

## A Guide to the PLAZA 3.0 Plant Comparative Genomic Database

Klaas Vandepoele

### Abstract

PLAZA 3.0 is an online resource for comparative genomics and offers a versatile platform to study gene functions and gene families or to analyze genome organization and evolution in the green plant lineage. Starting from genome sequence information for over 35 plant species, precomputed comparative genomic data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, and genomic colinearity information within and between species. Complementary functional data sets, a Workbench, and interactive visualization tools are available through a user-friendly web interface, making PLAZA an excellent starting point to translate sequence or omics data sets into biological knowledge. PLAZA is available at <http://bioinformatics.psb.ugent.be/plaza/>.

**Key words** Gene family, Orthology, Gene functions, Comparative genomics, Plants

---

### 1 Introduction

With the advent of so-called second- and third-generation sequencing technologies, the price for whole-genome sequencing has dropped substantially during the last decade. While in the pre-“next-generation sequencing” era almost exclusively genomes from a handful of model systems were sequenced, the decrease in cost and new technologies producing longer reads have allowed numerous plant species with agricultural, economic, or environmental importance to have their genome being sequenced [1]. The availability of complete genome sequences has significantly altered our view on the complexity of plant genomes, but also generated new insights into gene functions, regulation, and genome evolution. Nevertheless, many challenges related to “understanding a new genome sequence” remain, especially for species with large genomes or lacking closely related sequenced relatives. Through the detection of similarities and differences with genomes of closely and more distantly related species, both conserved as well as novel genome features can be explored [2–4]. More

generally, comparative genomic approaches are pivotal to transfer biological knowledge from well-studied model species to non-model organisms and to gain insights into the evolution of specific genes or entire metabolic and signaling pathways [5]. However, such comparisons require high-quality data repositories to efficiently compare genes across different plant clades or to mine conserved gene functions [6].

PLAZA is an online resource for plant comparative genomics (<http://bioinformatics.psb.ugent.be/plaza/>) and offers a versatile platform to study gene functions, gene families, or genome organization and evolution. Precomputed comparative genomic data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, and genomic colinearity information within and between species. Based on the integrated genome information from more than 35 plants, different complementary functional data sets, and interactive visualization tools that are available through a user-friendly web interface, PLAZA is an excellent starting point to translate sequence or omics data sets into biological knowledge. Apart from PLAZA 3.0, which focuses on sequenced genomes of dicots and monocots [7], pico-PLAZA is integrating genomes from green, red, and brown algae and diatoms [8].

Here, we present a practical guide to PLAZA, by first giving a brief overview of the different data types and tools present in the platform. Next, we demonstrate how to use PLAZA 3.0 Dicots to analyze different sets of hormone-responsive *Arabidopsis* genes and to translate biological information to other species. This example represents a general protocol to analyze any gene set generated using a plant omics technology such as RNA-Seq, ChIP-Seq, or a proteomics-based assay. For the analysis of RNA-Seq data sets for plants lacking genome sequence information, we refer to TRAPID, an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes [9].

---

## 2 Materials

### 2.1 Genome Sequence Information

PLAZA 3.0 has been divided into a monocot- and dicot-centric section containing 31 and 16 species, respectively. Both databases contain ten shared organisms, which can serve as reference species to link between both sections or as out-groups. A complete overview of the available species can be found at [http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/genes/status](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/genes/status), while specific release information can be found by clicking a species name on the same page. PLAZA 3.0 Dicots includes 1,087,713 genes, of which 93.1% are protein encoding. These protein-coding genes are part of 26,192 multigene gene families (51.4% multispecies gene families). PLAZA 3.0 Monocots contains 537,114 genes,

of which 93.7% are protein encoding. These protein-coding genes are clustered in 19,612 multigene gene families (74.3% multispecies gene families).

## **2.2 Gene Families and Phylogenetic Trees**

Gene families are delineated by first computing the protein sequence similarity through an all-against-all BLAST ( $e$ -value cutoff  $1e-05$ , retaining the top 500 hits) and then by applying Tribe-MCL [10] and OrthoMCL [11] to cluster genes in families and subfamilies, respectively. For each (sub-)family, multiple sequence alignments are generated and stored that help to unveil conserved protein domains. Precomputed approximately maximum-likelihood phylogenetic trees generated using FastTree [12] allow users to explore orthologous and paralogous relations between genes in detail.

## **2.3 Functional Annotation Data**

In PLAZA, Gene Ontology (GO) is used to assign cellular components, molecular functions, and biological processes to genes. Apart from primary annotations obtained from external databases, also homology- and orthology-based GO projections are applied to transfer GO annotations with experimental evidence types. Whereas orthology-based projection starts from tree-based or integrative orthology gene associations, homology-based projection starts from functional terms that are enriched per gene family and which are subsequently assigned to other family members lacking this term [7]. Recently, also MapMan has been included as an additional ontology to describe gene functions, together with transcription factor family classifications from PlnTFDB [13] and PlantTFDB [14]. Also InterPro domains are included to indicate the functional regions of encoded proteins. For each GO, MapMan, and InterPro term, also a dedicated page exists which summarizes per species the number of annotated genes, as well the associated gene families.

## **2.4 Help and Documentation**

In the PLAZA platform, different methods have been implemented to offer help to the user. Extensive documentation, tutorials, and frequently asked questions sections are accessible at the bottom of each page. An interactive glossary is integrated using mouseover events in the web browser, which offer one-line descriptions of terms, data types, and methods used in the platform.

## **2.5 Arabidopsis Hormone-Specific Marker Gene Sets**

The hormone-specific marker gene sets analyzed in Subheading 3 were retrieved from Supplemental Table S9 from Nemhauser et al. [15]. These gene sets comprise upregulated and downregulated genes upon treatment using six different plant hormones. The compounds assayed included abscisic acid (ABA), indole-3-acetic acid (IAA, auxin), 1-aminocyclopropane-1-carboxylic acid (ACC, ethylene precursor), zeatin (CK, cytokinin), brassinolide (BL, brassinosteroid), and methyl jasmonate (MJ, jasmonate). In the methods Subheadings 3.2 and 3.3, IAA upregulated and downregulated genes are studied. Note that some genes, such as

AT1G02200 in the ABA up data set and AT1G65400 in the ABA down data set, are obsolete and no longer supported as TAIR10 genes and therefore will not be imported.

## 2.6 Rice Auxin-Responsive Gene Set

A set of rice auxin/IAA-responsive genes was obtained from Jain and Khurana [16]. After uploading the locus genes reported in Supplemental Table S1, which contains auxin up- and downregulated genes, 210 upregulated and 71 downregulated genes were identified. Note that some genes in Supplemental Table S1 from [16] do not represent a valid rice gene identifier and will not be uploaded in the Workbench experiment.

---

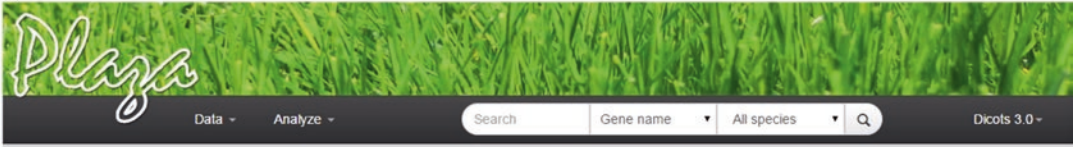
## 3 Methods

### 3.1 A PLAZA Quick Start

On the PLAZA 3.0 Dicots start page, available via [http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/), an integrated search function is available, making it possible to search for genes using gene symbols/names, gene identifiers, or functional descriptions. Optionally, the search species can be selected using a second drop-down menu (by default, the search operates on all species). Functional annotations can be searched using GO, InterPro, or MapMan descriptions or identifiers. After searching a specific gene (e.g., enter “E2Fa,” select “Gene Name,” and restrict to species “*Arabidopsis thaliana*”), a list of matches is reported. After clicking ATE2FA, with gene identifier AT2G36010, the corresponding gene page is shown ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/genes/view/AT2G36010](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/genes/view/AT2G36010)) (Fig. 1). Whereas the Overview section reports basic structural annotation and gene family information, the Descriptions section shows free-text gene function information. The PLAZA Toolbox gives an overview of the different data types and views associated with a given gene and is also available for a gene family or functional category. The Toolbox on the gene page allows exploring a gene’s colinearity in other species, local organization of homologous genes, phylogenetic tree, or orthologs. In addition, the Toolbox also hosts different views to facilitate sequence retrieval, browse BLAST hits, open the gene in the genome browser Genome View, and retrieve collinear gene pairs. Finally, the multi-tab table at the end of the page summarizes all available functional information, separated over GO, InterPro, MapMan, SignalP, and PlnTFDB/PlantTFDB.

On the E2Fa gene page, following the gene family link in the Overview section opens the corresponding gene family (*see Note 1*) page ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/gene\\_families/view/HOM03D001329](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/gene_families/view/HOM03D001329)). Whereas the Keywords box offers a quick view of the different (consensus) functional terms that are linked to this family, an interactive pie chart

A



Gene: **AT2G36010** (*Arabidopsis thaliana*)

Overview

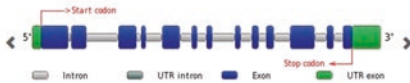
**Gene Identifier** AT2G36010  
**Transcript Identifier** AT2G36010.2  
**Gene Type** Coding gene  
**Location** 2 : 15119688-15122893 : positive

Family

**Gene family** HOM03D001329  
 (117 genes in 30 species)  
 Virdiplantae specific family  
**Subfamily** ORTH003D0017466  
 (9 genes in 7 species)  
 Rosids 1 specific family  
**Duplication type** Block duplicate

Identifiers

Identifier	Name
alias	ATE2FA
alias	E2F3
alias	E2F transcription factor 3
uniprot	Q9FNY0



Descriptions

**Description** E2F transcription factor 3

**Curated Summary** Member of the E2F transcription factors, (cell cycle genes), key components of the cyclin D/retinoblastoma/E2F pathway.

Show more...

B

**Toolbox**

**Explore**

- ...the colinearity of this gene with other genomes.
- ...the local gene organization for homologous genes.
- ...the phylogenetic tree of the homologous gene family.
- ...the orthologs using the Integrative Orthology Viewer.

**View**

- ...sequences.
- ...the multiple sequence alignment of the gene family.
- ...BLAST hits against the PLAZA database.
- ...BLAST hits against NCBI's protein database.
- ...the gene in Genomeview, a genome browser
- ...all colinear gene pairs.

C

Biological Process

GO term	Evidence(s)	Provider	Description	Source
GO:0045893	IDA	UniProt	positive regulation of transcription, DNA-dependent	1 2 3 4 5 6 7 8 9 10 11 12 13 14
GO:0051446	IDA	UniProt	positive regulation of meiotic cell cycle	1 2 3 4 5 6 7 8 9 10 11 12 13 14
GO:0006270	RCA	Gene Ontology	DNA replication initiation	1
GO:0006275	RCA	Gene Ontology	regulation of DNA replication	1
GO:0008283	RCA	Gene Ontology	cell proliferation	1
GO:0009165	RCA	Gene Ontology	nucleotide biosynthetic process	1
GO:0010090	RCA	Gene Ontology	trichome morphogenesis	1
GO:0042023	RCA	Gene Ontology	DNA endoreplication	1
GO:0051302	RCA	Gene Ontology	regulation of cell division	1

**Fig. 1** Overview of the PLAZA gene page. (a) Structural annotation and gene family information. (b) The Toolbox lists additional tools and data types that can be explored starting from this gene. (c) Gene Ontology functional annotation overview (molecular function data not shown)

shows the gene family content per species. Next to the chart, also information about the smallest encompassing phylogenetic clade is reported, which makes it possible to identify if a gene family is clade specific or found in, e.g., all Viridiplantae. The Toolbox again lists different extra items to View (multiple sequence alignment, similarity heatmap, or genome-wide organization) or to Explore (the local gene organization for homologous genes, the phylogenetic trees of this gene family, and the expansion/depletion of species in this gene family). Finally, the table at the end of the page lists the different genes part of the family, including outlier information and a customizable download function.

On the E2Fa gene page, analysis of the Gene Ontology biological process table reveals that functional annotations with experimental (cell background colored purple) and computational reviewed or electronic (cell background colored green and red, respectively) evidence are available. Following the link to the *GO page*, “DNA replication initiation” ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/go/view/GO-0006270](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/go/view/GO-0006270)) allows identifying other genes functionally annotated with this term. Whereas the selection box “Primary data” refers to functional annotations obtained from primary data sources, “All data” refers to primary annotations as well as transferred (or projected) annotations using gene orthology and homology information [7]. By following the link “View... the associated gene families,” it is possible to explore the gene families containing one or more genes annotated with this GO term (61 families, see [http://bioinformatics.psb.ugent.be//plaza/versions/plaza\\_v3\\_dicots/go/view\\_gene\\_gf/GO-0006270](http://bioinformatics.psb.ugent.be//plaza/versions/plaza_v3_dicots/go/view_gene_gf/GO-0006270)). In the same table, the phylogenetic profile, which depicts the presence or absence of a family in a species, allows to explore the presence of homologues in different flowering plants, as well as in more primitive species like mosses (see, e.g., “ppa” *Physcomitrella patens*) and green algae (e.g., “cre” *Chlamydomonas reinhardtii*). By clicking the table header on “#associated genes,” which activates the embedded sort function, it becomes immediately clear that most families, containing two or more genes annotated with DNA replication initiation, contain homologues in nearly all species.

Finally, on the page “Gene families associated with a GO term,” clicking the most abundant family, i.e., HOM03D000391 ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/gene\\_families/view/HOM03D000391](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/gene_families/view/HOM03D000391)), containing mini-chromosome maintenance proteins, and selecting “View... the genome wide organization of this gene family,” allows to rapidly explore the physical location of these genes on the *Arabidopsis* genome (or any other species, when adjusting the species in the WGMapping tool). For example, when changing for this gene family, the species to *Glycine max*, the high fraction of block duplicates (12/18, or 66%) indicates that large-scale or whole-genome duplication played an important role in the increased copy number of this family in soybean.

## 3.2 Analysis of Hormone-Specific *Arabidopsis* Marker Genes Using the PLAZA Workbench

The PLAZA Workbench makes it possible for users to efficiently analyze multiple genes stored in an experiment (*see* **Note 2**). Apart from various import and export options, different functional and comparative analyses can be performed starting from a Workbench experiment. Apart from browsing gene families and gene functions associated with a specific experiment, also GO enrichment analysis can be performed to identify overrepresented functions. Furthermore, gene sets, families, and functions can be compared between different experiments, making it possible to ask more complex biological questions and generate new hypotheses.

### 3.2.1 Upload Genes as New Workbench Experiments

Starting from the sets of hormone-specific *Arabidopsis* genes described by Nemhauser, Hong, and Chory [15] (*see* Subheading 2.5), we will generate two Workbench experiments covering genes transcriptionally modulated after IAA treatment (denoted IAAup and IAAdown). First, go to the PLAZA 3.0 Dicots main page ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/)), and select from the top menu Analyze—Workbench. Next, click the Register button and complete the required fields, including an active e-mail address. Next, use your e-mail address and the received password to log in into the PLAZA Workbench. To create a new Workbench experiment, execute the following steps:

1. In the box “Add new experiment,” enter the experiment name “IAAup” and click “Create experiment.”
2. In the box “Current Experiments,” select experiment “IAAup.”
3. In the Actions box, select “Import using gene identifiers.”
4. From the Excel file (Subheading 2.5), select the genes from column “IAAup” and select “Copy” (Ctrl-C).
5. Paste the list of *Arabidopsis* gene identifiers in the input box and select “Import genes.”
6. A set of 198 *Arabidopsis* genes is now stored in the IAAup Workbench experiment.

Repeat these steps to generate the experiment “IAAdown” (*see* Subheading 2.5 for details).

### 3.2.2 Identify Gene Families Associated with IAA-Regulated Genes

Starting from the IAAup Workbench experiment, “View... associated gene families” allows exploring the regulated genes using their corresponding gene families. After sorting the families by clicking the “#associated genes” in the table header, we observe two families (HOM03D000122 and HOM03D000031) containing more than ten IAAup genes, while most families contain only one IAAup gene. Note that this information can also be graphically displayed, using the button “View bar charts” below the table. Using the mouseover function in the PLAZA website reveals that the HOM03D000122

family contains 13 AUX/IAA proteins. Apart from exploring families and functions, it is also possible to study the evolutionary conservation of the IAA-regulated genes in different plant clades, such as flowering plants, mosses, and algae, based on the presented homology information. For example, exploring the phylogenetic profile of the HOM03D000122 family, containing 13 IAAup genes, using this table reveals that it has homologues in all species apart from the two algae *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus* (cre and olu, respectively). By clicking the column “species” and sorting ascending, it becomes clear that seven IAAup genes are part of families only containing Brassicaceae homologues. When analyzing the phylogenetic profiles of the families associated with IAAdown genes (via the IAAdown Workbench experiment), all associated gene families are present in 22 or more species, indicating that despite the fact that most IAA-regulated genes are part of families present in most flowering plants (e.g., having homologues in dicots and monocots), also several IAA-upregulated genes are evolutionary more recent.

**3.2.3 Functional Analysis**  
*Using GO and InterPro*  
*Annotations: Identifying*  
*Hormone-Response*  
*Regulators in the IAAup*  
*and IAAdown Experiment*

Apart from browsing the functions of the associated gene families for the IAAup genes, the option “View... the associated functional annotation” in the Toolbox makes it possible to study gene functions using InterPro and GO annotations. The upper table reports the associated InterPro data and makes it possible to further subdivide the IAAup genes based on protein domain information, while the lower table covers associated GO data. To analyze, e.g., the different transcription factors present in the IAAup data set, the following steps can be applied:

1. Use the browsers search function (Ctrl-F) and search for “transcription factor” (TF). Matches will be found in both the InterPro and the GO table.
2. Look for the GO entry “GO:0003700—sequence-specific DNA binding transcription factor activity” reporting that 42 genes (or 21 % of the IAAup genes) are annotated with this term.
3. Follow the link behind the “42” genes and select, in the new window, the Toolbox option “Create new experiment from this subset or add subset to existing experiment.” Generate a new experiment, called IAAupTF, containing the selected 42 genes.
4. Go back to the Workbench main screen, and select the experiment “IAAupTF” and again select “View... the associated functional annotation.” The table with associated InterPro data now summarizes the different types of TFs present, including 14 genes with an AUX/IAA protein domain and eight AP2/ERF domain proteins.
5. Repeat the **steps 1–3** to generate a new Workbench experiment containing IAAdown TFs (called IAAdownTF,  $n=9$  genes).



Note that, based on the associated functional annotation (InterPro data), no AUX/IAA protein domain-containing TFs are present in IAA<sub>down</sub>TF, revealing a functional separation of up- and downregulated transcription factors during auxin response.

### 3.2.4 Gene Ontology Enrichment Analysis

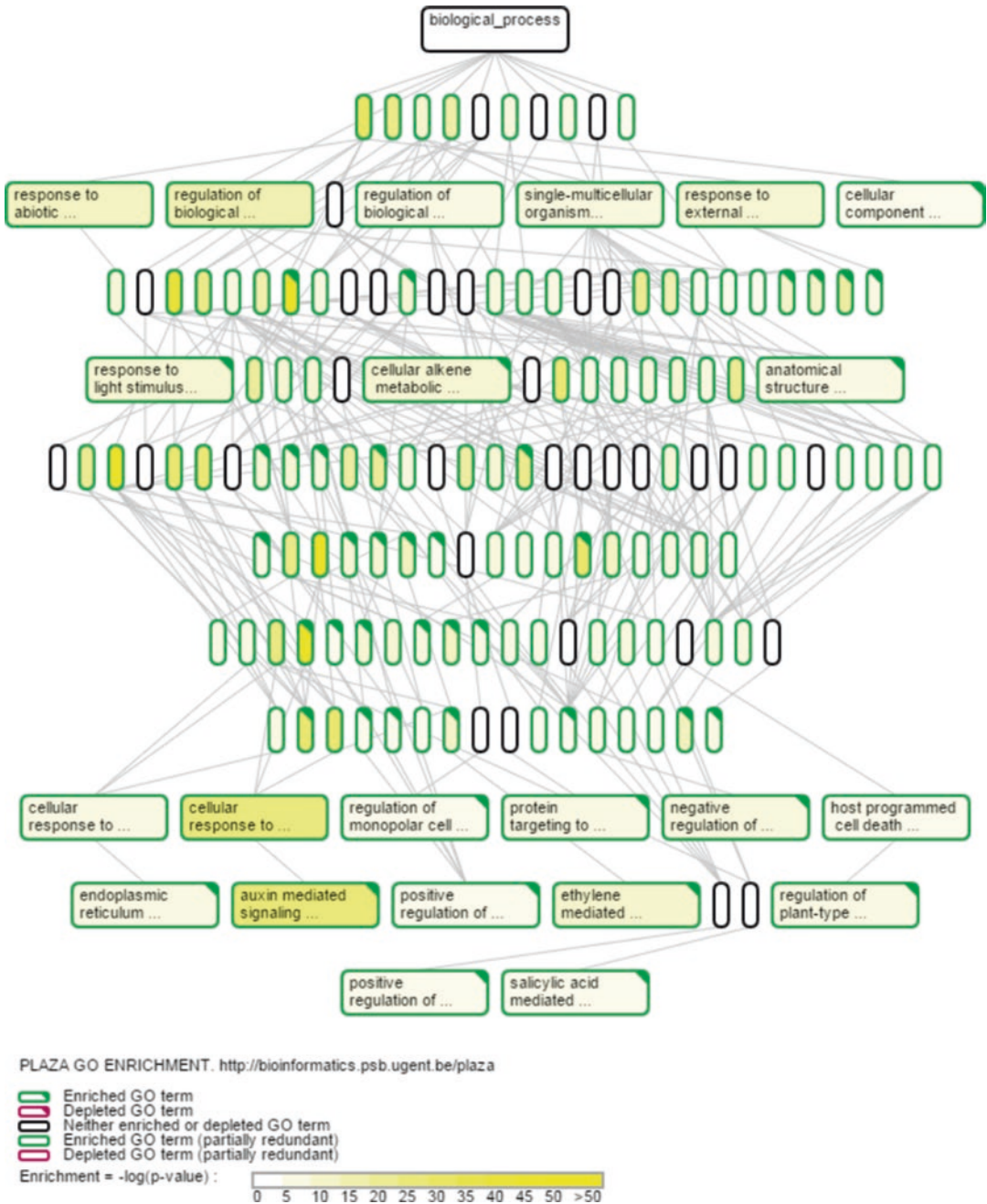
Because exploring the associated functional annotations for Workbench experiments with many genes can result in a large table with many InterPro domains and GO terms, the analysis of over-represented functional terms using enrichment analysis offers an efficient alternative to explore gene functions. To perform a GO enrichment analysis and investigate the obtained results, execute the following steps:

1. Starting from the Workbench main screen, select the experiment “IAA<sub>up</sub>” and select “View... the GO enrichment.”
2. After all calculations are ready, three main results sections are shown:
  - (a) A toolbox with links to enrichment graphs and a download option
  - (b) Interactive bar chart displays for GO types molecular function (MF), biological process (BP), and cellular component (CC)
  - (c) A GO enrichment data table

Whereas the GO enrichment table depicts the different GO-type fold enrichment values (in log<sub>2</sub> scale) and statistical significance values (Bonferroni-corrected *p*-values), the enrichment graphs summarize the overrepresented GO terms and use different color codes to annotate enriched or depleted functional terms (Fig. 2). The bar charts display, again per GO type, the log<sub>2</sub> fold enrichments and associated *p*-values (bars and black line, respectively). Whereas the graphs offer an intuitive graphical overview on the overrepresented GO terms, the data table is most useful to further dissect gene functions, for example, by isolating specific gene sets and storing them in a new Workbench experiment.

To study the role of IAA<sub>up</sub> genes in organ development and to identify how they mechanistically contribute to this process, apply the following steps:

1. From the GO enrichment output, search for development-related GO terms using the browser search function (Ctrl-F), and search for “development.”
2. Identify “organ development,” click on the subset ratio values (20.73%, which indicates that one fifth of these IAA<sub>up</sub> genes are annotated with this GO term), and select the Toolbox option “Create new experiment from this subset.” Save these 40 genes in a new experiment with name “IAA<sub>up</sub>\_organ\_dev.”



**Fig. 2** Gene Ontology enrichment graph for *Arabidopsis* genes upregulated after IAA treatment. Graphical overview of the functional enrichment analysis executed on the IAAup genes using the Gene Ontology biological process terms. Boxes filled with *yellow* indicate significantly enriched functions ( $p$ -value cutoff 0.001). Collapsed *yellow* boxes represent significant GO terms for which a more specific and significant GO term is displayed

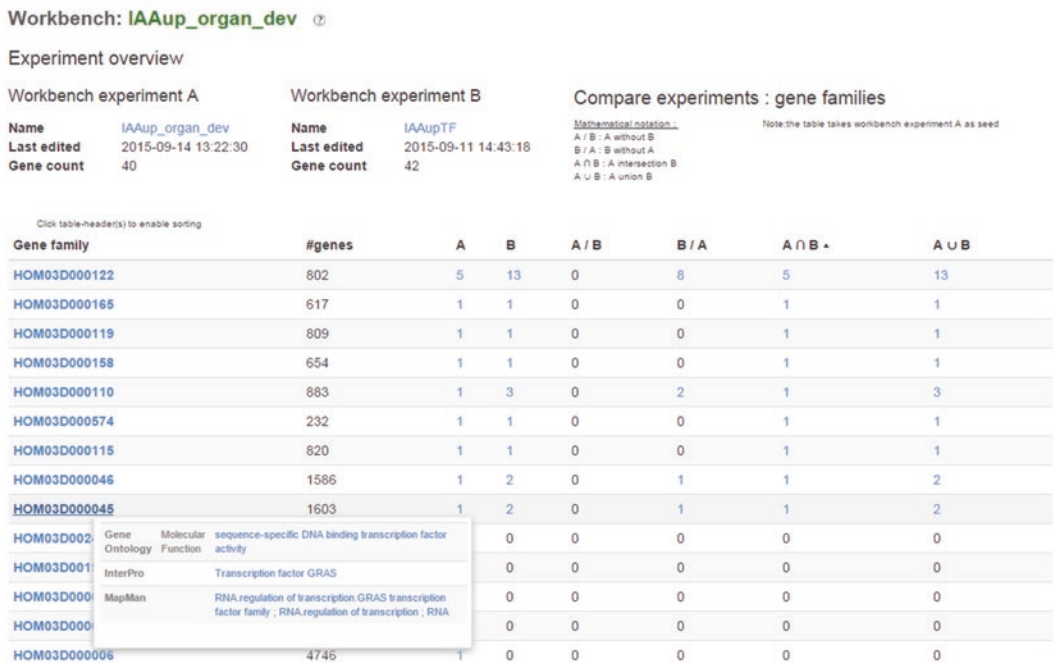
3. For the new experiment “IAAup\_organ\_dev,” apply the GO enrichment procedure, and study the overrepresented GO molecular function terms (*see* Table 1).
4. Apart from mechanisms related to plant hormone biology like “auxin efflux transmembrane transporter activity” and “cytokinin dehydrogenase activity,” 32.50% of all these genes are annotated as transcription factors (“sequence-specific DNA binding transcription factor activity,” 2.36 log<sub>2</sub> fold enrichment). This result reveals that auxin has a strong impact on transcriptional control of plant development [17].

**Table 1****GO molecular function (MF) enrichment data table for experiment “IAAup\_organ\_dev”**

GO type	GO term	Log <sub>2</sub> enrichment <sup>a</sup>	p-value	Subset ratio	Description	Shown
MF	GO:0003700	2.36	1.62E-06	32.50%	Sequence-specific DNA binding transcription factor activity	V
MF	GO:0001071	2.35	1.64E-06	32.50%	Nucleic acid binding transcription factor activity	X
MF	GO:0046983	3.03	1.28E-05	20.00%	Protein dimerization activity	V
MF	GO:0005515	1.03	3.54E-04	55.00%	Protein binding	X
MF	GO:0010329	5.82	0.0017	5.00%	Auxin efflux transmembrane transporter activity	V
MF	GO:0005516	3.2	0.0026	10.00%	Calmodulin binding	V
MF	GO:0080161	5.32	0.0034	5.00%	Auxin transmembrane transporter activity	X
MF	GO:0010487	9.41	0.0044	2.50%	Thermospermine synthase activity	V
MF	GO:0015562	4.88	0.01	5.00%	Auxin efflux transmembrane transporter activity	X
MF	GO:0016768	8.41	0.01	2.50%	Spermine synthase activity	V
MF	GO:0019139	6.6	0.03	2.50%	Cytokinin dehydrogenase activity	V
MF	GO:0005102	3.55	0.04	5.00%	Receptor binding	V
MF	GO:0015291	2.53	0.04	7.50%	Secondary active transmembrane transporter activity	V
MF	GO:0004714	5.95	0.05	2.50%	Transmembrane receptor protein tyrosine kinase activity	V

<sup>a</sup>Log<sub>2</sub> enrichment refers to the overrepresentation of a GO term and is calculated by taking the log<sub>2</sub> of the subset ratio (frequency of the GO term in the Workbench experiment) divided by the genome ratio (frequency of the GO term in the genome)

5. Although we again could create a new experiment from this functional subset, the PLAZA Workbench makes it possible to directly compare, e.g., the “IAAup\_organ\_dev” genes with those from the “IAAupTF” experiment.
6. Starting from the Workbench main screen, select the experiment “IAAup\_organ\_dev” (40 genes), and select in the Toolbox “Compare... with other PLAZA workbench experiments.” Select as second experiment “IAAupTF” (42 genes). In the box “Select comparison mode,” select “Compare... genes.”
7. In the resulting output table, select “ $A \cap B$ ” to see the 13 genes matching both criteria, i.e., IAAup genes involved in organ development and having transcription factor activity.
8. Finally, by repeating **step 5** but now selecting “Compare... gene families” in the box “Select comparison mode,” we can easily identify the 9 gene families associated with the 13 genes found in both experiments. Note that clicking the headers allows sorting the table, whereas performing a mouseover on the gene family identifier shows the associated functional annotations (Fig. 3).



**Fig. 3** Compare two Workbench experiments at the gene family level. The mouseover shows the functional annotation for gene family HOM03D000045, which codes for GRAS transcription factors. The gene shared between the experiments “IAAup\_organ\_dev” and “IAAupTF,” which is part of the GRAS transcription factor family, is LOM2 (LOST MERISTEMS 2, AT3G60630), a gene experimentally characterized as being involved in maintenance of shoot apical meristem identity and root hair cell tip growth ([http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/genes/view/AT3G60630](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/genes/view/AT3G60630))

### 3.3 Translating Biological Information from Model to Crop Using Integrative Orthology

Whereas the PLAZA gene families make it possible to study the conservation of specific genes in different plant clades, they also make it possible to identify orthologs in other species. Although orthologs are strictly defined as homologues separated by a speciation event [18], orthologs are frequently used to search for genes with conserved functions in different species. In plants, utilization of orthology is not trivial, due to a wealth of paralogs (homologous genes created through a duplication event) in almost all plant lineages [19]. The frequent whole-genome duplications in several lineages result in the establishment of one-to-many and many-to-many orthologs (or co-orthologs).

#### 3.3.1 Exploring Integrative Orthology for a Single Gene

Starting from the gene AT4G17350, which encodes a plant protein of unknown function DUF828 and which is part of the IAAup experiment, clicking the Toolbox link “Explore... the orthologs using the Integrative Orthology Viewer” on the gene page opens the integrative orthology viewer (*see Note 3*). Browsing the information for the rice species *Oryza sativa* ssp. *Japonica* in the ortholog table lists the different methods supporting the predicted orthologous genes. Clicking the diamond in the right column opens the integrative orthology viewer which graphically depicts the different inference methods. The overview reports that rice gene OS02G44040 is predicted to be an ortholog by the four integrated methods (Fig. 4). Furthermore, a second *Arabidopsis* gene is displayed, indicating that many-to-many orthology exists between these *Arabidopsis* and rice genes. Selecting a candidate rice ortholog, by clicking the diamond symbol, activates links in the Linkout box containing more information about the different methods. For example, clicking the rice ortholog OS02G44040 and following the link “Tree-based ortholog—More information” opens the phylogenetic Tree Explorer for the associated gene family HOM03D000709 (*see Note 4*). Using the zoom-X and zoom-Y options in the *Archaeopteryx* tree viewer and disabling the “dynamic hiding” option makes it possible to identify the sub-tree containing the gene AT4G17350 and its orthologs in monocots, including the rice genes OS02G44040 and OS10G41060, both identified as tree-based orthologs.

#### 3.3.2 Identifying Rice Orthologs for the *Arabidopsis* IAAup Genes

Within the PLAZA Workbench, the integrative orthology method is also present to efficiently identify orthologs for a large set of genes. Using the protocol below, we will first identify orthologs in *O. sativa* ssp. *Japonica* starting from the *Arabidopsis* genes in the IAAup experiment and subsequently compare this with auxin-responsive genes that were experimentally determined [16].

1. Starting from the Workbench main screen, select the experiment “IAAup” and select “View... the orthologous genes using the PLAZA integrative method.”







2. Select as target species “*O. sativa* ssp. *Japonica*.”
3. Use the default settings that will report all types of orthologous relationships considering all evidence types. Also keep the default setting of minimum one required evidence type. Select “Retrieve genes.”
4. The resulting integrative orthology table reports for each *Arabidopsis* gene the orthologous rice gene including a graphical overview of the evidence types (legend at the bottom of the page). Note that in some cases, multiple rice orthologs can be predicted, whereas for some *Arabidopsis* genes, no rice orthologs can be found. The latter scenario holds for gene AT1G64405, which is part of a Brassicaceae-specific gene family (see [http://bioinformatics.psb.ugent.be/plaza/versions/plaza\\_v3\\_dicots/gene\\_families/view/HOM03D010987](http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/gene_families/view/HOM03D010987)).
5. Select the “Download results” button, save the tab-delimited text file, and open this file in a spreadsheet application like Microsoft Excel.
6. Select the rice genes from column B “Orthologous\_genes,” and store these in a new Workbench experiment called “IAAup\_rice\_orthologs” (148 genes). Note that browsing the associated gene families returns a list of families very similar to

Integrative Orthology Viewer: **AT4G17350** (*Arabidopsis thaliana*) 





Mapping organism  
Orthologs overview

Oryza sativa ssp. japonica  
[Return to orthologs overview](#)

Orthology Overview

	OS02G44040	OS10G41060	OS10G41860
AT4G17350			
AT5G47440			

Legend

-  Tree-based orthology
-  Orthologous gene family
-  Anchor point
-  Best-Hits-and-Inparalogs(BH)family

Linkout

- [Tree-based orthology: More information](#)
- [Orthologous gene family: More information](#)
- [Anchor point: More information](#)
- [Best-Hits-and-Inparalogs\(BH\)family: More information](#)

AT4G17350 vs. OS02G44040

**Fig. 4** Integrative Orthology Viewer. Starting from the query gene AT4G17350, the orthologs in *Oryza sativa* are displayed together with the evidences from the different orthology prediction methods reported in the legend. For the selected AT4G17350-OS02G44040 orthologous gene pairs, additional information is available via the Linkout box to explore detailed orthology information per method

the ones identified in Subheading 3.2.2, apart from families containing Brassicaceae-specific IAAup genes, which are logically absent in rice.

Finally, we will compare how well these rice orthologs overlap with a set auxin-upregulated genes obtained through microarray transcript profiling.

1. Starting from the rice auxin-responsive gene set (*see* Subheading 2.6), create a Workbench experiment “rice\_IAAup” containing the upregulated genes (210 rice genes after processing). *See* Note 5 for more information about importing genes in a Workbench experiment using external gene identifiers.
2. To compare these experimentally determined auxin-responsive genes with the rice orthologs from Workbench experiment “IAAup\_rice\_orthologs,” select in the Toolbox “Compare... with other PLAZA workbench experiments.”
3. After selecting “IAAup\_rice\_orthologs” as second experiment, first determine how many genes are shared between both experiments. Overall, 21 genes are shared between both experiments, which is much more than expected by chance and highly significant (hypergeometric distribution  $p < 2.63e-24$ ;  $n = 40,738$   $m = 210$   $k = 147$   $-x = 21$  bigger or equal).
4. Comparing these two Workbench experiments at the gene family level (using the “Comparison mode—Gene families”) and using the clickable header for sorting reveals that 12 families are shared. Interestingly, the HOM03D000709 family contains the unknown rice gene OS10G41060, which is part of the predicted rice orthologs (experiment “IAAup\_rice\_orthologs,” based on the *Arabidopsis* IAAup gene AT4G17350) and also belongs to the auxin-upregulated gene set (experiment “rice\_IAAup”). This comparative analysis reveals that, despite the fact that no experimental GO biological process annotations are available for OS10G41060 (nor its *Arabidopsis* ortholog AT4G17350), these genes show a conserved auxin response.

---

## 4 Notes

### 1. PLAZA gene families

All protein-coding genes are stored in gene families based on sequence similarity inferred through BLAST [20]. A gene family is defined as a group of two or more homologous genes. A graph-based clustering method (Markov clustering implemented in Tribe-MCL [10]) was used to delineate gene families based on BLAST protein similarities. Although this method is very well suited for clustering large sets of proteins derived from multiple species, high false-positive rates caused by the

potential inclusion of spurious BLAST hits have been reported [21]. Therefore, we applied a post-processing procedure by tagging genes as outliers if they showed sequence similarity to only a minority of all family members. The OrthoMCL method [11] was applied to build subfamilies based on the same protein similarity graph. Because OrthoMCL models orthology and in-paralogy (duplication events postdating speciation) based on a reciprocal best-hit strategy, the final protein clusters will be smaller than the Tribe-MCL clusters because out-paralogs (homologues from duplication events predating speciation) will not be grouped. Therefore, from a biological point of view, subfamilies or out-paralogs can be considered as different subtypes within a large protein family.

## 2. PLAZA Workbench

To analyze multiple genes in batch, we have developed a PLAZA Workbench enabling the analysis of different comparative and functional properties for user-defined gene sets. Hundreds of genes can easily be uploaded through a list of (internal or external) gene identifiers or based on a sequence similarity search. For example, this last option enables users to map an EST or assembled RNA-Seq data set from a non-model organism to a reference genome annotation present in PLAZA. For gene sets saved by the user in the Workbench, detailed information about functional annotation (InterPro and GO), associated gene families, block and tandem gene duplicates, and gene structure is provided. In addition, the GO enrichment tool allows to determine whether a user-defined gene set is overrepresented for one or more GO terms.

## 3. Integrative Orthology Viewer

The Integrative Orthology Viewer displays for a query gene and its predicted in-paralogs, the associated orthologs, including the support from the four different orthology inference methods (a BLAST-, protein clustering-, phylogenetic tree-, and collinearity-based approach). In addition, all links are provided to explore the supporting evidence and specific details of the individual predictions.

## 4. Tree Explorer

The Tree Explorer makes it possible to display and analyze the phylogenetic trees calculated for the different gene families in PLAZA. This page used the *Archaeopteryx* tree viewer, which is a Java applet and requires at least Java 1.5. To determine which version of Java your web browser is using, please visit [www.javatester.org](http://www.javatester.org). Also, make sure your browser supports NPAPI technology required for Java applets (e.g., Internet Explorer, Firefox, or Safari). By default, the Tree Explorer displays the phylogenetic tree with protein domain information. Alternatively, also gene structure information can be shown, which makes it possible to compare exon-intron structures



between closely and more distantly related homologues. Finally, the phylogenetic tree showing speciation/duplication events shows annotated tree nodes including information about speciation and duplication events.

#### 5. Conversion external gene identifiers

The PLAZA Workbench supports the creation of new experiments by importing gene identifiers from different sources. Whereas importing genes using PLAZA gene identifiers is the fastest, also external identifiers provided by the original gene annotation providers are supported (as long as the latter were made available in the export or bulk download files of these data providers). For example, rice gene LOC\_Os01g18360 will be automatically recognized by the system as PLAZA gene identifier OS01G18360. Only in case no one-to-one mapping was found between an external identifier and a PLAZA gene identifier, the user is asked to select the appropriate gene. Note that the import of genes through external identifiers is slower than using PLAZA gene identifiers and might take some minutes for large sets of genes.

---

## Acknowledgments

I thank Michiel Van Bel for excellent technical assistance and maintenance of the PLAZA platform, all long-term PLAZA users for their feedback and Annick Bleys for help in preparing the manuscript. This work was supported by the Multidisciplinary Research Partnership “Bioinformatics: From Nucleotides to Networks” Project (no 01MR0410W) of Ghent University.

## References

1. Michael TP, Jackson S (2013) The first 50 plant genomes. *Plant Genome* 6. doi: [10.3835/plantgenome2013.3803.0001in](https://doi.org/10.3835/plantgenome2013.3803.0001in)
2. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martinez-Garcia PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, de Jong PJ, Mockaitis K, Main D, Langley CH, Neale DB (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* 196:891–909
3. Vlad D, Kierzkowski D, Rast MI, Vuolo F, Dello Ioio R, Galinha C, Gan X, Hajheidari M, Hay A, Smith RS, Huijser P, Bailey CD, Tsiantis M (2014) Leaf shape evolution through duplication, regulatory diversification, and loss of a homeobox gene. *Science* 343:780–783
4. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345:1249721
5. Hardison RC (2003) Comparative genomics. *PLoS Biol* 1:156–160
6. Vandepoele K, Van de Peer Y (2005) Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol* 137:31–42
7. Proost S, Van Bel M, Vanechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res* 43:D974–D981

8. 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
9. Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K (2013) TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. *Genome Biol* 14:R134
10. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584
11. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189
12. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490
13. Pérez-Rodríguez P, Riaño-Pachón DM, Corréa LGG, Rensing SA, Kersten B, Mueller-Roeber B (2010) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res* 38:D822–D827
14. Jin J, Zhang H, Kong L, Gao G, Luo J (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 42:D1182–D1187
15. Nemhauser JL, Hong F, Chory J (2006) Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell* 126:467–475
16. Jain M, Khurana JP (2009) Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. *FEBS J* 276:3148–3162
17. Guilfoyle TJ, Hagen G (2007) Auxin response factors. *Curr Opin Plant Biol* 10:453–460
18. Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Biol* 19:99–113
19. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. *Trends Plant Sci* 14:680–688
20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
21. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2:e383