Supplementary material for manuscript "Dissecting plant genomes with the PLAZA comparative genomics platform"

Michiel Van Bel[1,2+], Sebastian Proost[1,2+], Elisabeth Wischnitzki[1,2], Sara Mohavedi[1,2], Christopher Scheerlinck[3], Yves Van de Peer[1,2] and Klaas Vandepoele[1,2,*]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium,

[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium and

[3]Department Industrial Sciences BME-CTL, University College Ghent, Ghent University Association, Schoonmeersstraat 52, B-9000 Ghent, Belgium

[*]To whom correspondence should be addressed. Tel: +32 9 3313822; Fax +32 9 3313809; Email: klaas.vandepoele@psb.vib-ugent.be

[+]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Supplementary Table 1. Gene and gene family data content for PLAZA 2.5

Species	Genes	Coding genes	RNA genes	Pseudo genes	Transposons	GO (1)	Interpro (2)	Genes in non singleton gf	Genes in multi-species gf
Lotus japonicus	69,647	43,146	45	0	26,456	19,770	24,738	25,716	25,174
Medicago truncatula	57,587	45,197	776	0	11,614	17,836	21,999	38,494	30,844
Glycine max	46,509	46,464	45	0	0	32,616	38,517	45,982	45,652
Malus domestica	95,230	63,546	0	0	31,684	47,889	44,063	58,790	55,288
Fragaria vesca	34,809	34,809	0	0	0	17,053	21,388	30,833	26,756
Manihot esculenta	30,800	30,748	52	0	0	20,434	24,222	30,132	29,965
Ricinus communis	31,221	31,221	0	0	0	16,901	20,285	24,455	23,315
Populus trichocarpa	41,521	41,476	45	0	0	25,633	30,585	37,777	36,847
Arabidopsis thaliana	33,602	27,416	1,359	924	3,903	22,874	21,467	26,118	25,932
Arabidopsis lyrata	32,670	32,670	0	0	0	21,429	23,557	30,870	29,264
Carica papaya	28,072	28,027	45	0	0	13,914	16,265	22,531	21,005
Theobroma cacao	46,269	28,882	45	0	17,342	16,653	20,039	27,575	25,306
Vitis vinifera	26,644	26,504	88	52	0	20,244	19,035	23,268	22,682
Oryza sativa ssp. japonica	57,874	42,211	92	0	15,571	26,014	24,735	37,391	37,020
Oryza sativa ssp. indica	59,430	49,202	39	0	10,189	26,518	29,345	44,310	43,609
Brachypodium distachyon	26,678	26,632	46	0	0	17,805	20,832	25,687	25,256
Sorghum bicolor	34,686	34,609	77	0	0	21,502	24,601	31,921	31,384
Zea mays	39,597	39,190	0	323	84	21,926	25,700	35,221	32,528
Selaginella moellendorffii	22,285	22,285	0	0	0	12,417	15,995	17,392	13,783
Physcomitrella patens	36,137	28,097	72	0	7,968	14,283	16,275	21,287	17,787
Ostreococcus lucimarinus	7,805	7,805	0	0	0	4,737	5,807	7,408	7,340
Ostreococcus tauri	8,116	7,994	122	0	0	3,966	5,094	6,797	6,663
Micromonas sp. RCC299	10,276	10,204	72	0	0	5,953	7,120	8,144	7,872
Volvox carteri	15,544	15,544	0	0	0	6,082	8,410	13,782	12,041
Chlamydomonas reinhardtii	16,841	16,788	53	0	0	8,509	8,973	13,666	11,796
Total	909,850	780,667	3,073	1,299	124,811	462,958	519,047	685,547	645,109

(1) # genes with at least one GO term

(2) # genes with at least one InterPro domain

Supplementary Table 2. List of rosid core gene families.
See file core_families_rosids.xlsx

Supplementary Table 3. List of monocot core gene families.
See file core_families_monocots.xlsx

Supplementary Table 4. List of green plant core gene families.
See file core_families_greenplants.xlsx

Supplementary Table 5. Species-specific gene families.
Number of species-specific gene families (and associated genes),including singletons. The lowest counts are indicated in light grey whereas the highest counts are in dark grey.

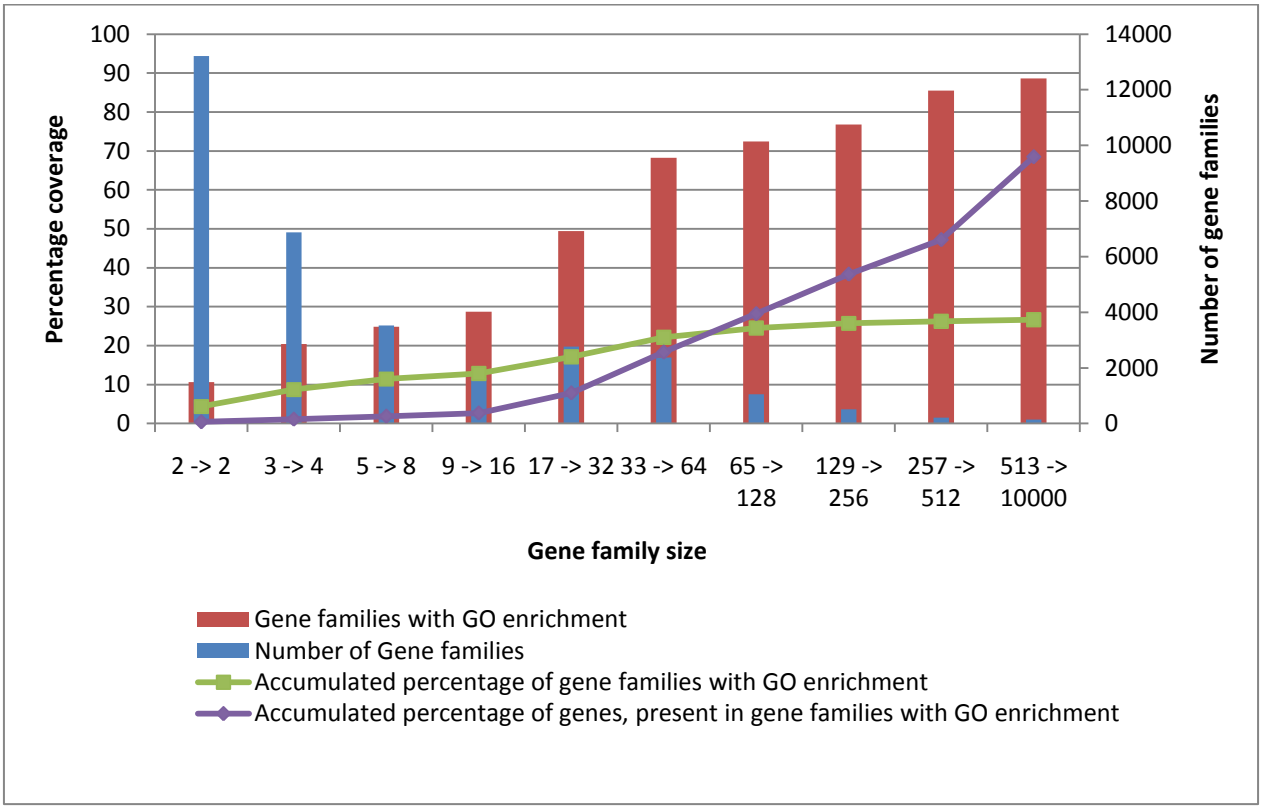
Species	Gene families	Genes
Arabidopsis lyrata	2,206	3,406
Arabidopsis thaliana	1,362	1,484
Brachypodium distachyon	1,078	1,376
Carica papaya	5,865	7,022
Chlamydomonas reinhardtii	3,562	4,992
Fragaria vesca	4,867	8,053
Glycine max	616	812
Lotus japonicus	17,540	17,972
Malus domestica	5,760	8,258
Manihot esculenta	662	783
Medicago truncatula	8,417	14,353
Micromonas sp. RCC299	2,162	2,332
Oryza sativa ssp. indica	5,154	5,593
Oryza sativa ssp. japonica	4,963	5,191
Ostreococcus lucimarinus	424	465
Ostreococcus tauri	1,240	1,331
Physcomitrella patens	7,687	10,310
Populus trichocarpa	3,983	4,629
Ricinus communis	7,186	7,906
Selaginella moellendorffii	5,662	8,502
Sorghum bicolor	2,802	3,225
Theobroma cacao	1,546	3,576
Vitis vinifera	3,431	3,822
Volvox carteri	2,034	3,503
Zea mays	4,855	6,662
Sum	105,064	135,558

Supplementary Table 6. Clade-specific gene families.
Number of clade-specific families (and associated genes). The lowest counts are indicated in light grey whereas the highest counts are in dark grey. The third column is the normalized number of genes (by taking the number of species in the phylogenetic clade into account).

Clade	Species	Gene families	Genes	Normalized genes
Arabidopsis	2	1,289	5,416	2,708
Chlamydomonadales	2	2,332	8,491	4,246
Euphorbiaceae	2	121	451	226
Galegoids	2	43	702	351
Oryza	2	5,197	14,433	7,217
Ostreococcus	2	795	1,783	892
PACCMADClade	2	302	935	468
Rosaceae	2	139	1,098	549
BEPClade	3	152	801	267
Brassicales	3	12	60	20
Malpighiales	3	48	219	73
Mamiellales	3	947	3,332	1,111
Papilionoideae	3	36	460	153
Malvids	4	1	4	1
Chlorophyta	5	134	958	192
Monocots	5	711	8,855	1,771
N2FixingClade	5	0	0	0
Fabids	8	1	15	2
Rosids	12	47	1,605	134
Eudicots	13	85	3,185	245
Angiosperms	18	202	15,105	839
VascularPlants	19	102	7,851	413
LandPlants	20	534	78,976	3,949
GreenPlants	25	1,116	200,837	8,033
Sum		143,46	355,572	

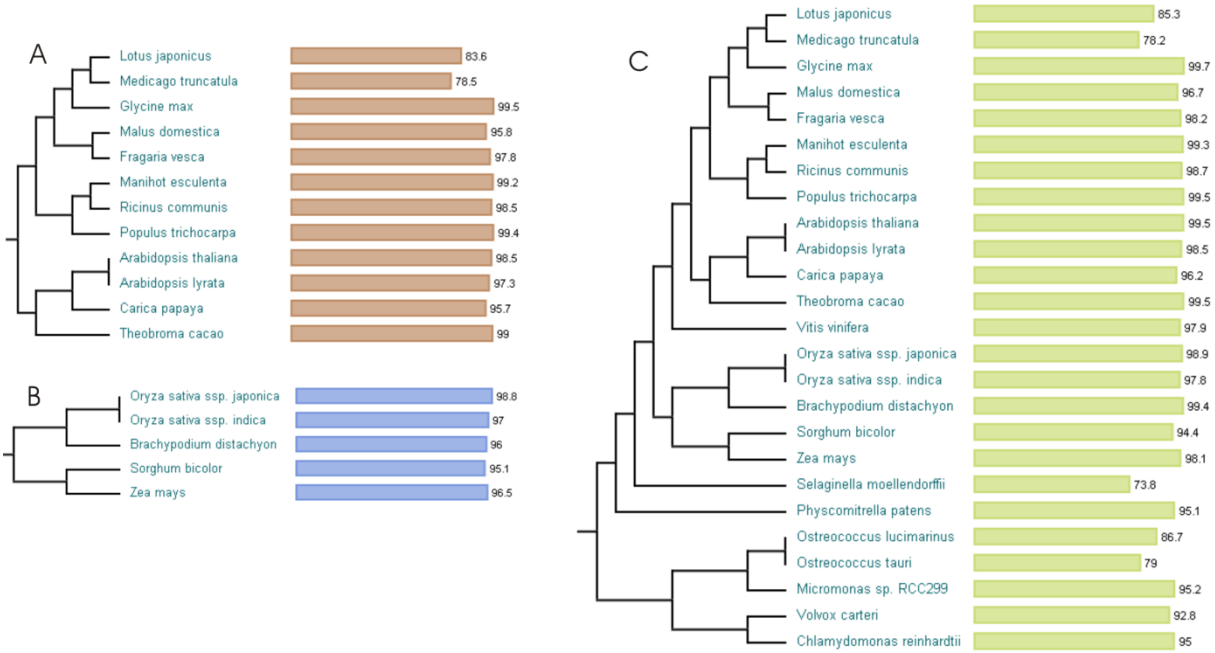
Supplementary Figure 1. Gene family coverage by GO enrichment, organized by gene family size.

Cumulative gene coverage is the combined sum of all genes that are in a gene family with GO enrichment.



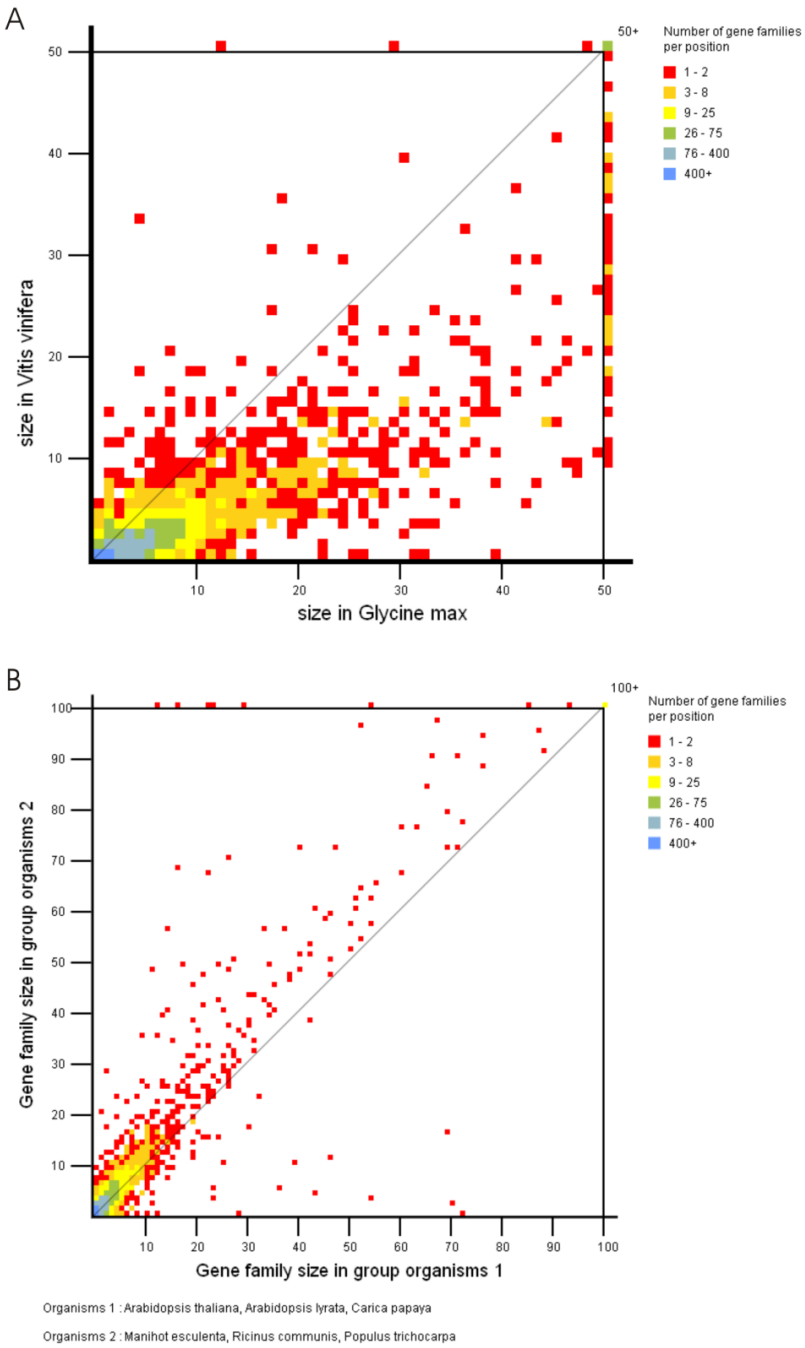
Supplementary Figure 2. Core gene family coverage.

Core gene family coverage in all PLAZA organisms, using the 6,316 rosid (A), 7,076 monocot (B), and 2,928 green plant core gene families (C). Coverage is expressed as percentage of the core gene families having the indicated organism. Data are based on Supplementary Tables S2, S3, and S4 for A, B, and C, respectively.



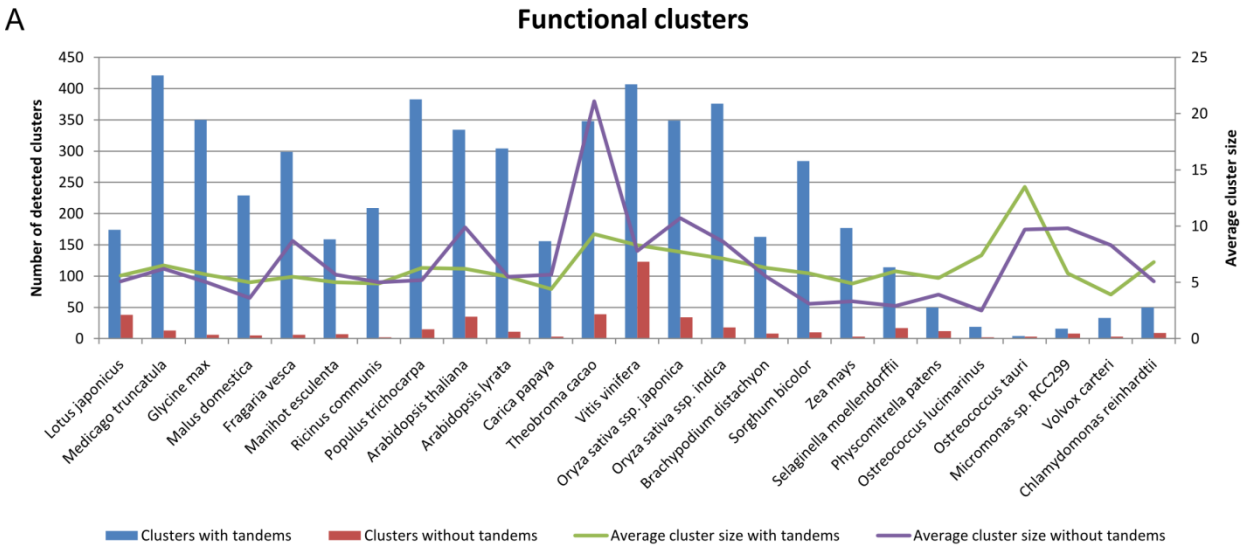
Supplementary Figure 3. Gene family Expansion Plot.

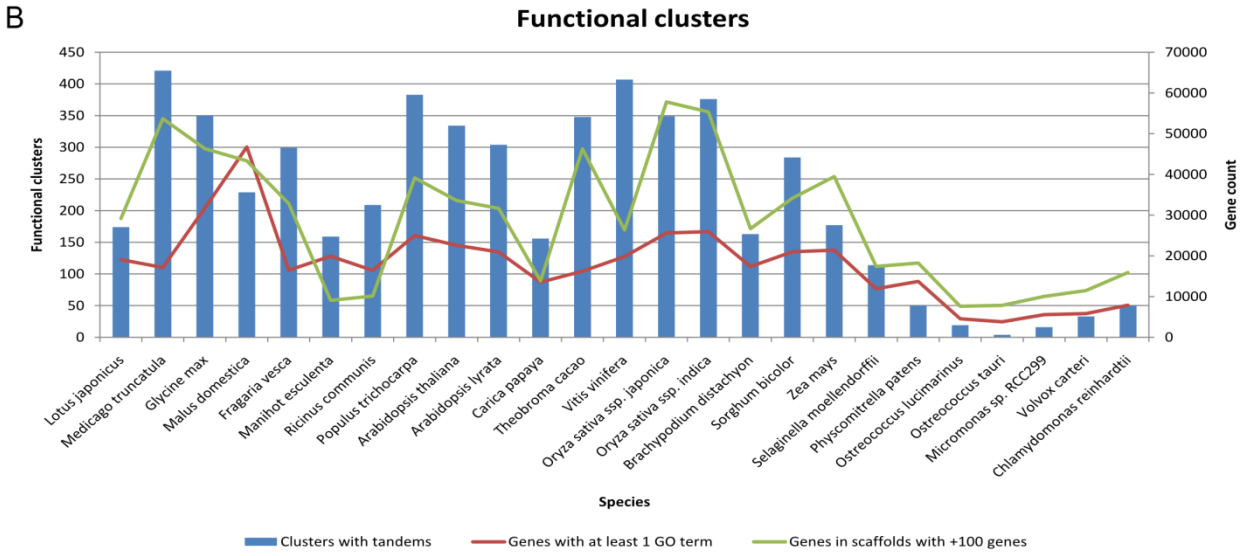
(A) The gene copy number in *Vitis vinifera* and *Glycine max*, within each gene family, is indicated by the position of a dot, whereas the color indicates the number of gene families with these gene copy numbers. (B) Density plot between two sets of organisms, Brassicales (*Arabidopsis thaliana, Arabidopsis lyrata,* and *Carica papaya*) versus Malpighiales (*Manihot esculenta, Ricinus communis,*and *Populus trichocarpa*).



Supplementary Figure 4. Summary functional clusters.

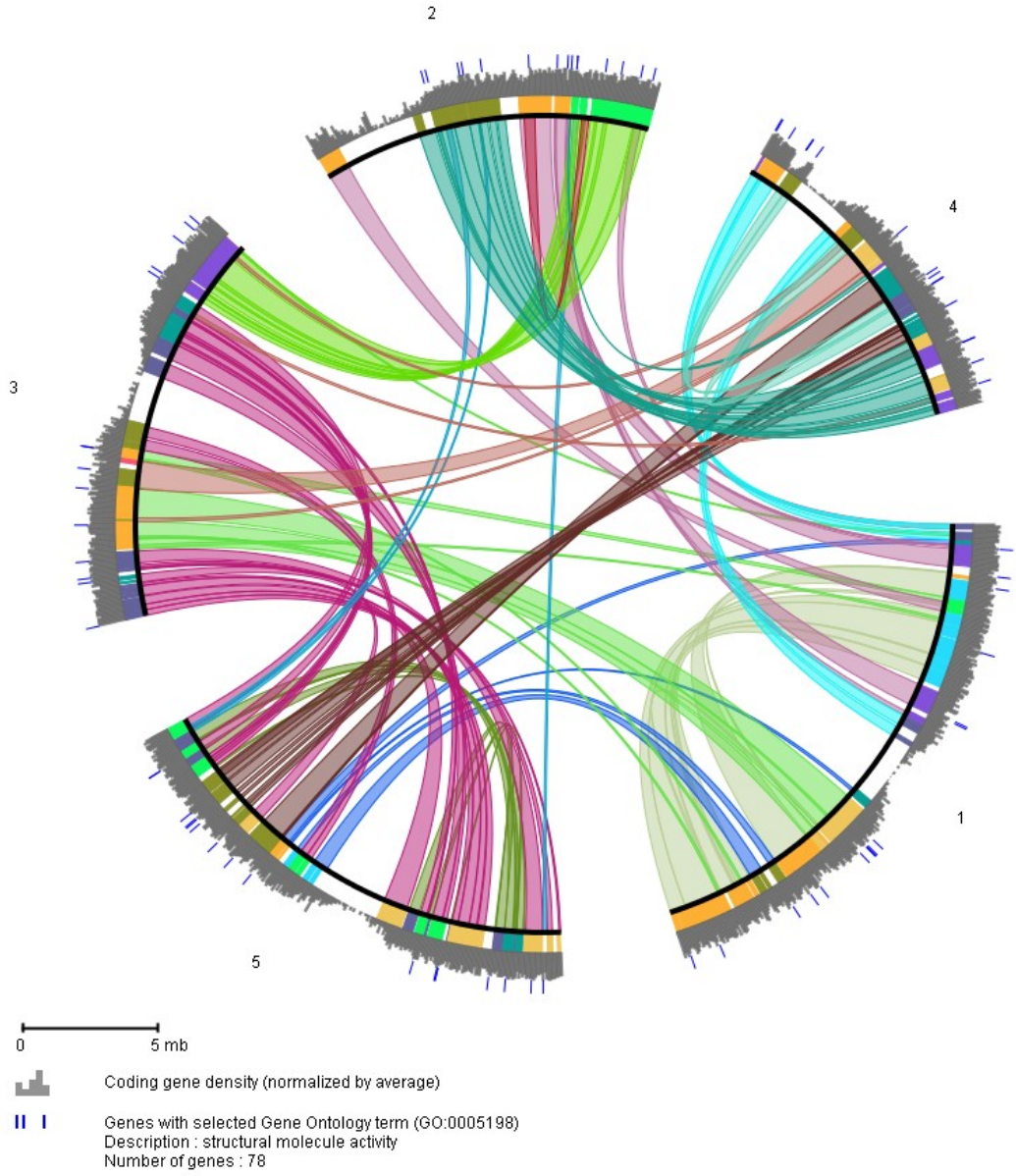
(A) Number of detected functional clusters per species with associated average gene cluster size. (B) Number of detected functional clusters per species with number of genes with at least one GO term and number of genes in large scaffolds . Species are ordered according to the PLAZA phylogenetic tree.





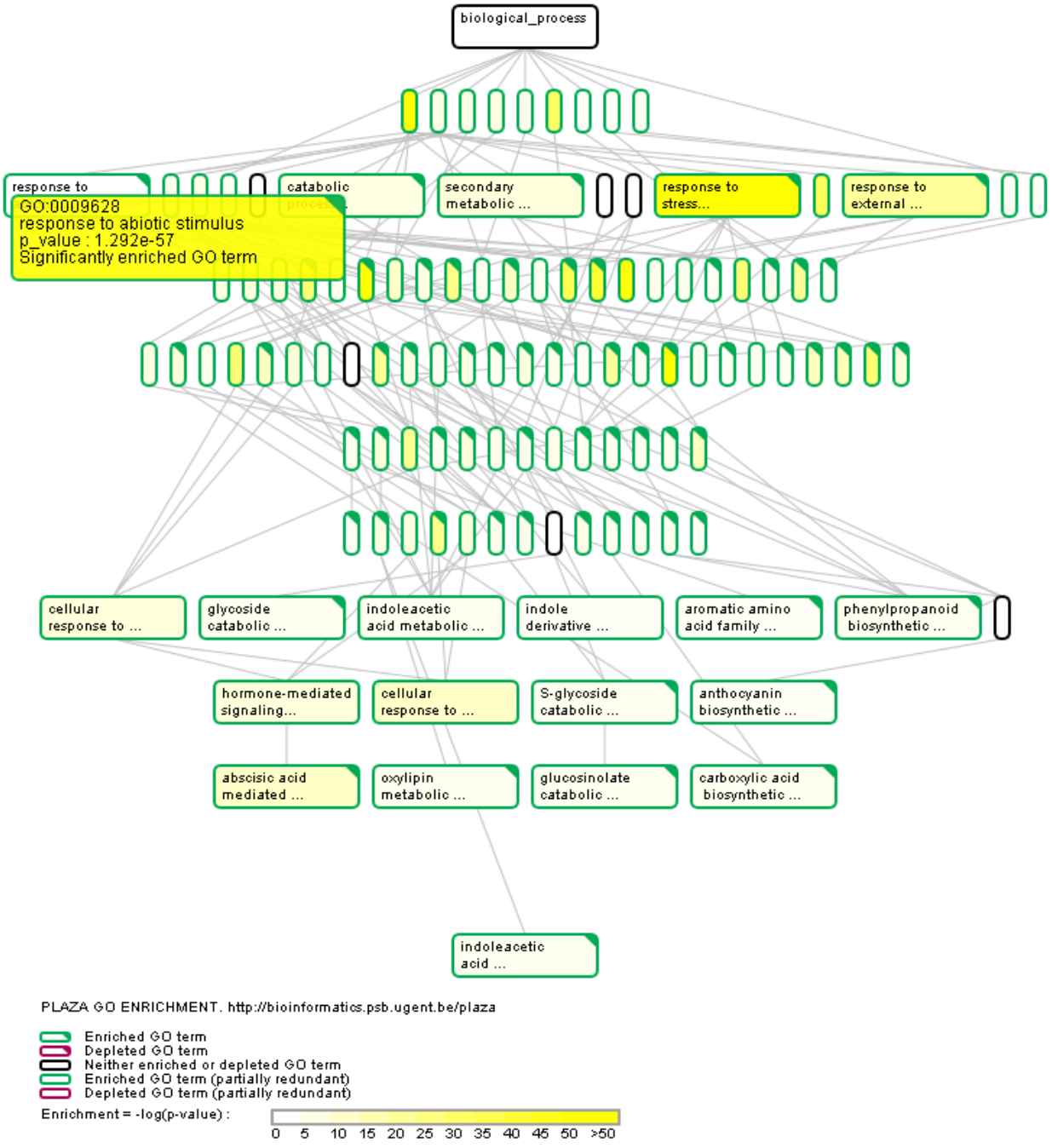
Supplementary Figure 5. Circle Plot.

Plot showing colinear regions within *Arabidopsis thaliana* (inner circle) and between *Arabidopsis thaliana* and *Arabidopsis lyrata* (colored border of circle, indicating different *Arabidopsis lyrata* chromosomes). Also displayed are the coding gene density (grey blocks on the border of the circle) and a selected GO term (GO:0005198) (blue stripes on border of the circle). Coloring of colinear regions within *Arabidopsis thaliana* is based on start/end chromosomes, and only those colinear regions (both intra- and inter-species) with a Ks-value between 0.3 and 2 (corresponding with 3R duplication event) are shown.



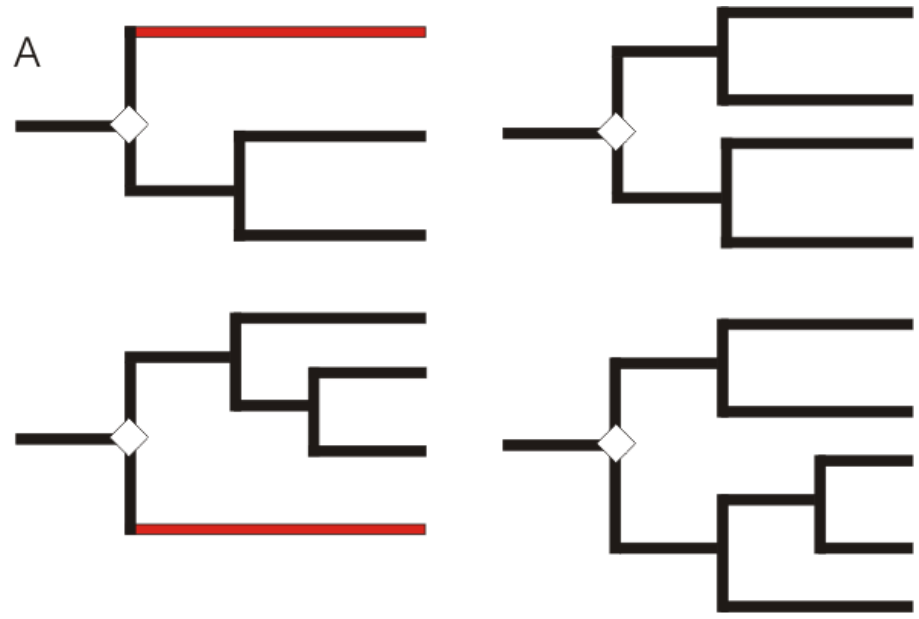
Supplementary Figure 6. Gene Ontology enrichment graph.

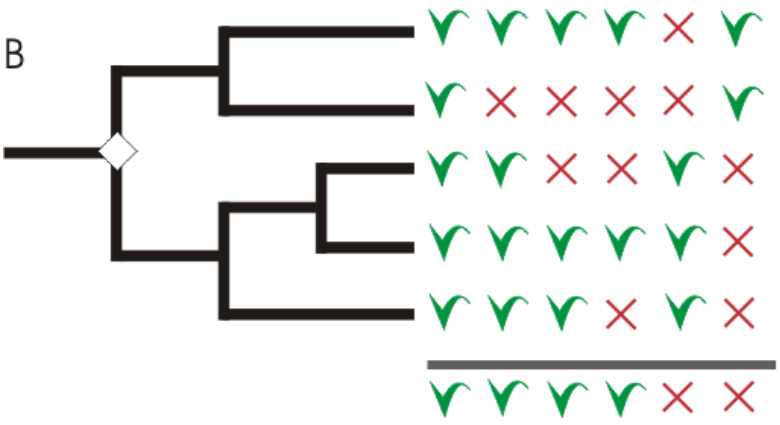
GO enrichment graph for a set of *Arabidopsis thaliana* genes that are differentially expressed under drought stress conditions. The enlarged rectangle depicting additional information about the enriched functional category, is shown when hovering over the image with the mouse pointer.



Supplementary Method 1. Selection of core gene families.

Detection of phylogenetic clades for which core gene families can be selected (A). Red branches indicate problematic subclades for which no core gene families can be determined because of the presence of a single species (leaves). (B) Indication of when the presence of a gene family gene within a phylogenetic clade that is accepted as contributing to the core gene family for that clade.

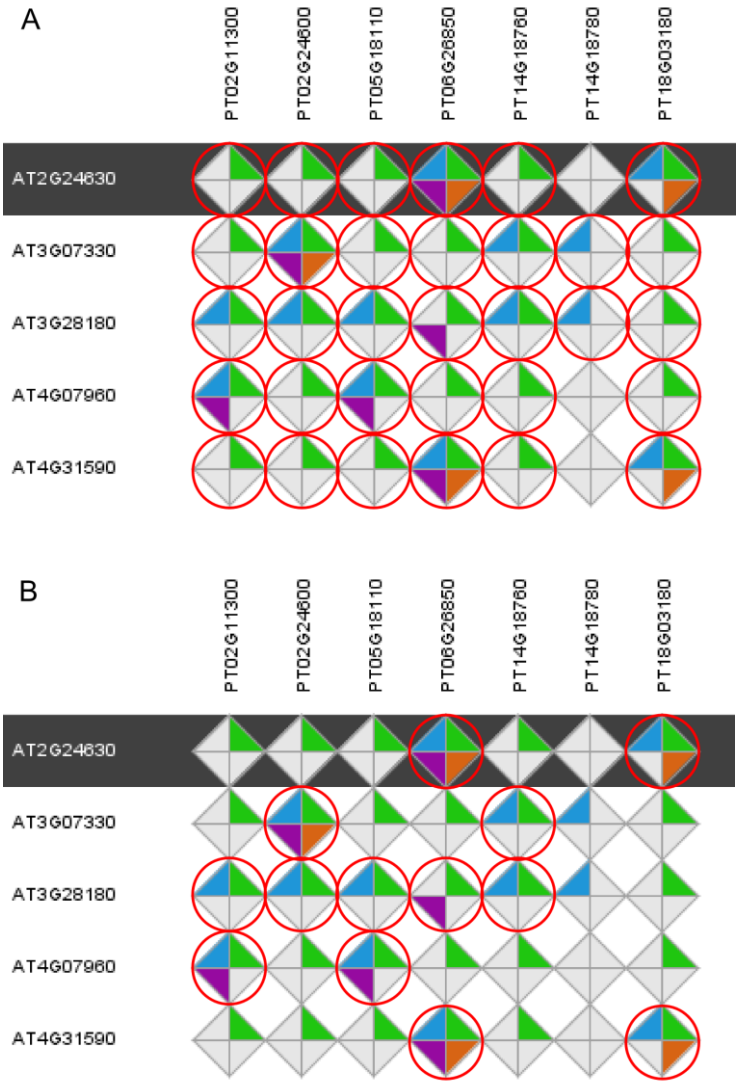




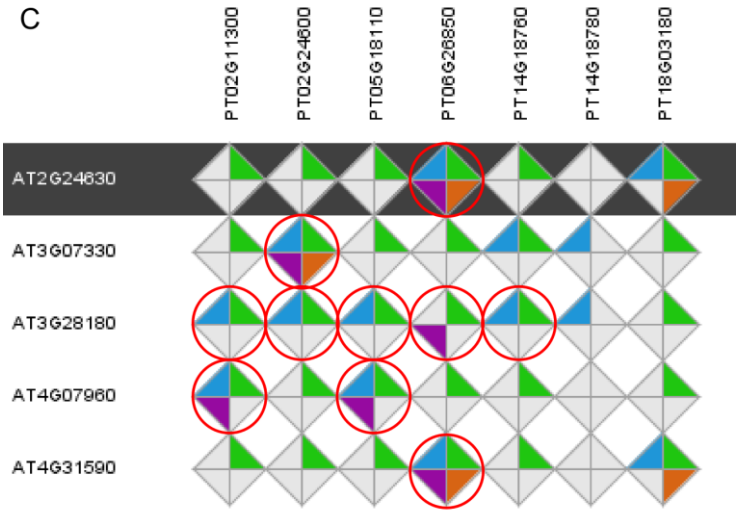
Supplementary Method 2. Different strategies for the Integrative Orthology detection.

Red Circles indicate the orthology prediction retained by a specific selection strategy. Majority-voting is a simple concept in which, per query gene, only those orthologous relations are kept that have the maximum of supporting orthologous types.

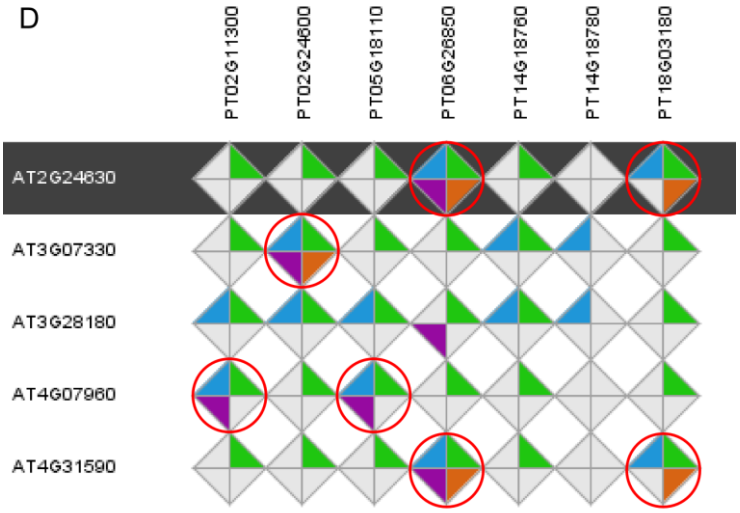
(A) The first strategy requires at least one type of orthology support, possibly leading to an overestimation of the number of orthologs. (B) The second strategy requires at least two types of orthology support. (C) The third strategy requires at least two types of orthology support, but with majority voting. (D) The fourth strategy requires at least three types of orthology support. (E) The fifth strategy requires at least three types of orthology support, but with majority voting.



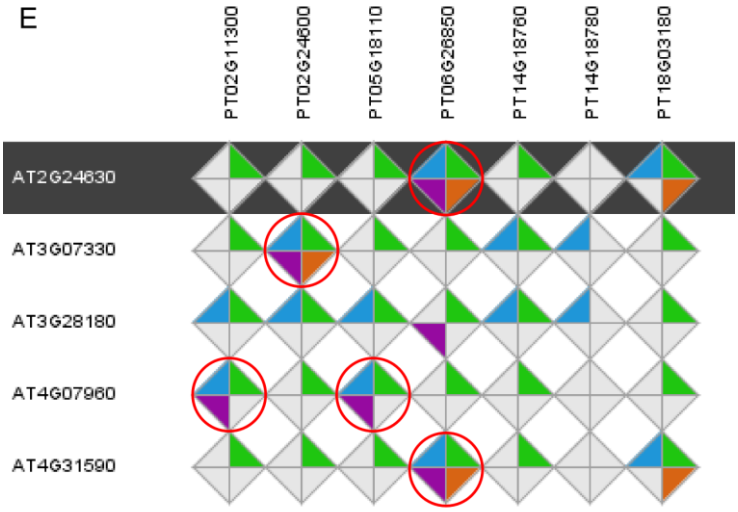
C


PT02G11300
PT02G24600
PT05G18110
PT06G26850
PT14G18760
PT14G18780
PT18G03180
AT2G24630
AT3G07330
AT3G28180
AT4G07960
AT4G31590

D


PT02G11300
PT02G24600
PT05G18110
PT06G26850
PT14G18760
PT14G18780
PT18G03180
AT2G24630
AT3G07330
AT3G28180
AT4G07960
AT4G31590

E


PT02G11300
PT02G24600
PT05G18110
PT06G26850
PT14G18760
PT14G18780
PT18G03180
AT2G24630
AT3G07330
AT3G28180
AT4G07960
AT4G31590