



**HAL**  
open science

## **GreenPhylDB: a database for plant comparative genomics**

M.G. Conte, Sylvain Gaillard, Nadège Lanau, Mailys Rouard, Christophe Périn

► **To cite this version:**

M.G. Conte, Sylvain Gaillard, Nadège Lanau, Mailys Rouard, Christophe Périn. GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Research*, 2008, 36, pp.(Database issue) D991-D998. 10.1093/nar/gkm934 . hal-02665685

**HAL Id: hal-02665685**

**<https://hal.inrae.fr/hal-02665685v1>**

Submitted on 31 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# GreenPhylDB: a database for plant comparative genomics

M. G. Conte<sup>1</sup>, S. Gaillard<sup>2</sup>, N. Lanau<sup>1</sup>, M. Rouard<sup>3</sup> and C. Périn<sup>1,\*</sup>

<sup>1</sup>CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier, <sup>2</sup>INRA UMR Génétique et Horticulture (GenHort) - BP 60057 - 49071 Beaucouzé cedex and <sup>3</sup>Bioversity International - Commodities for livelihood programme Parc Scientifique Agropolis II, 34397 Montpellier - Cedex 5, France

Received August 14, 2007; Revised October 9, 2007; Accepted October 11, 2007

## ABSTRACT

**GreenPhylDB (<http://greenphyl.cirad.fr>) is a comprehensive platform designed to facilitate comparative functional genomics in *Oryza sativa* and *Arabidopsis thaliana* genomes. The main functions of GreenPhylDB are to assign *O. sativa* and *A. thaliana* sequences to gene families using a semi-automatic clustering procedure and to create 'orthologous' groups using a phylogenomic approach. To date, GreenPhylDB comprises the most complete list of plant gene families, which have been manually curated (6421 families). GreenPhylDB also contains all of the phylogenomic relationships computed for 4375 families. A total of 492 TAIR, 1903 InterPro and 981 KEGG families and subfamilies were manually curated using the clusters created with the TribeMCL software. GreenPhylDB integrates information from several other databases including UniProt, KEGG, InterPro, TAIR and TIGR. Several entry points can be used to display phylogenomic relationships for *A. thaliana* or *O. sativa* sequences, using TAIR, TIGR gene ID, family name, InterPro, gene alias, UniProt or protein/nucleic sequence. Finally, a powerful phylogenomics tool, GreenPhyl Ortholog Search Tool (GOST), was incorporated into GreenPhylDB to predict orthologous relationships between *O. sativa*/*A. thaliana* protein(s) and sequences from other plant species.**

## INTRODUCTION

Comparative genomics is the study of genomic relationships between different species and serves as a significant base for functional genomics. This is also the principle method to transfer gene function/annotation towards crop

species of agronomical importance in order to hasten the identification of genes of interest (1).

Many researchers have focused their investigations on model species, where major molecular and genetic resources are available, to discover gene function and/or to study specific biological processes. *Arabidopsis thaliana* and *Oryza sativa* (rice) have emerged as model plants for dicotyledonous and monocotyledonous species, respectively, because of their compact genome sizes of 130 Mbp for *A. thaliana* and 389 Mbp for rice. Furthermore, these two species offer a wide range of referenced plant genetic and molecular resources accessible through the web. The *A. thaliana* and *O. sativa* genomic sequencing consortiums have delivered high quality, full-length sequences for both species (2,3). This in turn has paved the way for genomic comparisons between monocots and dicots, and also enabled the transfer of valuable functional information to other crop species, including cereals.

However, several limitations still hamper efficient full genome comparative analysis. Plant genes are often members of large multigenic families, thereby complicating their functional analysis and transfer of annotation. Genetic redundancy and/or neo-functionalization is frequent and further complicates gene function assignment. Although *in silico* approaches have not overcome all of these difficulties, they should greatly help future comparative genomics for all plant scientists. Phylogenomics (4), a field combining genomic and phylogenetic analysis in a high-throughput manner, appears to be the most promising route towards achieving the comprehensive identification of orthology and paralogy relationships in full genomes.

We present in this article the development of GreenPhylDB, a database for comparative genomic analysis of the *O. sativa* and *A. thaliana* full genomes. GreenPhylDB is a web accessible, user-friendly comparative platform for plant genomes studies including family classification, phylogenomic analysis information and cross-reference links (<http://greenphyl.cirad.fr>). GreenPhylDB contains 6421 multigenic families

\*To whom correspondence should be addressed. Tel: +33 0 4 67 61 71 85; Fax: +33 0 4 67 61 56 05; Email: [perin@cirad.fr](mailto:perin@cirad.fr)

half-automatically clustered including 492 TAIR (<http://www.arabidopsis.org/>), 1903 InterPro (<http://www.ebi.ac.uk/interpro/>) and 981 KEGG (<http://www.genome.jp/kegg/>) families. A total of 6421 gene families were manually curated and 4375 subjected to phylogenetics analysis. Each family was analyzed using an automatic pipeline and bootstrap scores are provided as an indicator of tree/orthologs reliability. GreenPhylDB comes with a set of dedicated tools to search for *O. sativa* and *A. thaliana* orthologs using family, InterPro, KEGG, TAIR or sequences ID as queries. Finally, it also integrates a dedicated phylogenomics tool, GreenPhyl Ortholog Search Tool (GOST), to search for *A. thaliana* and *O. sativa* orthologs using protein sequences from other plant species.

## MATERIALS AND METHODS

### Genomic data

The datasets selected for analysis by our pipeline were provided by the J. Craig Venter Institute (JCVI, formerly known as TIGR) and The Arabidopsis Information Resource (TAIR). The pseudo-chromosome reference annotation layers for *A. thaliana* (Version 6) and *O. sativa* (Version 4) were downloaded, respectively from TAIR and JCVI websites. ([ftp://ftp.arabidopsis.org/home/tair/Sequences/whole\\_chromosomes/](ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/)) ([ftp://ftp.tigr.org/pub/data/Eukaryotic\\_Projects/o\\_sativa/annotation\\_dbs/pseudomolecules/version\\_4.0](ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_4.0))

### Database content

**Automatics clustering.** A total of 21 038 clusters were produced with full *A. thaliana* and *O. sativa* genomes using the four inflation levels (1.2, 2, 3 and 5) of the TribeMCL software (5). A total of 71 000 of the 80 644 sequences from these two genomes (88%) were integrated into at least one cluster at the lower inflation value. We found 9086 unclassified sequences (orphan), 79% belonging to *O. sativa* and 21% to *A. thaliana*. Retrotransposons and transposons sequences were also automatically filtered from the 21 038 clusters using searches for sentences containing 'transposons sequences', 'retrotransposons sequences' and 'containing TE sequences'.

**Manual curation.** Clusters were first curated at the lower inflation value (labeled 1.2). If the cluster was not consistent (i.e. containing sequences members of different families), we looked at higher inflation values until we resolved consistent clusters. Consistent clusters annotations were defined according to several external sources such as TAIR, InterPro, KEGG or DATF/DRTF and more rarely, UniProt (<http://www.expasy.uniprot.org/>) or Pubmed (<http://www.ncbi.nlm.nih.gov/sites/entrez>). We also curated subfamilies when external information was available.

**Phylogenomics analysis.** After manual annotation and validation, each family was phylogenomically analyzed using GreenPhyl Pipeline (Conte, M. *et al.* submitted) to

infer ortholog and paralog relationships between members of a given family. Briefly, our pipeline follows the established phylogenetic analysis workflow: we first tested family consistency using MUSCLE (6) and LEON (7) to filter misannotated sequences. The family multi-alignment was done with MAFFT (8), followed by RASCAL (9). A total of 100 bootstrapped trees were generated with SEQBOOT and the distance matrix was computed using PROTDIST from the same package PHYLIP (10). Tree construction was performed with PHYML (11) and trees were rooted with SDI (12) from the Forester package (13). Finally, ortholog inference using bootstrapped tree was done with DORIO from the Forester package (13).

The generated family trees were carefully examined to identify poorly resolved phylogenetic reconstruction. Indeed, a consistent family can present one or several 'evolutionary distinct' subgroups in the phylogenetic tree. In this particular case, we decided to delete the phylogenomics results and ran again the phylogenomics analysis at a higher inflation level to enforce the phylogenetic signal. This procedure was repeated until well-resolved trees were obtained.

Definitions of super-orthologs and ultra-paralogs are from Zmasek (13) *Super-orthologs*: 'Given a rooted gene tree with duplication or speciation assigned to each of its internal nodes, two sequences are super-orthologous if and only if each internal node on their connecting path represents a speciation event'. They have the highest probability to share a similar function in several species and can be used with high confidence for a direct annotation transfer.


*Ultra-paralogs*: 'Given a rooted gene tree with duplication or speciation assigned to each of its internal nodes, two sequences are ultra-paralogous if and only if the smallest subtree containing them both contains only internal nodes representing duplications'. Ultra-paralogs are mostly paralogs that have undergone recent duplications in a given species, either by tandem or segmental duplications.

### Programming and database implementation

GreenPhylDB runs on a MySQL (<http://www.mysql.com>) database using Structured Query Language (SQL). Web pages are generated via Perl CGI scripts (14,15) and are delivered by an Apache HTTP server (<http://httpd.apache.org>). Bioperl API has been used to deal with different data formats. Several Java applet visualization tools are also used to view sequence multi-alignments (Jalview) (16) and phylogenomic trees (ATV) (13).


### Database structure

GreenPhylDB database was designed to store the phylogenomic data produced during the automatic pipeline execution. MySQL tables were constructed around TIGR/TAIR ID as central entry points. Each sequence is linked to families and phylogenomics predictions through associated tables.

**Name: GRAS transcription factor family**  



Family id: 20939  
 Number: 113  
 Description: DATF and DRTF description  
 CrossReferences:  
 DATF  
 DRTF  
 PMID: [10341448](#)

**Phylogenomic analysis performed**


[- Annotation page -](#)  
**Sequences:** 

Total (with splice): 101 (ARATH : 35 - ORYZA : 66) Locus (without splice): 95 (ARATH: 33 - ORYZA: 62)

Get selected sequences in


**Group structure:** 

1.2	113 (101)
2	84 (88) 2550 (5) 4594 (3)
3	90 (82) 1235 (9) 4684 (3)
5	153 (48) 434 (22) 1169 (9) 1663 (7) 4720 (2)

**Phylogenomic search bar:** 

Search in this family  relationships with score above

[Comparative search](#)

**Gene list:** 

Seq Id	Locus Alias	Uniprot	KEGG EC	1.2	2	3	5	InterPro motif	Sequence annotation
<input checked="" type="checkbox"/> <a href="#">At1g07520.1</a>				113	84	90	153	<a href="#">IPR005202</a>	scarecrow transcription factor family protein, sim
<input checked="" type="checkbox"/> <a href="#">At1g07530.1</a>				113	84	90	153	<a href="#">IPR005202</a>	scarecrow-like transcription factor 14 (SCL14), id
<input checked="" type="checkbox"/> <a href="#">At1g14920.1</a>	GAI	<a href="#">Q9LQT8</a>		113	84	90	1663	<a href="#">IPR005202</a>	gibberellin response modulator (GAI) (RGA2) / gibb
<input checked="" type="checkbox"/> <a href="#">At1g21450.1</a>	SCL1			113	84	90	434	<a href="#">IPR005202</a>	scarecrow-like transcription factor 1 (SCL1), iden
<input checked="" type="checkbox"/> <a href="#">At1g50420.1</a>	SCL3			113	84	90	153	<a href="#">IPR005202</a>	scarecrow-like transcription factor 3 (SCL3), iden
<input checked="" type="checkbox"/> <a href="#">At1g50600.1</a>				113	84	90	434	<a href="#">IPR005202</a>	scarecrow-like transcription factor 5 (SCL5), simi

**Figure 1.** Partial view of the family entry page for the GRAS transcription factor family (fid = 20939). This family was validated based on the DRTF and DATF databases, with 35 *A. thaliana* and 62 rice loci and links are provided to these databases. The group structure shows that the GRAS family is consistent at the 1.2 level (cluster number 113). At higher stringency levels, the GRAS cluster is subdivided into five clusters (153, 434, 1169 and 1663) annotated as the LAS/SCR/SHR, PAT1, SCL6 and GAI subfamilies, respectively. Mouse movement over any group displays the name of the cluster if available (in this case the GAI subfamily). Beside each cluster the numbers of loci within the cluster are visible between brackets.

## RESULTS

### GreenPhyl database statistics

A total of 21 038 genes clusters were assembled using the TribeMCL pipeline software at the four inflation levels. At publication, 6421 clusters have been manually annotated including 64 from DRTF (17)/DATF (18) transcription factor databases, 492 from TAIR family list (19), 1903 from InterPro family list (20) and 981 from the KEGG (21) database. We considered a cluster species-specific if at least two sequences belonging to a single species were grouped together at the lowest inflation value ( $I = 1.2$ ). We found 703 and 116 rice- and *A. thaliana*-specific clusters, respectively. From the 6421 annotated clusters, 4375 have been phylogenomically analyzed. We found 398 649 phylogenomic relationships, including 50 032 orthologous associations with a score above 50%.

### Data visualization

**Family visualization.** Each cluster stored in GreenPhylDB is accessible through a specific webpage. Cluster information includes cluster name, cluster ID and number, cross-references to external family classification databases (ex: DRTF, DAFT, TAIR, InterPro family domain),

comments and publication links (Figure 1). The cluster information field has been manually curated and assignment of family name was based on external information. Direct access to all families annotated using the external classification databases is available via a drop down menu called 'Search Page'.

The cluster structure is also visible (i.e. the subdivision of any cluster at higher stringency levels of inflation) including the number of *A. thaliana* and *O. sativa* loci belonging to each cluster, together with the number of gene models and splice forms in each species. Cluster structure represents the subclassification of sequences at four clustering levels. One cluster at lower stringency level is often subdivided into different subgroups at higher stringency level. This field presents four levels of stringency (1.2, 2, 3 and 5) with the corresponding cluster numbers and the number of sequences in each cluster in brackets. Figure 1 shows cluster 113 (fid = 20939) corresponding to the GRAS transcription factor family. This cluster is subdivided into 3, 3 and 5 subclusters at inflation levels 2, 3 and 5, respectively. Each subcluster can be reached directly by clicking on the cluster ID number.

A family that has been phylogenetically analyzed at a given level displays the 'phylogenetic analysis performed'

<a href="#">Home</a>	<a href="#">Clustering list</a>	<a href="#">Search Page</a>	<a href="#">Statistics</a>	<a href="#">Search Tools</a>	<a href="#">Help</a>
----------------------	---------------------------------	-----------------------------	----------------------------	------------------------------	----------------------

**At1g14920.1**

gibberellin response modulator (GAI) (RGA2) / gibberellin-responsive modulator, identical to GAI GB:CAA75492 GI:2569938 (Arabidopsis thaliana) (Genes Dev. In press)

TAIR entry: [At1g14920.1](#) OryGenesDB entry: [At1g14920.1](#) UniProt entry: [Q9LQT8](#) T-DNA Express entry: [At1g14920](#)

[- Annotation page -](#)

**\*Locus alias: GAI**

**\*No GO information from IPR motif**

**Sequence classification:**

- 113 (MCL\_I1.2): GRAS transcription factor family
- 84 (MCL\_I2)
- 90 (MCL\_I3)
- 1663 (MCL\_I5): DELLA subfamily

**Similarity evidence (BBMH and INPARANOID )**

\* No similar BBMH

\* Query belong to Inparanoid group N\* 2725

Arabidopsis	Confidence value %	Oryza	Confidence value %
At2g01570.1	100	Os03g49990.1	100
At1g14920.1	39.01		

**GREENPHYL Phylogenomic predictions**

Orthology (o) Subtree-neighbor (n) SuperOrthologs (s) Distance (D)

UniProt	Alias	o	n	s	D
Os03g49990.1 Q7G7J6	SLR1	100	46	0	0.28567

UltraParalogy (p) Distance

UniProt	Alias	p	D
At2g01570.1 Q9SLH3	RGA	100	0.122
At1g66350.1 Q9C8Y3	RGL1	52	0.31678

Segmental duplication: Oryza list Arabidopsis list

At2g01570.1 Group: 13

[View Phylogenomic Tree](#)

Sequence:

```
>At1g14920.1_ARATH
MKRDHHHHHDKTKTMMNEEDDGNHDELLAVLGYKVRSSSEMAVQAQKLEQLEVMMSNV
QEDDLSQLATEVTHYPAELYTWLD SMLTD LNP P S S M A E Y D L K A I P G D A I L N Q F A D S A S
S S N Q G G G D T Y T T M K R L K C S N G V V E T T T A T A E S T R H V L V D S Q E N G V R L V H A L L A C A E A V
Q K E N L T V A E A L V K I G F L A V S Q I G A M R K V A T Y F A E A L A R R T Y R L S P S Q S P I D H S L S D T L Q
M H F Y E T C P Y L K F A H F T A N Q A I L E A F Q G K R V H V I D F S H S Q L Q W P A L M Q A L A L R P G G P P V
F R L T G I G P P A P D N F D Y L H E V G C K L A H L A E A I H V E F E Y R G F V A N T L A D L D A S H L E R P S E I
E S V A V S V F L H K L L G R P G A I D K V L G V V N I K P E I F T V V E Q S M H N S P I F L D R F T E S L H Y
Y S T L F D S L E G V P S G D K V M S E V Y L G K Q I C N V V A C D G P D R V E R H E T L S Q U R N R F G S A G F A A
A H I G S N A F Q A S M L L A L F N G G E C Y R V E E S D G C L M L G U H T R P L I A T S A M K L S T N *
```

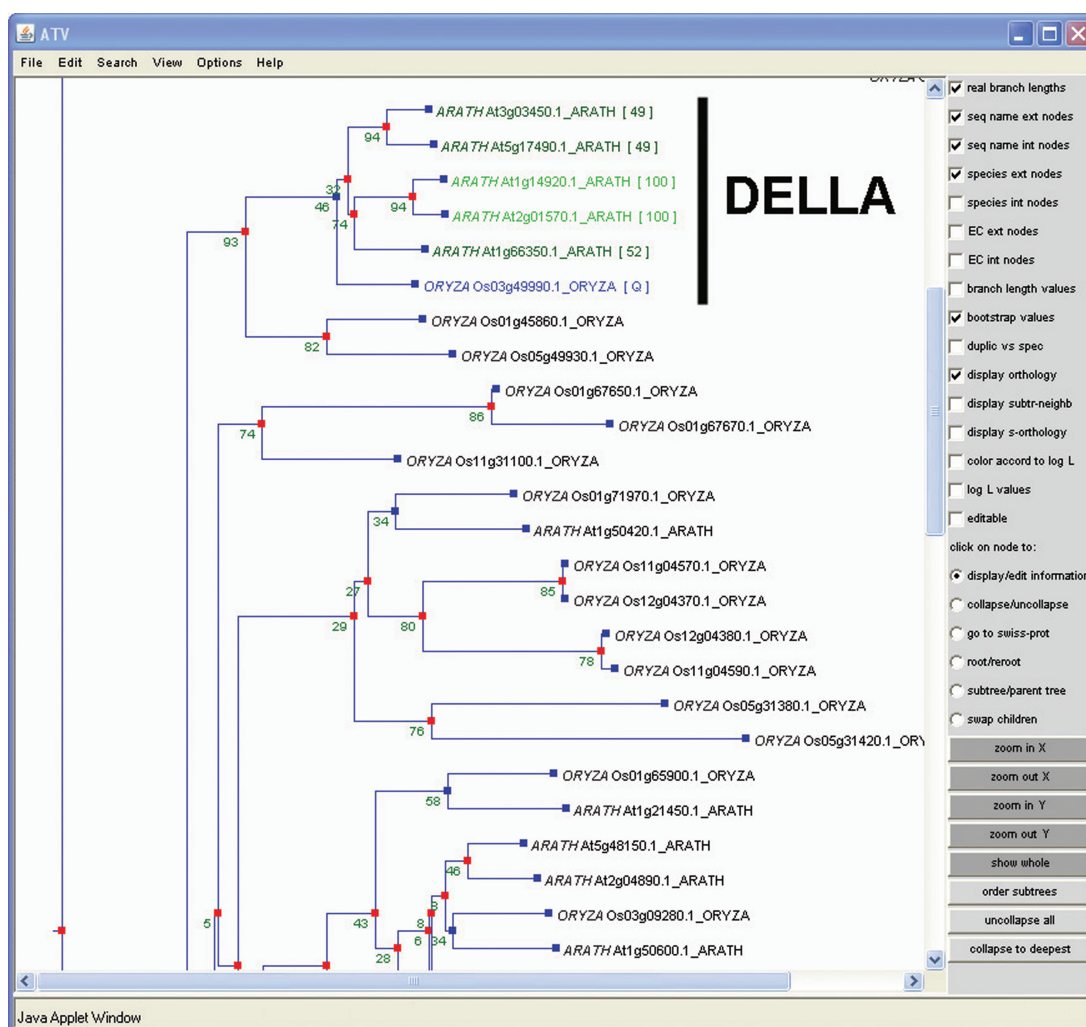
**Figure 2.** Sequence entry page for *At1g14920.1* (*GAI*). The *Os03g49990.1* (*SLR1*) rice gene is predicted as the *A. thaliana* *GAI* ortholog while *At2g01570.1* (*RGA*) and *At1g66350.1* (*RGL1*) are predicted as *A. thaliana* *GAI* Ultraparalogs. *GAI* classification inside cluster of several inflation values is visible in ‘sequence classification’ followed by ortholog similarity prediction by BBMH and Inparanoid, in this case in full agreement with the phylogenomic prediction. GreenPhyl phylogenomic prediction is separated into three sections; the ortholog prediction for the sequence, with the corresponding orthology (o), sub-tree neighbor (n), superorthologs (so) score (in%) and the genetic distance (D); the ultra-paralogs prediction for the query, in this case *RGA* and *RGL1* with the associated ultra-paralogy score (p); and finally if the query has tandem/segmental duplicated genes (using TIGR segmental duplications and OrygenesDB tandem duplication data). The phylogenomic tree is accessible through the ‘View phylogenomic tree’.

message and ‘selected for phylogenomics analysis’ when the family has been submitted. In Figure 1, the GRAS family was analyzed and a phylogenomic search bar lists all of the ortholog, super-ortholog and ultra-paralog groups above a user-defined bootstrap score (50% by default) identified for this family. The comparative search link can be used to compare ortholog predictions by similarity with the Best Blast Mutual Hits (BBMH) method, Inparanoid (<http://inparanoid.sbc.su.se/cgi-bin/index.cgi>) (22,23) and our phylogenetic method. The full alignment and the phylogenetic tree are available through two java applets, Jalview (16) and ATV (13), by clicking on ‘Family Multi-alignment’ and ‘Family tree’.

Finally, the family list of cluster’s members is visible and can be sorted by clicking on the head columns ‘Seq ID’, ‘Locus Alias’, ‘UniProt’, ‘InterPro motif’ or ‘Sequence annotation’. Most of the page features are

‘clickable’ and several help links are available through question marks. Sequence visualization details can be accessed by clicking directly on the sequence ID.

*Sequence visualization.* Each sequence stored in GreenPhylDB is accessible through a specific page that contains sequence information such as TIGR or TAIR sequence ID and annotations, cross references to external sequence databases, gene name (Alias), InterPro domains, Gene Ontology molecular function extracted from InterPro domain profiles, KEGG and EC classification. Cluster sequence classification at the four levels of inflation is also visible. For example, in Figure 2 the *A. thaliana* gene *At1g14920.1* belongs successively to groups 113 (GRAS family), 84, 90 and 1663 (*GAI* subfamily). Similarity ortholog predictions are visible in the next field ‘similarity evidence’ either by BBMH or



**Figure 3.** Partial view of the GRAS family phylogenetic tree. Two *Arabidopsis thaliana* genes (*At1g14920.1* and *At2g01570.1*) are predicted as orthologs to the query [Q] (*Os03g49990.1*) with a bootstrap support above 50%. Note that all DELLA proteins are members of the same clade.

Inparanoid. In Figure 2 the rice *GAI* ortholog, *SLRI*, predicted by phylogenetic analysis was also found by Inparanoid but not by BBMh. In the next field, GreenPhyl phylogenomic predictions are separated into two parts. The first contains the ortholog prediction for the sequence with the corresponding orthology (o), subtree neighbor (n), super-orthologs (so) score (in%) and the genetic distance (D). The second part contains the ultra-paralog predictions for the query. In this case, *A. thaliana* genes *At2g01570.1* and *At1g66350.1* are predicted as *GAI* ultra-paralogs with a score (p) of 100% and 52%, respectively. The whole phylogenomic tree is accessible through the 'View phylogenomic tree', where orthology relationships as well as duplication/speciation are visible using the ATV applet (Figure 3). Super-orthologs, ultra-paralogs and subtree-neighbors are three new concepts that were defined by Dr Zmasek (13) (see Materials and Methods section).

### GreenPhyl tools

**Search Toolbar.** A quick search toolbar is accessible at the top of each GreenPhylDB page. Users can retrieve

sequence information using sequence ID, locus name, alias name, UniProt ID or keywords. It is also possible to query with family information like a family name, InterPro domain, internal GreenPhyl cluster ID, KEGG ID or EC number.

**Family and phylogenomics search tools.** The drop-down menu entitled 'search tools' provides more advanced search possibilities. Users can retrieve family classification of a gene ID list using the 'Get classification of your ID list' tool. This tool can be used to find family or subfamily classifications in GreenPhylDB. 'Get sequence form InterPro profile' will extract a gene list based on InterPro domain profiles. Several InterPro ID can be combined by 'AND' or 'OR' in the search field; a useful feature for protein families defined as multi-domain-containing proteins. 'Get classification using BLAST search' is available to identify *O. sativa* or *A. thaliana* genes by BLAST. In addition, the BLAST output gives family classification information for the five best hits. Finally, the 'Get phylogenomic scores of your ID list' tool



**Figure 4.** Phylogenomic analysis of a non-*Arabidopsis*/rice protein sequence. The sequence of wheat *RHT1* gene (Q9ST59) is pasted into the text field of the phylogenomic search tool (A). Step 1, the sequence is tentatively attributed to GreenPhylDB clusters by BLASTP. In this case, *RHT1* belongs to the GRAS family and the DELLA subfamily (B), and the GRAS cluster was phylogenetically analyzed. The species name 'wheat' is then chosen and, after submission, the *RHT1* gene integration in the pre-computed tree is initiated (step 2). GOST produces an output list of the rice and *A. thaliana* orthologs (C) and the phylogenetic tree (D) with bootstrap scores (%).

can retrieve, whenever available, phylogenomics scores and groups of orthologs from an ID list.

*Phylogenomic analysis of another species gene.* A specific analysis tool named GOST was developed to predict *O. sativa*/*A. thaliana* orthologs using protein sequences from other species. Phylogenomic analyses follow a two-step procedure. First, the submitted protein sequence is aligned using BLASTP applied to all rice and *A. thaliana* sequences. A proposition of group classification is given based on the best BLASTP hits. The species name of the query must be correctly selected at this step as GOST compares the species tree with the gene tree to infer ortholog relationships. Then, GOST integrates the sequence into the previously saved family multi-alignment and creates the bootstrap file. An example of GOST output using the *RHT1* (Reduced Height 1) (Q9ST59) wheat gene belonging to the GRAS family is illustrated in Figure 4. Indeed, the *RHT1* is correctly predicted as an ortholog of one rice and four *A. thaliana* genes, which are DELLA proteins. This method is almost as fast as a similarity search and will help users working on unsequenced or partially sequenced plant species to obtain family classification and ortholog predictions from the two model species.

## DISCUSSION

GreenPhylDB offers several critical advantages over several recently described plant and eukaryote ortholog databases. First, most of these databases use pairwise distance comparison algorithm to determine orthology (24,25). If homology is inferred from similarity of several sequences, there is no way to be sure that they are phylogenetically connected and similarity methods cannot differentiate between paralogs and orthologs. The BBMH or Reciprocal Best Hit (RBH) search for orthologs, a popular strategy based on sequence similarity, generates false positives as similarity itself is not a reliable indicator of ortholog relatedness (26). Moreover, some of the databases deal with incomplete 'genomic repertoire' (13,27), using for instance UniProtDB accessions, and can falsely predict ortholog relations or even miss some true ortholog relations.

The only database comparable to GreenPhylDB, to our knowledge, is the orthologID database (28). GreenPhylDB nevertheless present several improvements and/or additional settings compared to orthologID. First, GreenPhylDB clustering is performed with TribeMCL, a more efficient software than other classical BLAST or PSI-BLAST methods. In addition, most of GreenPhylDB clusters were manually curated before any phylogenetic analysis to identify consistent clusters and subclusters of evolutionary-related sequences. GreenPhylDB also provides bootstrap support for ortholog predictions to quantify reliability of prediction and tree construction. Finally, users can insert their own sequences from another plant species and search for *O. sativa*/*A. thaliana* putative orthologs, a feature missing in all other plant ortholog databases.

GreenPhylDB was specifically designed for comparative functional analysis of plant orthologs and provides additional phylogenetics concepts such as ultraparalogy (14), a feature which is often synonymous with genetic redundancy and/or neo-functionalization. Each gene ID is then linked to the two most popular plant database for reverse genetics, T-DNA express (29) (see <http://signal.salk.edu>) and OrygenesDB (30) (see <http://orygenesdb.cirad.fr/index.htm>) for *A. thaliana* and *O. sativa*, respectively. External links including KEGG, TAIR, TIGR, InterPro, UniProt family were added to help cluster annotation and provide additional evidence of ortholog function.

Future plans include progressive integration of sequences from other full plant genomes and opening of the annotation section of GreenPhylDB to plant biologists requiring improved cluster classification of plant sequences. A full documentation is accessible and anyone willing to contribute to manual annotation of particular protein families is encouraged to contact [greenphylpdb@cirad.fr](mailto:greenphylpdb@cirad.fr).

## ACKNOWLEDGEMENTS

We wish to acknowledge support by the Generation Challenge Program (<http://www.generationcp.org>) and the Oryzon project of CIRAD (Centre de cooperation Internationale en Recherche Agronomique pour le Développement) for funding this research. We would also like to thank the CINES center (Centre Informatique National de l'Enseignement Supérieur, Montpellier, France) for technical support and for hosting the GreenPhyl database and website. Funding to pay the Open Access publication charges for this article was provided by CIRAD.

*Conflict of interest statement.* None declared.

## REFERENCES

- Irish, V.F. and Benfey, P.N. (2004) Beyond Arabidopsis. Translational biology meets evolutionary developmental biology. *Plant Physiol.*, **135**, 611–614.
- IRGSP. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
- AGI. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Eisen, J.A. and Fraser, C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.
- Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
- Thompson, J.D., Prigent, V. and Poch, O. (2004) LEON: multiple alignment Evaluation Of Neighbours. *Nucleic Acids Res.*, **32**, 1298–1307.
- Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
- Thompson, J.D., Thierry, J.C. and Poch, O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics (Oxford, England)*, **19**, 1155–1161.



10. Felsenstein. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. *Distributed by the author.*
11. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
12. Zmasek,C.M. and Eddy,S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**, 821–828.
13. Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
14. Christiansen,T., Torkington,N. and Wall,L. (1998) *Perl Cookbook*, 2nd edn. Sebastopol CA, ISBN 1565922433.
15. Guelich,S., Gundavaram,S. and Birznieks,G. (2000) *CGI Programming with Perl*, 2nd edn. Sebastopol CA, ISBN 1565922433.
16. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
17. Gao,G., Zhong,Y., Guo,A., Zhu,Q., Tang,W., Zheng,W., Gu,X., Wei,L. and Luo,J. (2006) DRTF: a database of rice transcription factors. *Bioinformatics*, **22**, 1286–1287.
18. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L. and Luo,J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.
19. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
20. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
21. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
22. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33** (Database issue), D476–D480.
23. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
24. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
25. Walker,N.S., Stiffler,N. and Barkan,A. (2007) POGs/PlantRBP: a resource for comparative genomics in plants. *Nucleic Acids Res.*, **35**, D852–D856.
26. Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.
27. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
28. Chiu,J.C., Lee,E.K., Egan,M.G., Sarkar,I.N., Coruzzi,G.M. and DeSalle,R. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, **22**, 699–707.
29. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P. *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.
30. Droc,G., Ruiz,M., Larmande,P., Pereira,A., Piffanelli,P., Morel,J.B., Dievart,A., Courtois,B., Guiderdoni,E. *et al.* (2006) OryGenesDB: a database for rice reverse genetics. *Nucleic Acids Res.*, **34**, D736–D740.