Published online 23 September 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab088

RapGreen, an interactive software and web package to explore and analyze phylogenetic trees

Jean-François Dufayard[1,2,*], Stéphanie Bocs[1,2], Valentin Guignon[2,3], Delphine Larivière[1,2], Alexandra Louis[4], Nicolas Oubda[1,2], Mathieu Rouard[2,3], Manuel Ruiz[1,2] and Frédéric de Lamotte[2,5]

[1]CIRAD, UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France, [2]French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France, [3]Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, France, [4]IBENS, Institut de Biologie de l'ENS, Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France and [5]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

Received June 09, 2021; Revised September 09, 2021; Editorial Decision September 12, 2021; Accepted September 13, 2021

ABSTRACT

RapGreen is a modular software package targeted at scientists handling large datasets for phylogenetic analysis. Its primary function is the graphical visualization and exploration of large trees. In addition, RapGreen offers a tree pattern search function to seek evolutionary scenarios among large collections of phylogenetic trees. Other functionalities include tree reconciliation with a given species tree: the detection of duplication or loss events during evolution and tree rooting. Last but not least, RapGreen features the ability to integrate heterogeneous data while visualizing and otherwise analyzing phylogenetic trees.

INTRODUCTION

Phylogenetic analysis is a crucial step in many biology projects. The proper interpretation of a phylogenetic tree depends as much on the robustness of the algorithms involved in the analysis as on the sophistication of the user interface. The latter is critical since most end-users have non-computer-science-related backgrounds.

The forerunner to RapGreen, the RAP software (1), was developed to (i) automatically reconcile phylogenetic trees with the species tree, (ii) display phylogenetic trees and (iii) explore phylogenetic tree collections using FamFetch, the HOBACGEN interface (2). Despite the effectiveness of the algorithms embedded in RAP, the learning curve to use FamFetch was steep; further, FamFetch was dedicated to exploring specific tree collections to which it was not possible to add user-created phylogenetic trees.

To offer a more comprehensive and user-friendlier service, we developed RapGreen. RapGreen is composed of three modules: (i) a Java package to compute analysis like tree reconciliation and rooting, and several statistics, (ii) a web interface (PHP, JS) to mine phylogenetic tree collections using tree patterns and (iii) a tree web visualization (PHP, JS) tool able to integrate heterogeneous data around tree topologies.

This new implementation addresses the drawbacks of the previous RAP version as it is easy to install and comes with a web interface. RapGreen is dedicated to analyze, display and explore any tree or tree collection, as long as they are in the Newick format. In addition, great attention has been paid to the user experience in terms of performance and ease of use. First, the system responds well even with large trees (several thousand leaves): the improved response time provides users with expanded exploring capabilities. Second, tree visualization allows both a schematic display of a tree or a more detailed vision by zooming into particular regions of the tree: annotations are displayed with a level of details adapted to the zoom setting. The modularity of RapGreen allows the use of the tree visualization, tree exploration, and tree analysis tools independently from each other.

MATERIALS AND METHODS

RapGreen is a modular software that contains three distinct features enabling it to reconcile, visualize and search gene trees corresponding to a specific pattern.

InTreeGreat, an integrative tree visualization

In the post-genomics era, datasets are becoming increasingly larger. Their analysis can be cumbersome and requires

*To whom correspondence should be addressed. Tel : +33 4 67 61 65 12; Mob: +33 6 16 78 31 64; Fax : +33 4 67 61 56 05;
Email: jean-francois.dufayard@cirad.fr

very efficient and easy-to-use interfaces so that users are able to drive the phylogenetic analysis of their own datasets in an autonomous fashion. InTreeGreat (Figure 1, (3)) is a Javascript/PHP interface, compatible with every standard web browser without requiring any plugin nor add-on. Its functionalities include the display and exploration of any tree in Newick or Newick extended format (e.g. NHX), branch and leaf coloring, displaying of branch length and branch support (or any other branch labels), and the integration and visualization of heterogeneous data (e.g. annotations and expression profiles). As an example, https://github.com/SouthGreenPlatform/rap-green/wiki/Examples-of-installed-services shows how to visualize micro-syntenies along a phylogenetic tree. This has been developed with the collaboration of the Genomicus project (4).

Tree pattern matching

Tree pattern matching consists of the definition of an evolutionary scenario (called tree pattern) that is then searched for in a collection of phylogenetic trees (see Figure 1). It may be used, among other examples, to retrieve orthologous candidates in large comparative genomic datasets, and/or to construct queries to find recently duplicated genes in specific taxa or gene losses at a defined point of the species history. The pattern is defined as a subtree that is exhaustively mined for in the whole gene tree collection, and each of the pattern's node or branch can be constrained on its nature (duplication/speciation if inferred in trees by RapGreen, or duplication/speciation/horizontal transfer if inferred in trees by another software), or by desired or undesired taxonomic levels. A description of functionalities is available here: https://github.com/SouthGreenPlatform/rap-green/wiki/Tree-pattern-matching-user-guide.

The tree pattern search algorithm, initially available in the FamFetch software, has been implemented as (i) a Javascript/PHP user interface to edit patterns and explore results, and (ii) a Java daemon (part of the Java package described in the next section) that can be installed on any Java-compatible infrastructure to manage the computational part using a client/server architecture between for example a computing cluster and a webserver. The most important difference with the former FamFetch version is that any user-built tree collection can be added to the search space. Results may be visualized in the InTreeGreat interface or exported in Newick format to be used with other visualization clients or analysis tools.

Phylogenetic tree analyses

A careful phylogenetic analysis requires finding the most reliable tree root and inferring events like gene duplications or losses which occurred during evolution. The RapGreen Java package meets these requirements as it allows the manipulation and comparison of phylogenetic trees. The main features are (i) the tree pattern matching daemon that handles the computational side of tree pattern matching on a computing cluster; (ii) the tree reconciliation module that allows the determination of gene duplications by comparison with a given corresponding species tree, and the prediction of paralogy and orthology relationships with several related statistics: https://github.com/SouthGreenPlatform/rap-green/wiki/About-gene-pair-statistics

Tree reconciliation also provides the ability to choose a root minimizing the number of predicted duplications and losses.

The tree data structure implementation is recursive, and algorithms have been re-implemented from the RAP 2005 version (1). This new implementation is compatible with the most recent JDK versions (JDK11 for example) and benefits from recent improvements of the Java virtual machine: (i) The tree pattern algorithm relies on a recursive unordered tree pattern matching method that has been adapted to phylogenetic tree specificities. It takes as an input a collection of phylogenetic trees, a tree pattern that will be mined as a subtree in the collection, and a species tree allowing any taxonomic level to be used in the pattern edition. The output is a subcollection of phylogenetic trees annotated with the pattern occurrences. It is also possible to get the list of matching sequences retrieved by the pattern. The main improvement of this algorithm since its former version is a full indexation of species in the species tree, allowing a huge number of taxa to be taken into account. (ii) The tree reconciliation algorithm is a simple reimplementation of the former version (1). It is based on a parsimony method: it predicts duplications minimizing their number, allowing poorly supported branches to be collapsed in the phylogenetic trees and in the species tree.

RESULTS

Tree pattern matching use case

RapGreen is able to address complex evolutionary questions. A use case is searching for orthologous genes between barley (*Hordeum vulgare*) and cultivated rice (*Oryza sativa*) under the scenario that barley genes have been duplicated at least twice in the recent barley history (three copies or more). Such a request can be formulated defining a tree pattern (see Figure 2A), on top of the tree, one sees that the clade under the duplication (red square) is constrained to contain only barley genes, and at least three of them (consequently two duplications). On the bottom, at least one rice gene is expected, orthologous to barley genes, because they diverged after a speciation event (green lozenge).

Searching for such a tree pattern usually takes no longer than one second (tested in a forest of 300 000 trees). The list of resulting matching trees (see Figure 2B) can be explored, and each tree can be displayed individually (see Figure 2C). In this example, the GreenPhylDB family GP000072 contains one occurrence of this pattern: one rice gene linked to three barley genes while the whole tree contains 2308 genes.

Each gene on the tree is clickable, linked to the GreenPhylDB page dedicated to this gene. For example Figure 2D shows the protein domain composition of each gene: we observe the intraspecific diversity in comparison to close orthologous diversity. In this case, one copy of the barley gene seems to have lost some of the protein domains conserved between the rice and other barley copies.

Returning to the list of matching trees, one may download the whole list of matching sequences in a CSV format file (see Figure 2E). This file contains several columns identifying the family of the sequence, a mapping number with

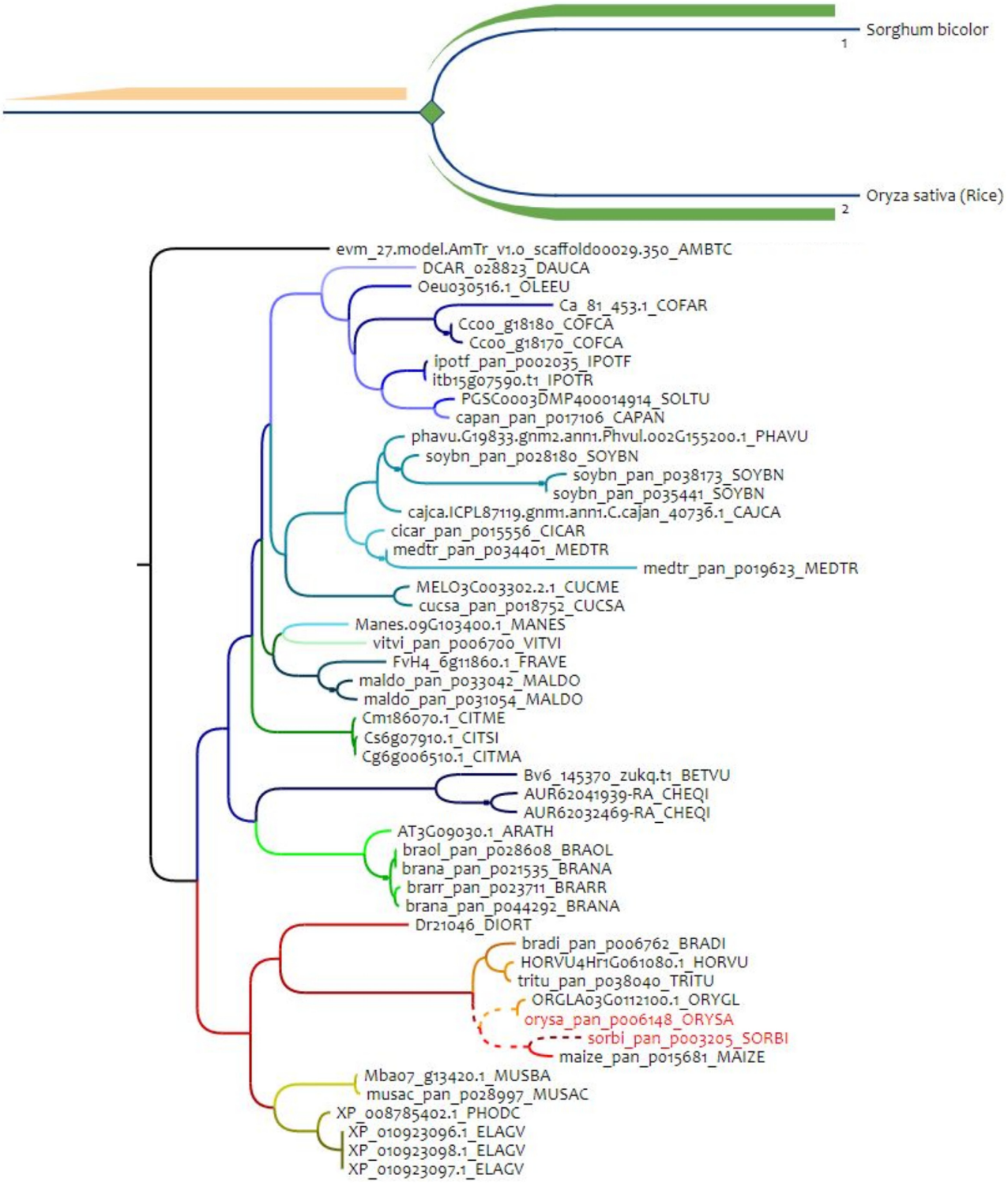
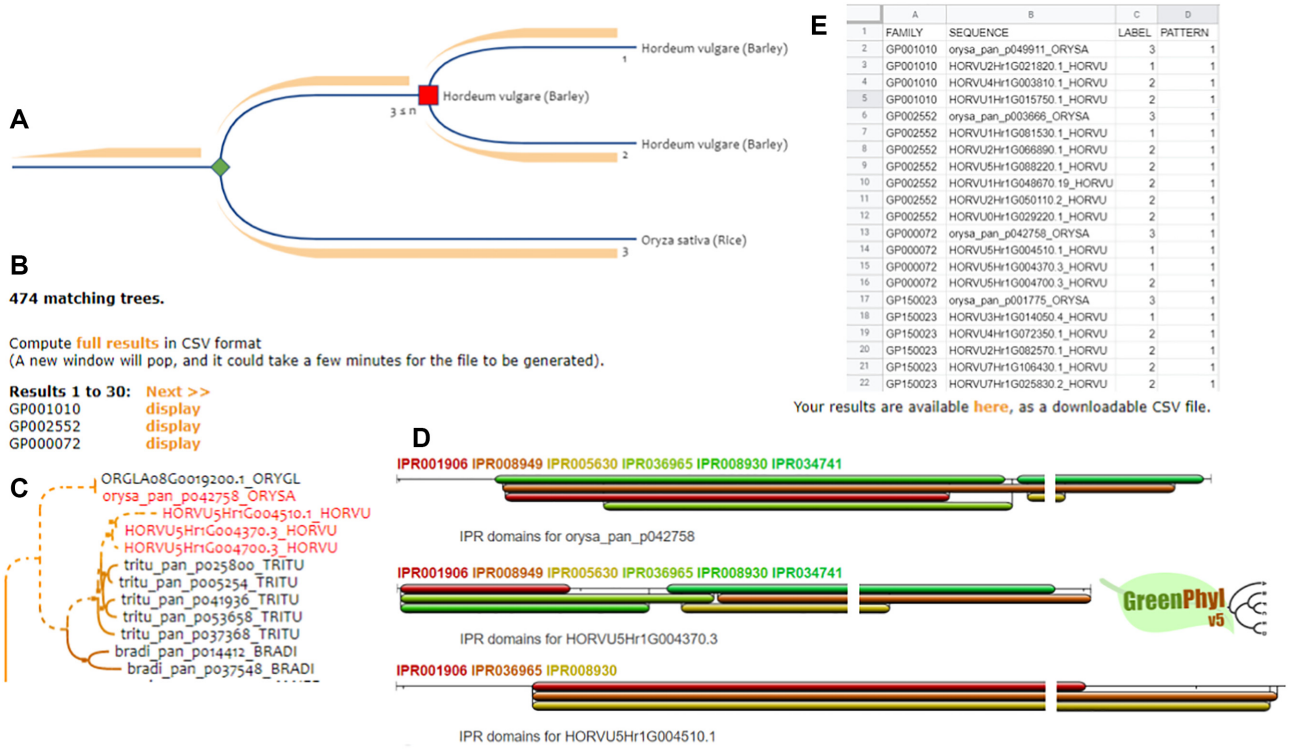


Figure 1. Top: the tree pattern edition interface, which allows to define an evolutionary scenario and search for it within a given phylogenetic tree collection. Bottom: the InTreeGreat interface shows an example of a phylogenetic tree of plant species, including the retrieved subtree (in dotted lines) matching the pattern shown on top (source: GreenPhylDB v5 (8,9)). Tree branches are colored following a color code defined in the species tree. The tree can be documented with heterogeneous related data, like functional annotations or gene expression profiles. Menus and related annotations are not displayed here to simplify the illustration.



	A	B	C	D
1	FAMILY	SEQUENCE	LABEL	PATTERN
2	GP001010	orysa_pan_p049911_ORYSA	3	1
3	GP001010	HORVU2Hr1G021820.1_HORVU	1	1
4	GP001010	HORVU4Hr1G003810.1_HORVU	2	1
5	GP001010	HORVU1Hr1G015750.1_HORVU	2	1
6	GP002552	orysa_pan_p003666_ORYSA	3	1
7	GP002552	HORVU1Hr1G081530.1_HORVU	1	1
8	GP002552	HORVU2Hr1G066890.1_HORVU	2	1
9	GP002552	HORVU5Hr1G088220.1_HORVU	2	1
10	GP002552	HORVU1Hr1G048670.19_HORVU	2	1
11	GP002552	HORVU2Hr1G050110.2_HORVU	2	1
12	GP002552	HORVU0Hr1G029220.1_HORVU	2	1
13	GP000072	orysa_pan_p042758_ORYSA	3	1
14	GP000072	HORVU5Hr1G004510.1_HORVU	1	1
15	GP000072	HORVU5Hr1G004370.3_HORVU	1	1
16	GP000072	HORVU5Hr1G004700.3_HORVU	2	1
17	GP150023	orysa_pan_p001775_ORYSA	3	1
18	GP150023	HORVU3Hr1G014050.4_HORVU	1	1
19	GP150023	HORVU4Hr1G072350.1_HORVU	2	1
20	GP150023	HORVU2Hr1G082570.1_HORVU	2	1
21	GP150023	HORVU7Hr1G106430.1_HORVU	2	1
22	GP150023	HORVU7Hr1G025830.2_HORVU	2	1

Figure 2. Schematic walkthrough of a tree pattern matching use case applied on GreenPhylDB v5. (**A**) pattern edition interface, the green lozenge represents a speciation node, and the red square a duplication node with a specific constraint (at least two duplications leading to minimum three barley genes); (**B**) resulting family list; (**C**) InTreeGreat viewer displaying one of the resulting pattern (the dotted lines represent the requested pattern); (**D**) rebound on GreenPhylDB protein domain display; (**E**) matching sequences for the 474 matching trees, in CSV format (each gene is labeled according to the pattern). (D and E) graphical displays are not included in the RapGreen package.

the pattern leaf, and the pattern number that can be greater than one if there is more than one matching pattern in the same tree.

This use case is reproducible using the website: https://github.com/SouthGreenPlatform/rap-green/wiki/Usecase:-GreenPhylDB-and-tree-pattern-matching

InTreeGreat use case

The nsLTP family (3) presents an opportunity for sequence-structure-function studies. However, finding links within heterogeneous data such as primary sequences, three-dimensional structures and functional annotations from the literature requires adapted tools. Furthermore, the difficulty increases with the sample size. Taking advantage of the RapGreen multiscale visualization features together with the enhanced knowledge representation through links to functional annotations, we were able to highlight links between certain groups of proteins and specific biological functions (3), and this despite the large size of the initial sample (800 proteins). In Figure 3, several examples of display options and annotations are presented; an interactive menu (A) allows users to display several pieces of information on branches and leaves, and to explore annotations available for this gene family.

This specific InTreeGreat interface is linked to a database dedicated to the nsLTP gene family for which manually curated annotations for each protein are provided (B). Users can explore Plant Ontology (11) or Gene Ontology (Gene Ontology Consortium 2021) annotations by a simple mouse-over the corresponding interactive texts. In this example, we choose to highlight several Plant Ontology keywords (leaf, seed, flower and root) with dedicated colors, as a colored board (B, to the right) aligned with tree leaves.

Clades in the tree can also be collapsed in order to summarize the information contained in large trees (D). Here, nsLTP proteins have been colored and collapsed by types (3) using the last common ancestor to group all corresponding annotated sequences.

This use case is reproducible using this website: https://github.com/SouthGreenPlatform/rap-green/wiki/Usecase:-Tree-visualization-with-InTreeGreat

DISCUSSION

Although the tree reconciliation algorithm was published in 2005, its functionalities (rooting, minimizing duplication and losses, collapsing incongruent and poorly supported branches, managing multifurcations) are comparable to recent methods like Treerecs (5). Importantly, RapGreen is currently the only phylogenetic package offering a tree pattern search.

The InTreeGreat phylogenetic tree viewer is one tool within an ecosystem of stand-alone and online interfaces that allow visualization of trees and correlated data. One example is the online tree viewer available on PLAZA (6) able to display the phylogenetic tree together with protein domains. Another example is Dendroscope (7): it allows full

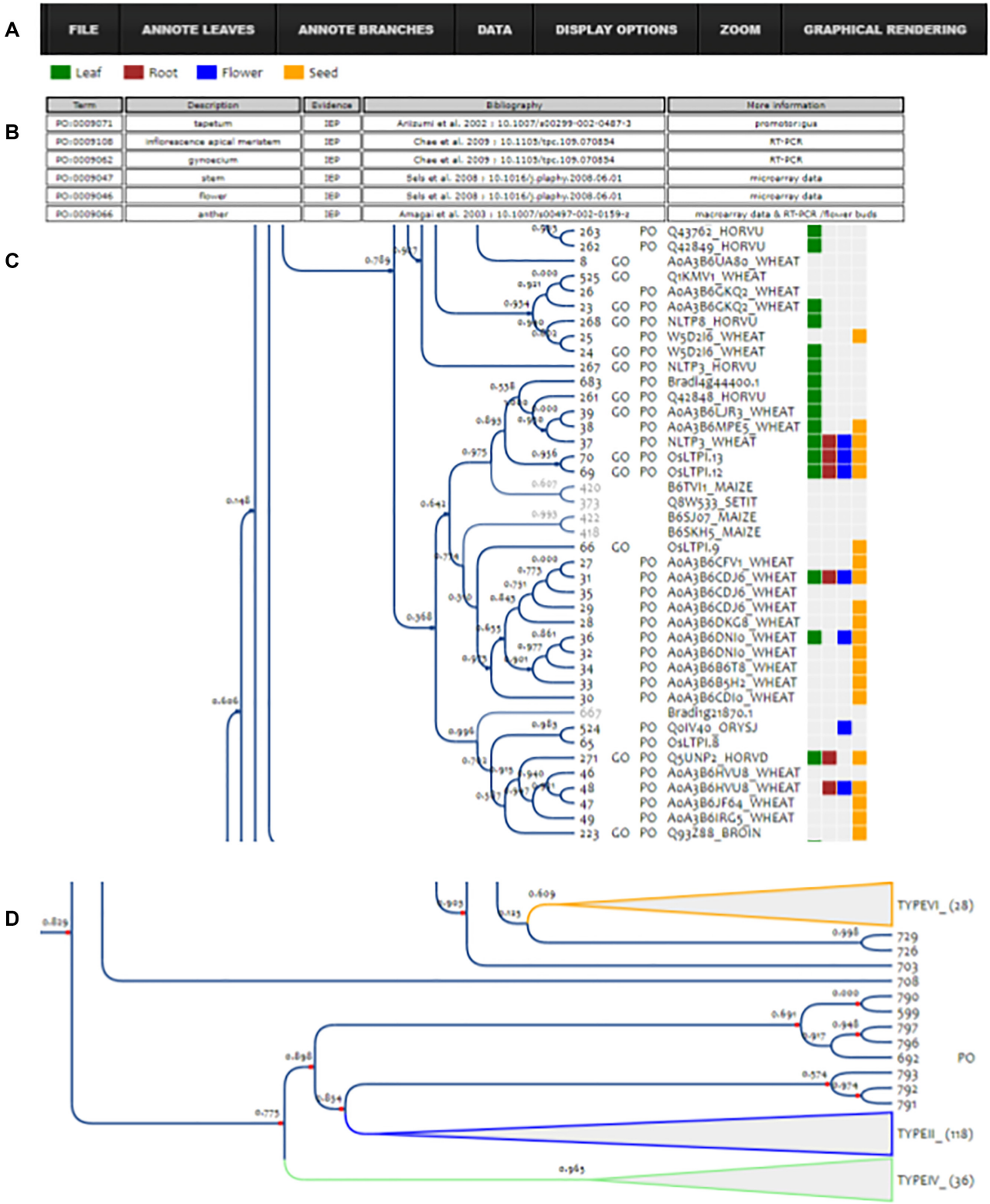


Figure 3. An example of InTreeGreat customization, for the nsLTP super-family (3). (**A**) Menu allowing access to coloring/collapsing/rendering tools. The legend is enriched automatically each time a color is associated with a keyword. (**B**) Popup providing annotations retrieved from Plant Ontology. (**C**) Partial display of the nsLTP family tree with several options annotations (GO for Gene Ontology, and PO for Plant Ontology, with annotations corresponding to the keywords: Leaf, Root; Flower and Seed) to the leaves, and a lesser contrast rendering for the unannotated genes. (**D**) Another partial display, figuring some collapsed nodes relative to nsLTP documented types.

customization of the graphical rendering of a tree. Each of these examples addresses specific needs; for its part, InTreeGreat highlights data correlations between phylogeny and heterogeneous annotations, and visualizations of the results from tree pattern matching.

RapGreen has been implemented on several public portals, notably GreenPhylDBv5 (8,9) which provides topology exploration for about ten thousand gene trees in plant genomes, and HOGENOM (10) containing >50 million sequences and one million gene trees for a wide set of sequenced organisms. With regards to visualization, InTreeGreat can handle trees of >20 000 leaves while remaining very responsive.

This application, therefore, brings improvements in the areas of portability, performance and user interface. In the short term, the tree pattern search functionality will be also integrated into the next release of Genomicus (4).

DATA AVAILABILITY

RapGreen is open-source and freely available on GitHub. The software consists of a Java package to analyze and compare phylogenetic trees, and two Javascript/PHP web interfaces: http://southgreenplatform.github.io/rap-green/

Graphical interfaces can be tested here: https://github.com/SouthGreenPlatform/rap-green/wiki/Examples-of-installed-services

The whole documentation is available here: https://github.com/SouthGreenPlatform/rap-green/wiki

ACKNOWLEDGEMENTS

This work was technically supported by the CIRAD–UMR AGAP HPC Data Center of the South Green Bioinformatics platform (https://www.southgreen.fr/) and was financially supported by French National Research Agency (ANR Genoplante): ANR- 08-GENO118 and ANR-10-BINF-03–04. We also acknowledge Paulette Lieby for her precious help in English writing.

FUNDING

French National Research Agency (ANR Genoplante) [ANR- 08-GENO118, ANR-10- BINF-03–04]; Institut National de la Recherche Agronomique.
Conflict of interest statement. None declared.

REFERENCES

1. Dufayard,J.-F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perrière,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
2. Perrière,G., Duret,L. and Gouy,M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.
3. Fleury,C., Gracy,J., Gautier,M.-F., Pons,J.-L., Dufayard,J.-F., Labesse,G., Ruiz,M. and de Lamotte,F. (2019) Comprehensive classification of the plant non-specific lipid transfer protein superfamily towards its sequence-structure-function analysis. *PeerJ*, **7**, e7504.
4. Nguyen,N.T.T., Vincens,P., Roest Crollius,H. and Louis,A. (2018) Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Res.*, **46**, D816–D822.
5. Comte,N., Morel,B., Hasić,D., Guéguen,L., Boussau,B., Daubin,V., Penel,S., Scornavacca,C., Gouy,M., Stamatakis,A. *et al.* (2020) Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics*, **36**, 4822–4824.
6. Van Bel,M., Diels,T., Vancaester,E., Kreft,L., Botzki,A., Van de Peer,Y., Coppens,F. and Vandepoele,K. (2017) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.*, **46**, D1190–D1196.
7. Huson,D.H. and Scornavacca,C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.*, **61**, 1061–1067.
8. Valentin,G., Abdel,T., Gaëtan,D., Jean-François,D., Matthieu,C. and Mathieu,R. (2020) GreenPhylDB v5: a comparative pangenomic database for plant genomes. *Nucleic Acids Res.*, **49**, D1464–D1471.
9. Guignon,V., Toure,A., Droc,G., Dufayard,J.-F., Conte,M. and Rouard,M. (2021) Correction to 'GreenPhylDB v5: a comparative pangenomic database for plant genomes'. *Nucleic Acids Res.*, **49**, 7203.
10. Penel,S., Arigon,A.-M., Dufayard,J.-F., Sertier,A.-S., Daubin,V., Duret,L., Gouy,M. and Perrière,G. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, **10**, S3.
11. Walls,R.L., Athreya,B., Cooper,L., Elser,J., Gandolfo,M.A., Jaiswal,P., Mungall,C.J., Preece,J., Rensing,S., Smith,B. *et al.* (2012) Ontologies as integrative tools for plant science. *Am. J. Bot.*, **99**, 1263–1275.