carteri*, *P. patens* and *S. bicolor* genes as source.

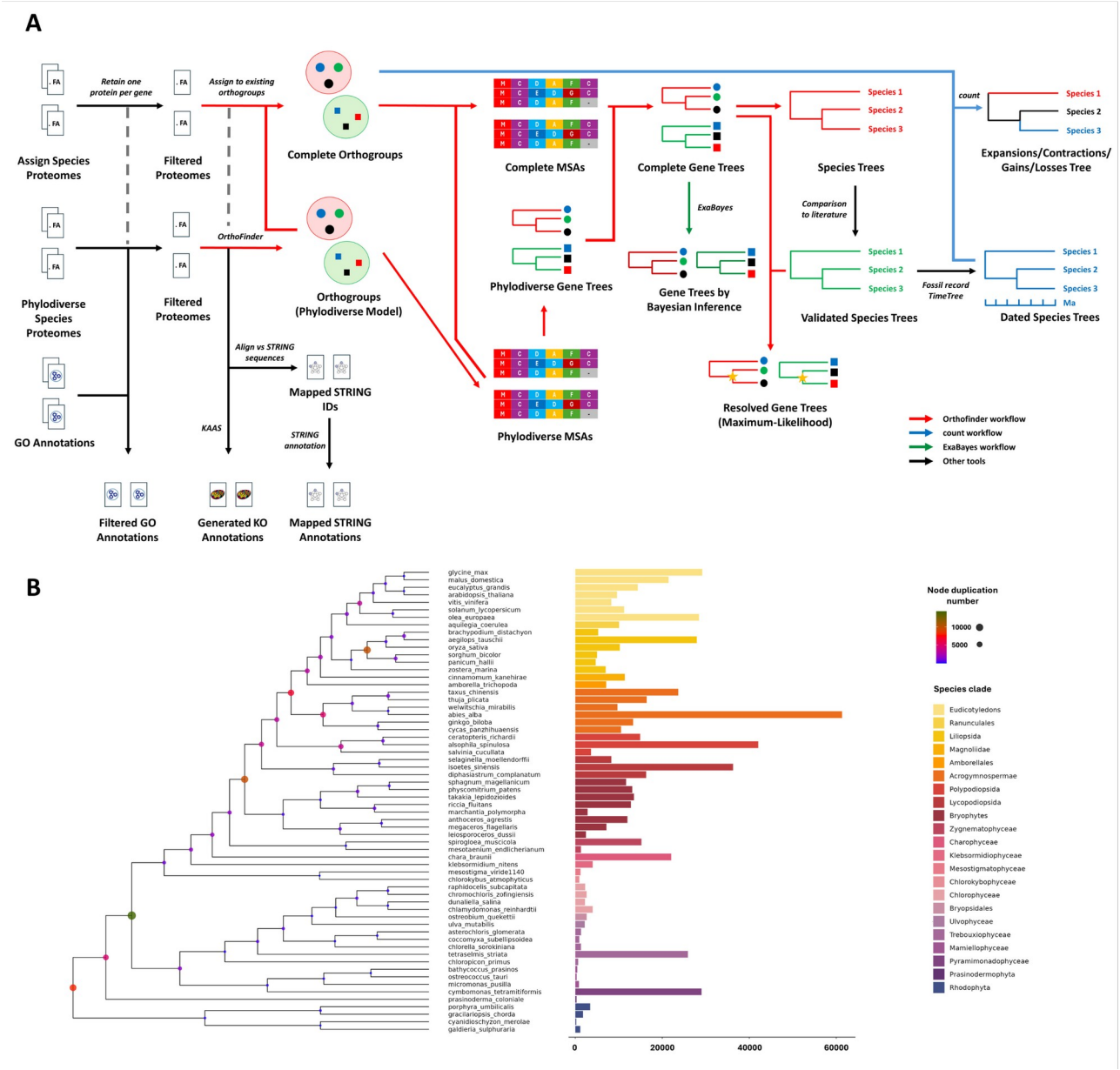


Figure 1. PharaohFUN's models creation and composition.

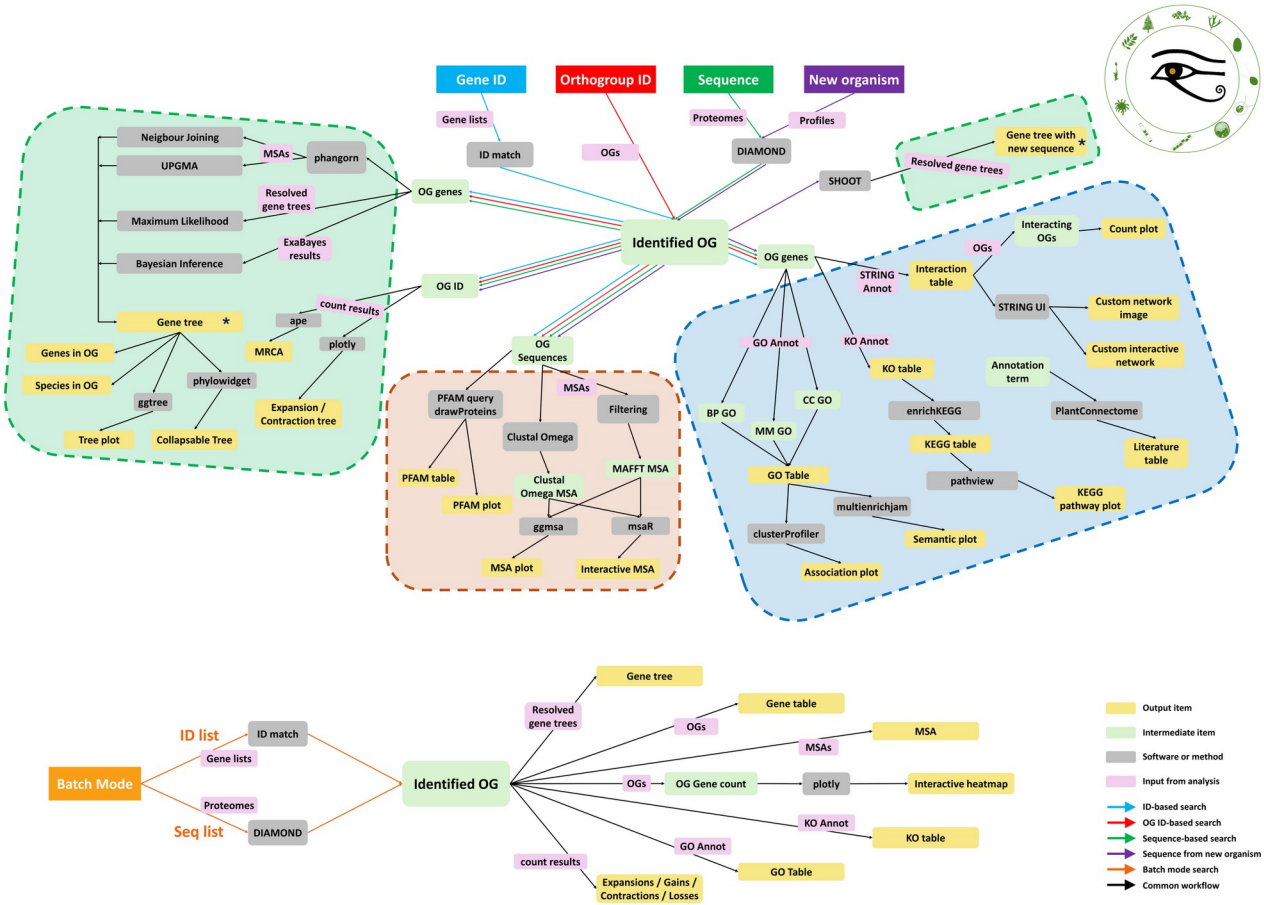


Figure 2. PharaohFUN architecture.

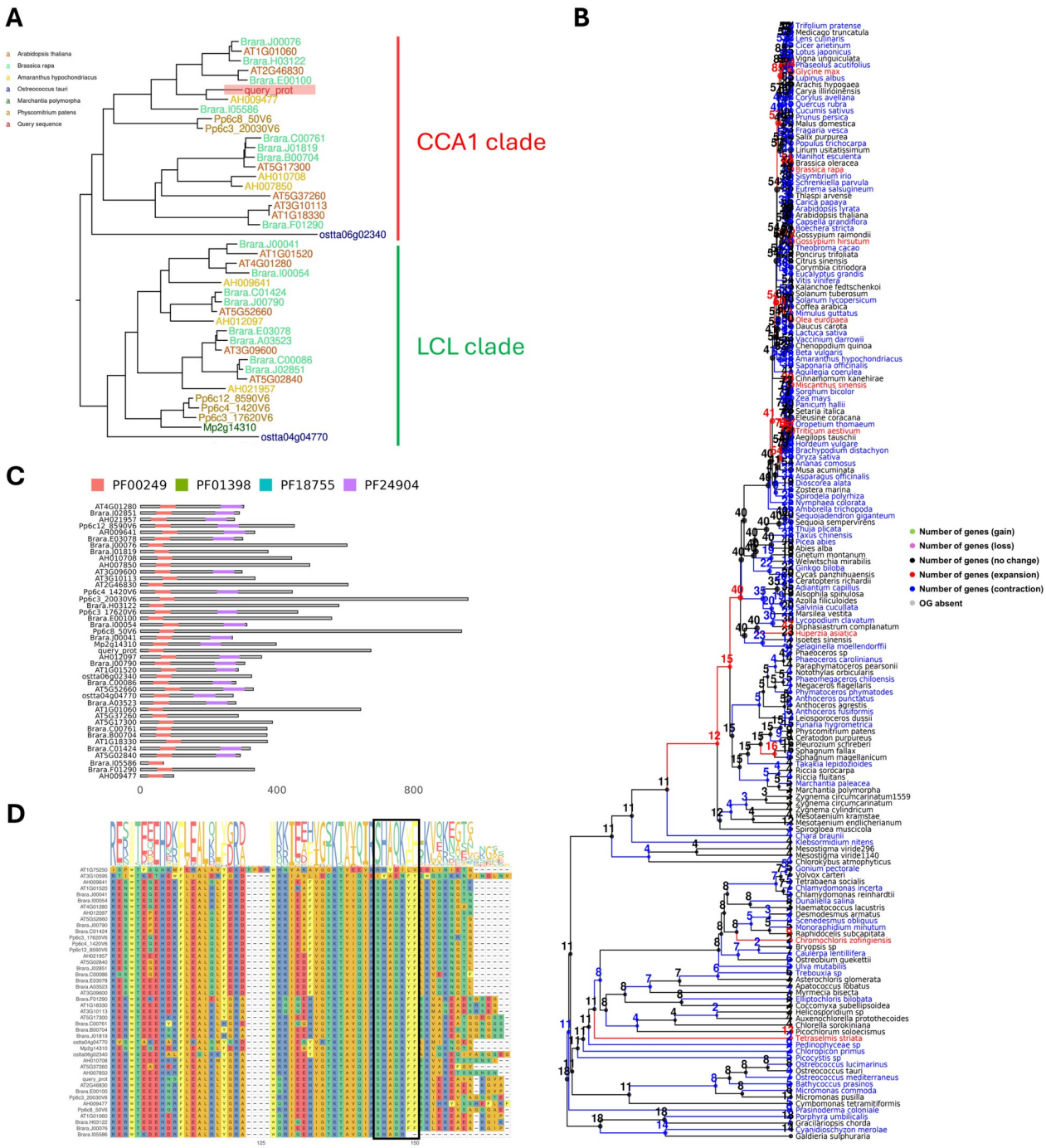


Figure 3. Overview of results for *Helianthus annus* putative CCA1 evolution using PharaohFUN.

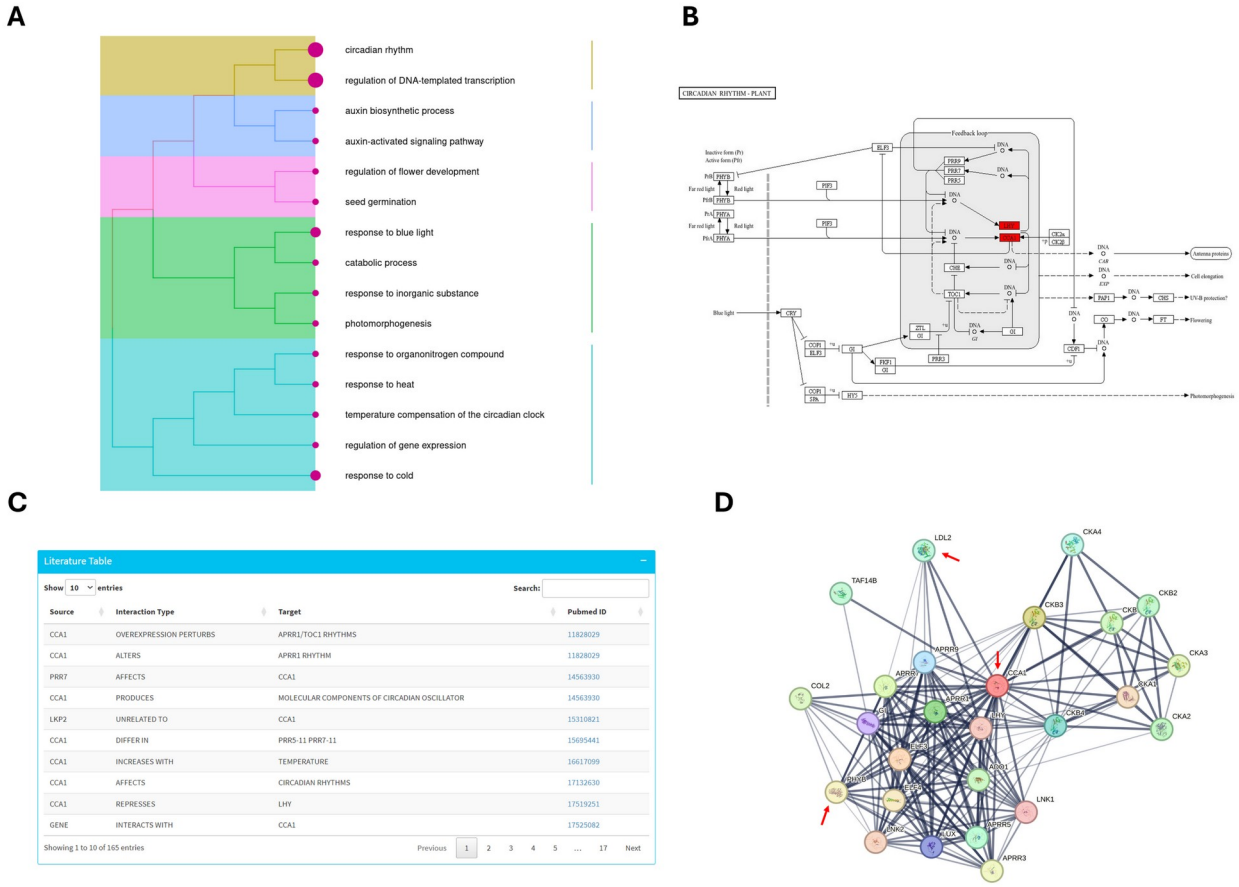


Figure 4. Overview of results for CCA1-clade functional annotation using PharaohFUN.

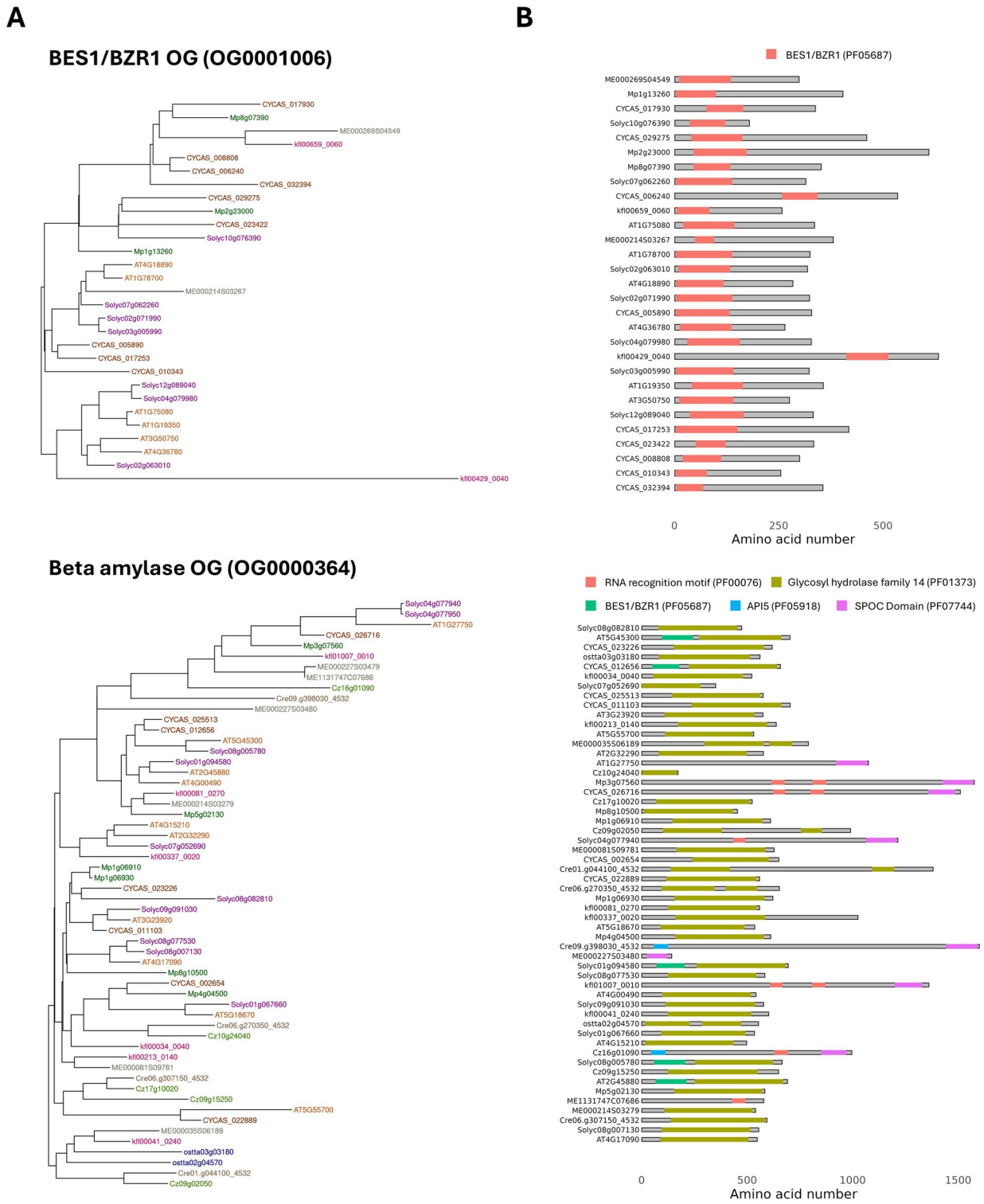


Figure 5. Results for the different OGs associated to BES1 family.

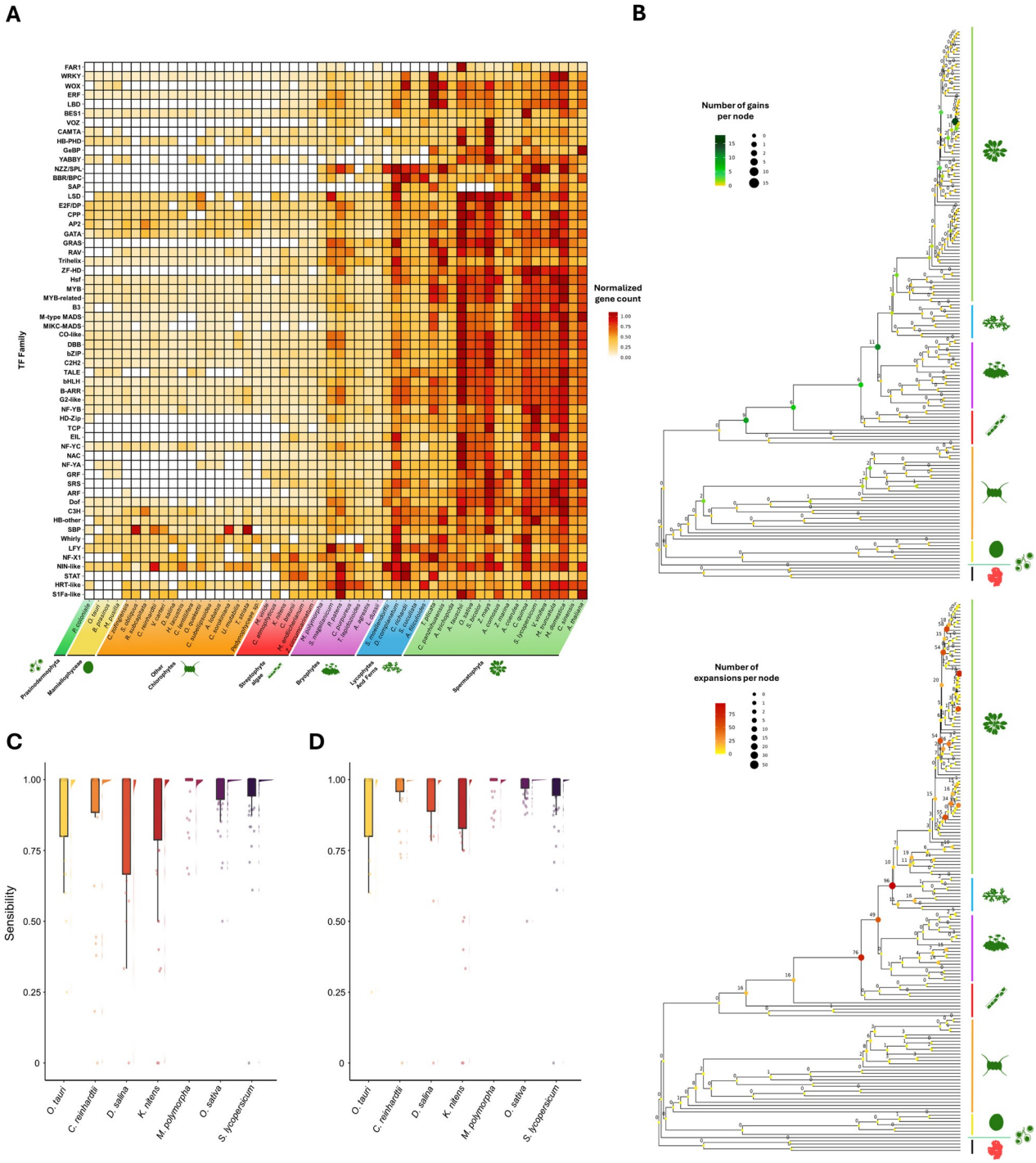


Figure 6. Evolution of the TF families of *A. thaliana* in Viridiplantae.

Table 1. TF families associated to the OGs that emerged in different ancestors of *A. thaliana.*

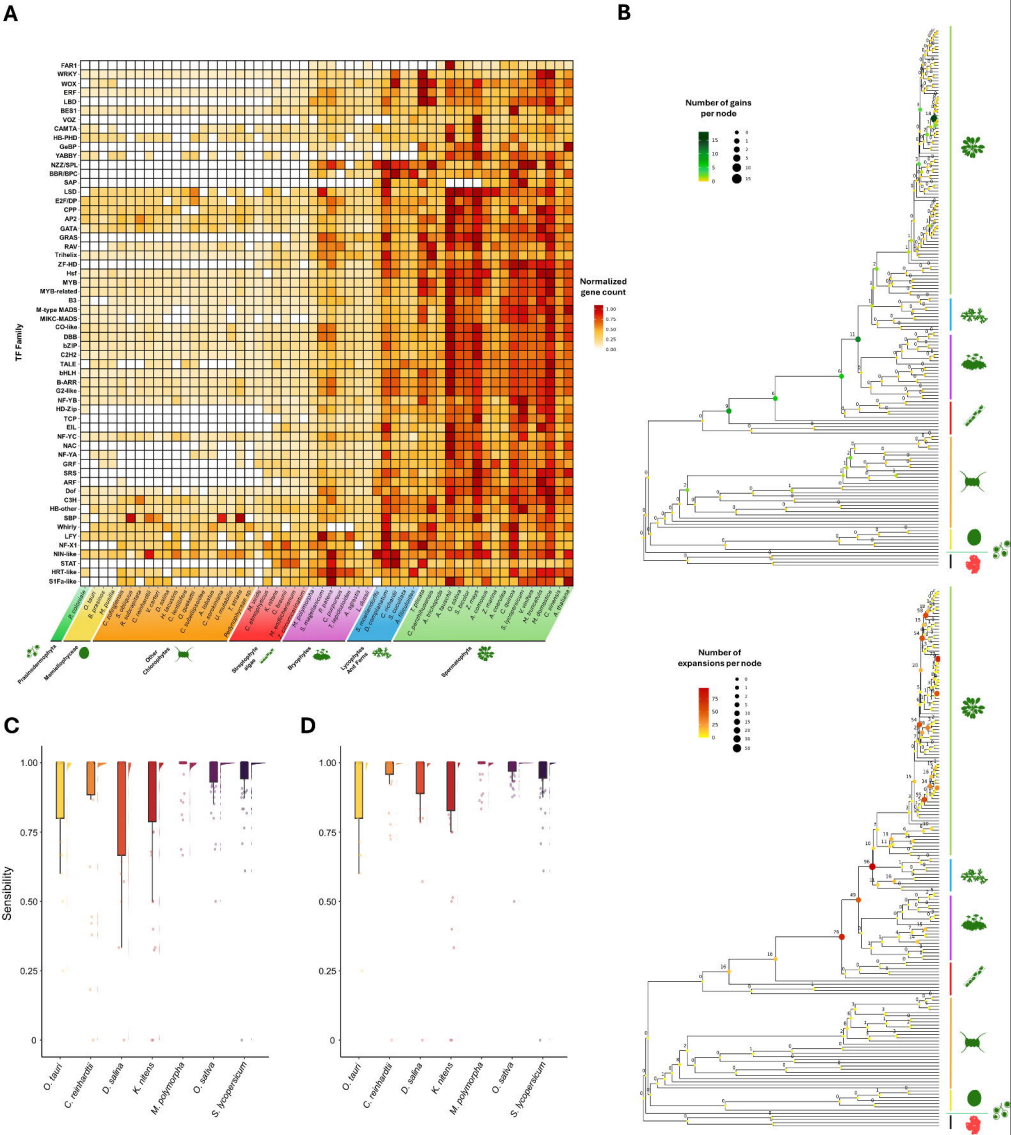
Node	TF family emergences
Root (Archaeplastida)	ARR-B / bHLH / bZIP / C2H2 / C3H / CO-like / CPP / DBB / E2F_DP / G2-like / GATA / HB-other / HSF / M-type MADS / MYB / MYB-related / NF-X1 / NF-YA / NF-YB / NF-YC / NIN-like / TALE / WRKY / YABBY
Viridiplantae	AP2 / B3 / Dof / ERF / SBP / Trihelix
Chlorophyta + Streptophyta	CAMTA / HD-zip / S1Fa-like / Whirly / WOX
Streptophyta	ARF / GRF / MICK MADS / RAV / SRS / ZF-HD
Phragmoplastophyta + Klebsormidiophyceae	BES1 / EIL / LBD / LFY / LSD / NAC / STAT
Phragmoplastophyta	HRT-like / TCP
Anydrophyta	BBR-BPC / GRAS / VOZ
Embryophyta	FAR1 / GeBP / NZZ_SPL
Tracheophyta	SAP

Table 2. Comparison between phylogenetic web-based tools. *****FunTree uses taxonomy data to guide reconciliation of trees, but species tree is not explicitly shown. ** Numbers correspond to the largest plant phylome. *** eggNOG-DB 6.0 does not incorporate a batch mode, but fully automatic functional annotation of sequences can be performed using eggNOG-mapper v2 on eggNOG-DB 5.0.

	PharaohFUN	**SHOOT**	**FunTree**	**Pico-PLAZA 3.0**	**eggNOG-DB 6.0**	**PhylomeDB V5**
Supported Organisms						
Angiosperms	75	17	*	3	102	14**
Non-angiosperms land plants	43	8	*	1	2	2**
Streptophyte algae	11	1	*	0	1	0**
Mamiellophyceae/ Piramimo	7	2	*	7	5	3**
Other Chlorophyte algae	32	2	*	9	12	1**
Prasinodermophytes	1	0	*	0	0	0**
Supported Features						
Gene Trees	Yes	Yes	No	Yes	Yes	Yes
Integrated Domains Info	Yes	No	Yes	Yes	Yes	Yes
Individual Domains Info	Yes	No	Yes	Yes	Yes	No
Ancestral state reconstructio	Yes	No	No	No	No	No

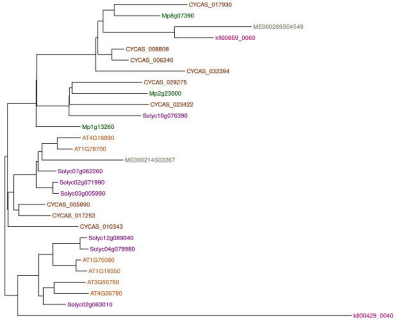
n of OG sizes						
Evolutionary events (Expansions / Contractions)	Yes	No	No	No	No	No
Integrated functional annotation for all genes	Yes	No	Shows complete record but gene correspondence is left to manual exploration	Shows complete record but gene correspondence is left to manual exploration	Yes	No
Individual functional annotation	Yes	No	Yes	Yes	Yes	No
Interactive exploration of alignments	Yes	No	Yes	Yes	No	Yes
Genomic positions	No	No	No	Yes	No	No
Include non-enzyme genes	Yes	Yes	No	Yes	Yes	Yes
Tree building methods	ML, distance-based, bayesian inference	ML	ML	ML	ML	ML
Annotation						
Type of Functional Annotation	GO, KEGG (including EC)	None	GO, EC, MACiE	GO, MapMan (including EC)	GO, KEGG (including EC), BiGG reaction, CAZy	None
Literature-based biological annotation	Yes	No	No	No	No	No
Interaction annotation	Yes	No	No	No	No	No
Input Types						
ID-based search	Yes	No	Yes	Yes	Yes	Yes

Sequence-based search	Yes	Yes	Beta	Only for registered users	Yes	Yes
OG-based search (or equivalent)	Yes	No	No	Yes	Yes	Yes
Batch search	Yes	No	No	Only for registered users	No***	No
Search by sequence from external species	Yes	Yes	No	No	No	No

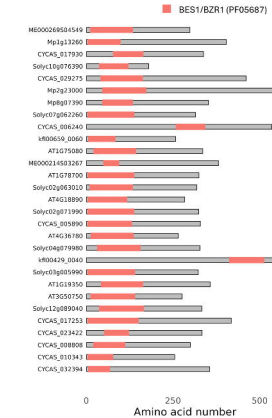

A
TF Family
FAR1
WRKY
WOX
ERF
LBD
BES1
VOZ
CAMTA
HB-PHD
GeBP
YABBY
NZZ/SPL
BBR/BPC
SAP
LSD
E2F/DP
CPP
AP2
GATA
GRAS
RAV
Trihelix
ZF-HD
Hsf
MYB
MYB-related
B3
M-type MADS
MIKC-MADS
CO-like
DBB
bZIP
C2H2
TALE
bHLH
B-ARR
G2-like
NF-YB
HD-Zip
TCP
EIL
NF-YC
NAC
NF-YA
GRF
SRS
ARF
Dof
C3H
HB-other
SBP
Whirly
LFY
NF-X1
NIN-like
STAT
HRT-like
S1Fa-like
Normalized gene count
1.00
0.75
0.50
0.25
0.00
P. patens
O. lauri
E. pallens
C. reinhardtii
D. salina
K. nitens
M. polymorpha
O. sativa
S. lycopersicum
A. thaliana
C. sinensis
A. coerulea
C. richardii
A. filiculoides
C. quitensis
A. fistulosa
U. mutica
E. crista
Polytrichum
M. vitrea
C. pennata
T. ferrea
M. endolithica
M. polymorpha
S. angustifolium
P. trichocarpa
C. clementina
C. sativus
L. esculentum
A. agrestis
S. dolicha
D. magellanica
C. moellendorffii
S. cruciatum
S. moellendorffii
L. tenuis
C. purpurea
Amborella
O. sativa
S. bicolor
A. comosus
A. tauschii
A. cunninghamii
A. fistulosa
C. richardii
S. trichocarpa
C. chinensis
M. domestica
M. truncatula
C. sinensis
A. thaliana
Prasinomonadophyta
Mamiellophyceae
Other Chlorophytes
Streptophyta algae
Bryophytes
Lycophyta And Ferns
Spermatophyta
B
Number of gains per node
Number of expansions per node
C
Sensibility
1.00
0.75
0.50
0.25
0
O. lauri
C. reinhardtii
D. salina
K. nitens
M. polymorpha
O. sativa
S. lycopersicum
D
1.00
0.75
0.50
0.25
0
O. lauri
C. reinhardtii
D. salina
K. nitens
M. polymorpha
O. sativa
S. lycopersicum

A

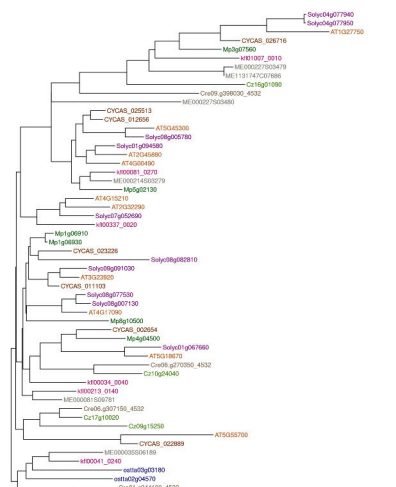
BES1/BZR1 OG (OG0001006)

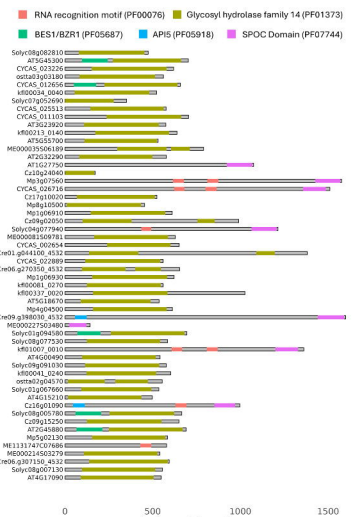


B

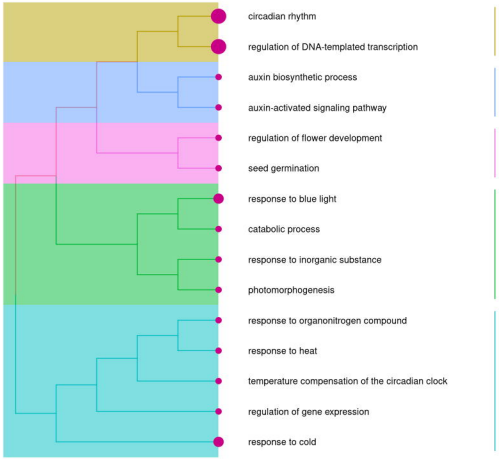


Beta amylase OG (OG0000364)

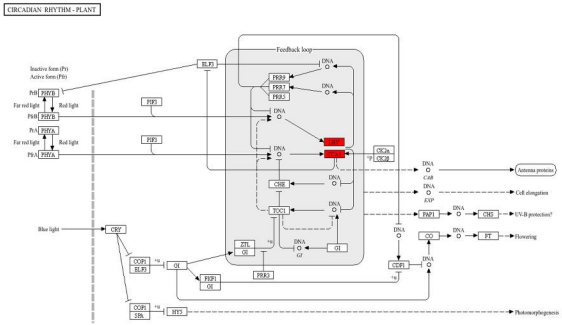




A



B



C

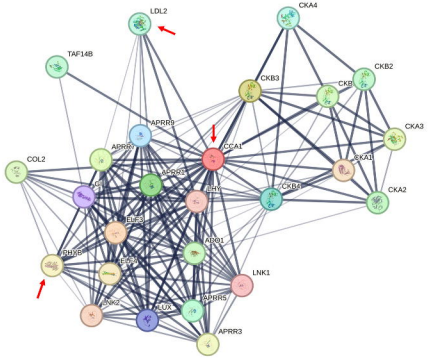
Literature Table

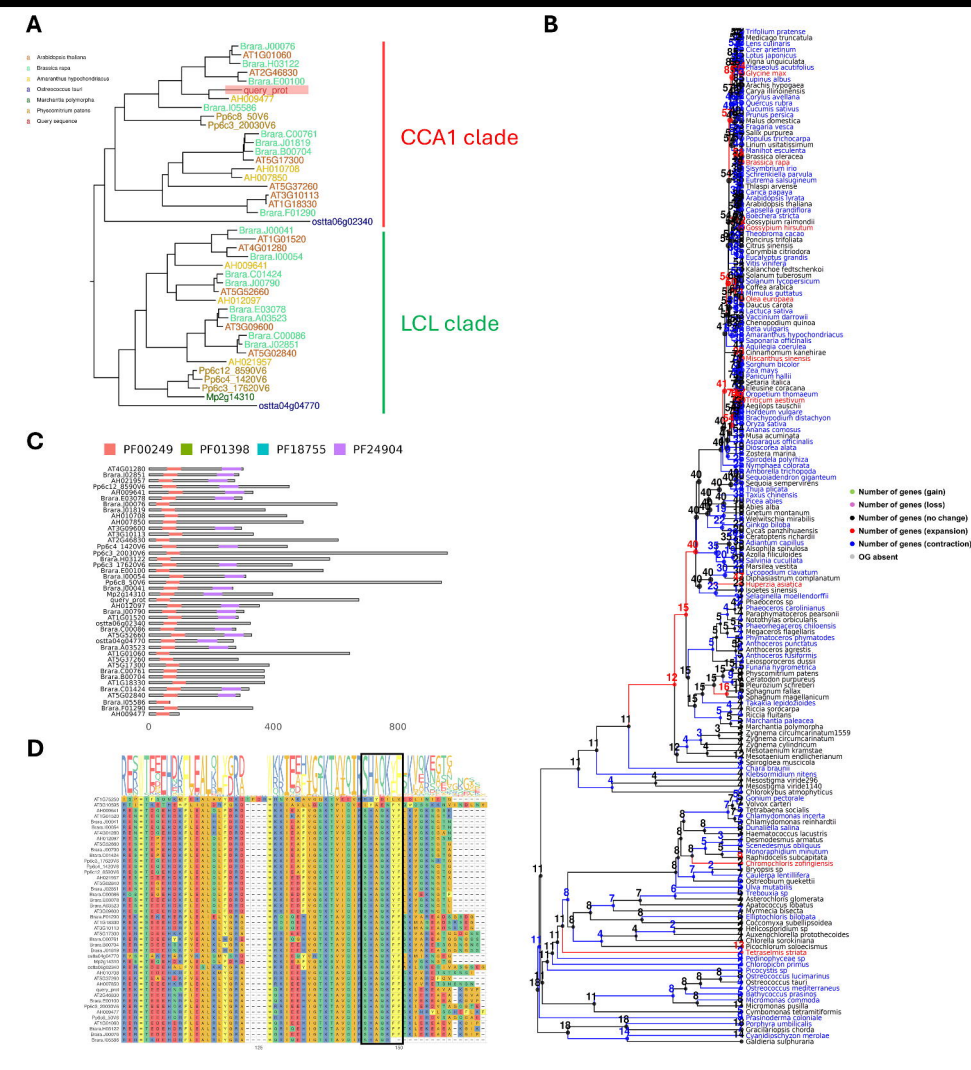
Show 10 entries Search:

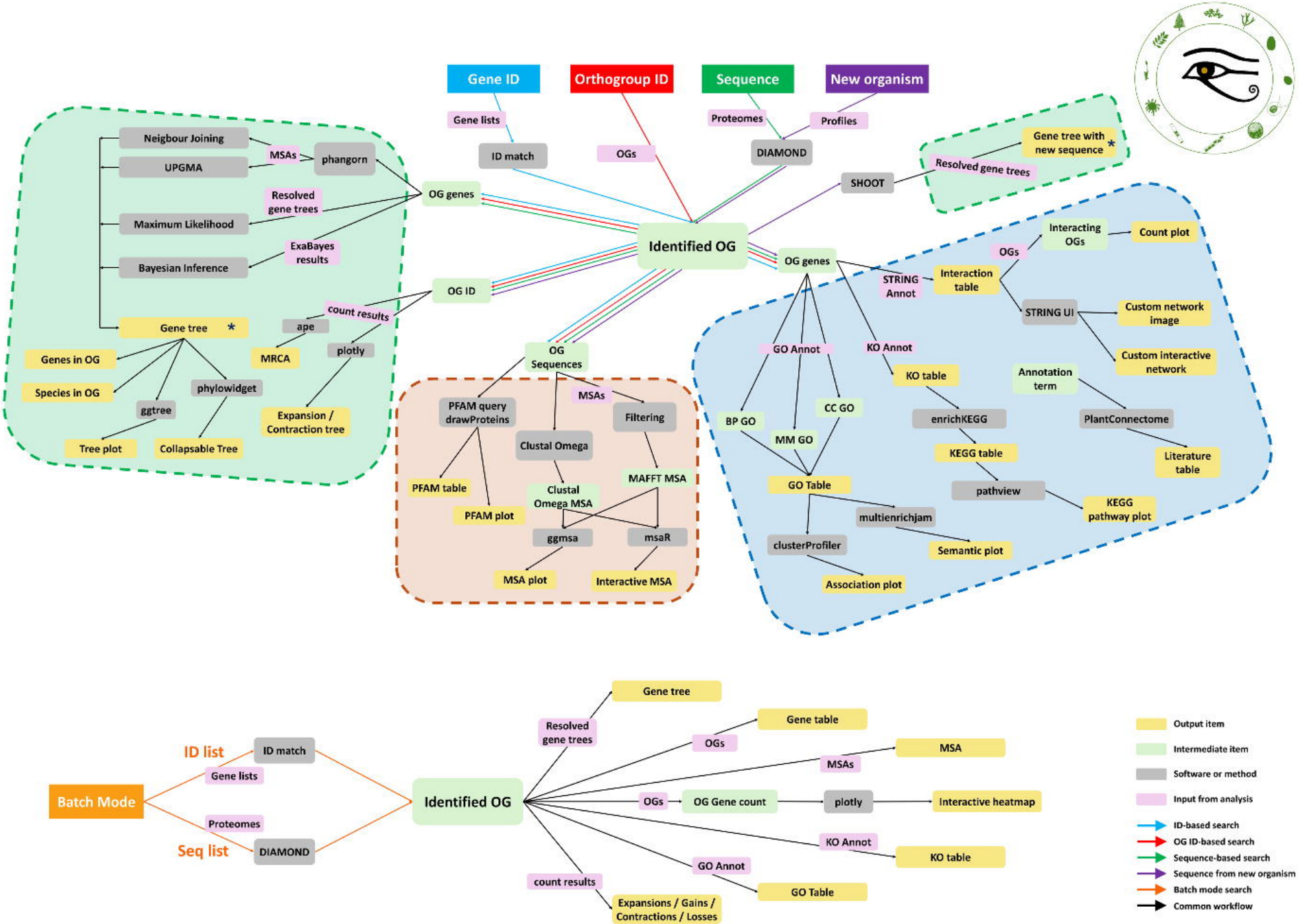
Source	Interaction Type	Target	Pubmed ID
CCA1	OVEREXPRESSION PERTURBS	APRR1/TOC1 RHYTHMS	11828029
CCA1	ALTERS	APRR1 RHYTHM	11828029
PRR7	AFFECTS	CCA1	14563930
CCA1	PRODUCES	MOLECULAR COMPONENTS OF CIRCADIAN OSCILLATOR	14563930
LKP2	UNRELATED TO	CCA1	15310821
CCA1	DIFFER IN	PRR5-11 PRR7-11	15695441
CCA1	INCREASES WITH	TEMPERATURE	16617099
CCA1	AFFECTS	CIRCADIAN RHYTHMS	17132630
CCA1	REPRESSES	LHY	17519251
GENE	INTERACTS WITH	CCA1	17525082

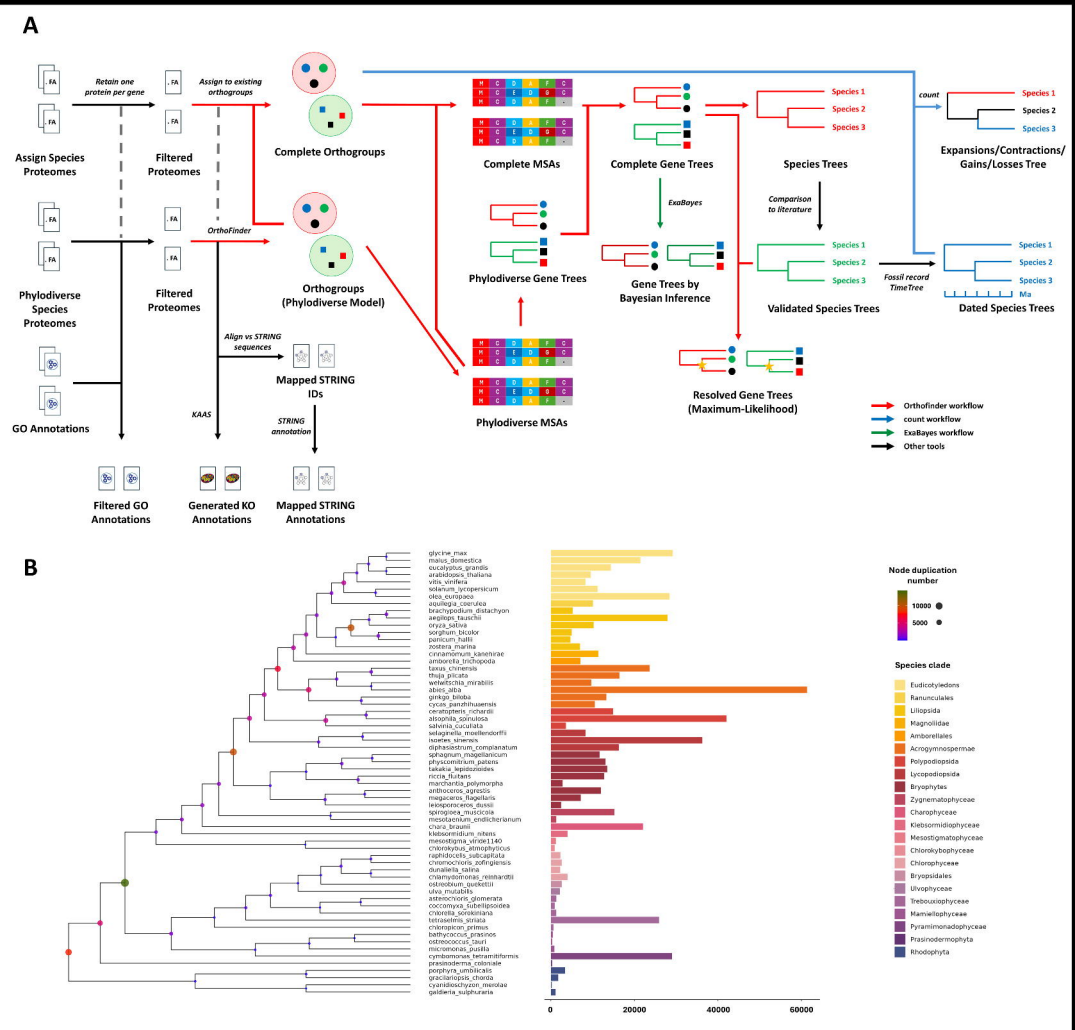
Showing 1 to 10 of 165 entries Previous 1 2 3 4 5 ... 17 Next

D




A
Arabidopsis thaliana
Brassica rapa
Amaranthus hypochondriacus
Ostreococcus tauri
Marchantia polymorpha
Physcomitrium patens
Query sequence
Brara.J00076
AT1G01060
Brara.H03122
AT2G46830
Brara.E00100
query_prot
AH009477
Brara.I05586
Pp6c8_50V6
Pp6c3_20030V6
Brara.C00761
Brara.J01819
Brara.B00704
AT5G17300
AH010708
AH007850
AT5G37260
AT3G10113
AT1G18330
Brara.F01290
ostta06g02340
CCA1 clade
Brara.J00041
AT1G01520
AT4G01280
Brara.I00054
AH009641
Brara.C01424
Brara.J00790
AT5G52660
AH012097
Brara.E03078
Brara.A03523
AT3G09600
Brara.C00086
Brara.J02851
AT5G02840
AH021957
Pp6c12_8590V6
Pp6c4_1420V6
Pp6c3_17620V6
Mp2g14310
ostta04g04770
LCL clade
B
Trifolium pratense
Medicago truncatula
Lens culinaris
Cicer arietinum
Lotus japonicus
Vigna unguiculata
Phaseolus acutifolius
Glycine max
Lupinus albus
Arachis hypogaea
Cicer arietinum
Corylus avellana
Quercus rubra
Cucumis sativus
Prunus persica
Malus domestica
Fragaria vesca
Salix purpurea
Populus trichocarpa
Linum usitatissimum
Manihot esculenta
Brassica oleracea
Brassica rapa
Sisymbrium irio
Schrenkiella parvula
Eutrema salsugineum
Thlaspi arvense
Carica papaya
Arabidopsis lyrata
Arabidopsis thaliana
Capsella grandiflora
Boechera stricta
Gossypium raimondii
Gossypium hirsutum
Theobroma cacao
Poncirus trifoliata
Citrus sinensis
Corymbia citriodora
Eucalyptus grandis
Vitis vinifera
Kalanchoe fedtschenkoi
Solanum tuberosum
Solanum lycopersicum
Coffea arabica
Mimulus guttatus
Olea europaea
Daucus carota
Lactuca sativa
Vaccinium darrowii
Chenopodium quinoa
Beta vulgaris
Amaranthus hypochondriacus
Saponaria officinalis
Aquilegia coerulea
Cinnamomum kanehirae
Miscanthus sinensis
Sorghum bicolor
Zea mays
Panicum hallii
Setaria italica
Eleusine coracana
Oropetium thomaeum
Triticum aestivum
Aegilops tauschii
Hordeum vulgare
Brachypodium distachyon
Oryza sativa
Ananas comosus
Musa acuminata
Asparagus officinalis
Dioscorea alata
Zostera marina
Spirodela polyrhiza
Nymphaea colorata
Amborella trichopoda
Sequoiadendron giganteum
Sequoia sempervirens
Thuja plicata
Taxus chinensis
Picea abies
Abies alba
Gnetum montanum
Welwitschia mirabilis
Ginkgo biloba
Cycas panzhihuaensis
Ceratopteris richardii
Adiantum capillus
Alsophila spinulosa
Azolla filiculoides
Salvinia cucullata
Marsilea vestita
Lycopodium clavatum
Diphasiastrum complanatum
Huperzia asiatica
Isoetes sinensis
Selaginella moellendorffii
Phaeoceros sp
Phaeoceros carolinianus
Paraphymatoceros pearsonii
Notothylas orbicularis
Phaeothegaceros chilonensis
Megaceros flagellaris
Phymatoceros phymatodes
Anthoceros punctatus
Anthoceros agrestis
Anthoceros fusiformis
Leiosporoceros dussii
Fumaria hygrometrica
Physcomitrium patens
Ceratodon purpureus
Pleurozium schreberi
Sphagnum fallax
Sphagnum magellanicum
Takakia lepidozioides
Riccia sorocarpa
Riccia fluitans
Marchantia paleacea
Marchantia polymorpha
Zygnema circumcarinatum1559
Zygnema circumcarinatum
Zygnema cylindricum
Mesotaenium kramstae
Mesotaenium endlicherianum
Spirogloea musicola
Chara braunii
Klebsormidium nitens
Mesostigma viride296
Mesostigma viride1140
Chlorokybus atmophyticus
Ionium pectorale
Volvox carteri
Terrabaena socialis
Chlamydomonas incerta
Chlamydomonas reinhardtii
Dunaliella salina
Haematococcus lacustris
Desmodesmus armatus
Scenedesmus obliquus
Monoraphidium minutum
Raphidocelis subcapitata
Chromochloris zohingiensis
Bryopsis sp
Cauleria lenticillifera
Ostreobium quekettii
Ulva mutabilis
Trebouxia sp
Asterochloris glomerata
Apatococcus fibulus
Myrmecia bisecta
Eliptochloris bilobata
Coccomyxa subellipsoidea
Helicosporidium sp
Auxenochlorella protothecoides
Chlorella sorokiniana
Picochlorum soloecismus
Tetraselmis striata
Pedinophyceae sp
Chloropicon primus
Picocystis sp
Ostreococcus lucimarinus
Ostreococcus tauri
Ostreococcus mediterraneus
Bathycoccus prasinos
Micromonas commoda
Micromonas pusilla
Cymbomonas tetramitiformis
Pyramiodermia coloniale
Porphyra umbilicalis
Gracilariopsis chorda
Cyanidioschyzon merolae
Galdieria sulphuraria
Number of genes (gain)
Number of genes (loss)
Number of genes (no change)
Number of genes (expansion)
Number of genes (contraction)
OG absent
C
PF00249
PF01398
PF18755
PF24904
AT4G01280
Brara.J02851
AH021957
Pp6c12_8590V6
AH009641
Brara.E03078
Brara.J00076
Brara.J01819
AH010708
AH007850
AT3G09600
AT3G10113
AT2G46830
Pp6c4_1420V6
Pp6c3_20030V6
Brara.H03122
Pp6c3_17620V6
Brara.E00100
Brara.J00054
Pp6c8_50V6
Brara.J00041
Mp2g14310
query_prot
AH012097
Brara.J00790
AT1G01520
ostta04g02340
Brara.C00086
AT5G52660
ostta04g04770
Brara.A03523
AT1G01060
AT5G37260
AT5G17300
Brara.C00761
Brara.B00704
AT1G18330
Brara.C01424
AT5G02840
Brara.I05586
Brara.F01290
AH009477
0
400
800
D
126
193


Gene ID
Orthogroup ID
Sequence
New organism
Gene lists
ID match
OGs
Proteomes
DIAMOND
Profiles
SHOOT
Resolved gene trees
Gene tree with new sequence *
Identified OG
OG genes
OG ID
Neigbour Joining
UPGMA
MSAs
phangorn
Resolved gene trees
Maximum Likelihood
ExaBayes results
Bayesian Inference
Gene tree *
count results
ape
MRCA
plotly
Genes in OG
Species in OG
ggtree
phylowidget
Tree plot
Collapsable Tree
Expansion / Contraction tree
OG Sequences
PFAM query drawProteins
MSAs
Filtering
Clustal Omega
PFAM table
PFAM plot
Clustal Omega MSA
MAFFT MSA
ggmsa
msaR
MSA plot
Interactive MSA
OG genes
STRING Annot
Interaction table
OGs
Interacting OGs
Count plot
STRING UI
Custom network image
Custom interactive network
GO Annot
KO Annot
KO table
Annotation term
BP GO
MM GO
CC GO
enrichKEGG
PlantConnectome
KEGG table
Literature table
GO Table
pathview
KEGG pathway plot
multienrichjam
clusterProfiler
Semantic plot
Association plot
Batch Mode
ID list
Seq list
ID match
Gene lists
Proteomes
DIAMOND
Identified OG
Resolved gene trees
Gene tree
OGs
Gene table
MSAs
MSA
OGs
OG Gene count
plotly
Interactive heatmap
KO Annot
KO table
GO Annot
GO Table
count results
Expansions / Gains / Contractions / Losses
Output item
Intermediate item
Software or method
Input from analysis
ID-based search
OG ID-based search
Sequence-based search
Sequence from new organism
Batch mode search
Common workflow


A
Assign Species Proteomes
Retain one protein per gene
Filtered Proteomes
Assign to existing orthogroups
Complete Orthogroups
Phylodiverse Species Proteomes
OrthoFinder
Orthogroups (Phylodiverse Model)
GO Annotations
Filtered GO Annotations
KAAS
Generated KO Annotations
Align vs STRING sequences
Mapped STRING IDs
STRING annotation
Mapped STRING Annotations
Complete MSAs
Phylodiverse Gene Trees
Phylodiverse MSAs
Complete Gene Trees
ExaBayes
Gene Trees by Bayesian Inference
Resolved Gene Trees (Maximum-Likelihood)
Species Trees
Comparison to literature
Validated Species Trees
Fossil record
TimeTree
Dated Species Trees
count
Expansions/Contractions/ Gains/Losses Tree
Species 1
Species 2
Species 3
Ma
Orthofinder workflow
count workflow
ExaBayes workflow
Other tools
B
glycine_max
malus_domestica
eucalyptus_grandis
arabidopsis_thaliana
vitis_vinifera
solanum_lycopersicum
olea_europaea
aquilegia_coerulea
brachypodium_distachyon
aegilops_tauschii
oryza_sativa
sorghum_bicolor
panicum_hallii
zostera_marina
cinnamomum_kanehirae
amborella_trichopoda
taxus_chinensis
thuja_plicata
welwitschia_mirabilis
abies_alba
ginkgo_biloba
cycas_panzhihuaensis
ceratopteris_richardii
alsophila_spinulosa
salvinia_cucullata
selaginella_moellendorffii
isoetes_sinensis
diphasiastrum_complanatum
sphagnum_magellanicum
physcomitrium_patens
takakia_lepidozioides
riccia_fluitans
marchantia_polymorpha
anthoceros_agrestis
megaceros_flagellaris
leiosporoceros_dussii
spirogloea_muscicola
mesotaenium_endlicherianum
chara_braunii
klebsormidium_nitens
mesostigma_viride1140
chlorokybus_atmophyticus
raphidocelis_subcapitata
chromochloris_zofingiensis
dunaliella_salina
chlamydomonas_reinhardtii
ostreobium_quekettii
ulva_mutabilis
asterococcus_glomerata
coccomyxa_subellipsoidea
chlorella_sorokiniana
tetraselmis_sinaia
chlorapicon_primus
bathycoccus_prasinos
ostreococcus_tauri
micromonas_pusilla
cymbomonas_tetramitiformis
prasinoderma_coloniale
porphyra_umbilicalis
gracilariopsis_chorda
cyanidioschyzon_merolae
galdieria_sulphuraria
0
20000
40000
60000
Node duplication number
10000
5000
Species clade
Eudicotyledons
Ranunculales
Liliopsida
Magnoliidae
Amborellales
Acrogymnospermae
Polypodiopsida
Lycopodiopsida
Bryophytes
Zygnematophyceae
Charophyceae
Klebsormidiophyceae
Mesostigmatophyceae
Chlorokybophyceae
Chlorophyceae
Bryopsidales
Ulvophyceae
Trebouxiophyceae
Mamiellophyceae
Pyramimonadophyceae
Prasinodermophyta
Rhodophyta