Supplementary Information

Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants

Riet De Smet[1,2], Keith L. Adams[1,3], Klaas Vandepoele[1,2], Marc C. E. Van Montagu[1,2,*], Steven Maere[1,2], Yves Van de Peer[1,2,*]

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

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

[3] Department of Botany, University of British Columbia, 6270 University Blvd, Vancouver, BC, V6T 1Z4, Canada

* Corresponding author

SI 1. Verification of single-copy status in *Brassica rapa*, *Solanum lycopersicum* and *Musa accuminata*, three species not included in the PLAZA dataset

Protein coding sequences of all annotated *Brassica rapa* (Chinese cabbage), *Solanum lycopersicum* (tomato) and *Musa accuminata* (banana) genes were retrieved: for *Brassica rapa* from the Brassica Database (http://solgenomics.net/organism/Solanum_lycopersicum/genome), for *Solanum lycopersicum* from the Sol Genomics Network (http://solgenomics.net/organism/Solanum_lycopersicum/genome) and for *Musa accuminata* from the Banana genome consortium (http://banana-genome.cirad.fr/). A BLAST database was constructed based on all protein coding genes of the 17 angiosperm genomes in PLAZA 2.5 augmented with the *Brassica rapa*, *Solanum lycopersicum* and *Musa accuminata* genes. An all-against-all BLAST analysis was run for these 20 genomes using BLASTP. The output of this BLAST analysis was used to assign *Brassica rapa*, tomato and banana genes to the OrthoMCL orthologous groups (OGs) and to distinguish inparalogs from outparalogs in these genomes. First, for the mapping of the *Brassica,* tomato and banana genes to the OGs we considered for each query-gene in the respective genomes the top-scoring BLAST hits and calculated for the top-scoring BLAST hits the overlap with the existing OGs using the Jaccard coefficient (intersection/union). The query-gene was assigned to families that had a Jaccard coefficient exceeding 0.5. As such we could map 27,644 out of the 41,019 *Brassica rapa* genes to OGs, 20,537 out of the 34,727 tomato genes and 17,386 out of the 36,549 banana genes. To derive whether the *Brassica*, tomato and banana genes were single-copy in the families to which they were assigned we used a strategy similar to the Inparanoid method (1). If for a certain *Brassica*, tomato or banana query gene, a gene from the same genome ranked higher in the BLAST output than genes from the other genomes, it was considered to be an inparalog, otherwise it was considered to be an outparalog. Using this strategy we could assign 2610 *Brassica* genes, 2722 tomato genes and 2616 banana genes to (mostly) single-copy OGs. For tomato the large majority (83% or 2267 out of 2722) of the genes assigned to these (mostly) single-copy OGs were single-copy, for *Brassica rapa* this was 66% (1734 out of 2610) and for banana this was 71% (1848 out of 2616) (Fig. S1). The lower percentage of single-copy genes identified for *Brassica rapa* and banana can be explained by the higher number of WGD events that has occurred in these organisms (2, 3).

Motivation for the choice of Jaccard threshold

For the assignment of the *Brassica rapa*, tomato and banana genes to the OrthoMCL OGs we rely on a cut-off on the Jaccard coefficient to decide whether the query-gene shows BLAST-similarity to a sufficient number of genes within the OGs. We aim to set this cutoff sufficiently high to ensure that the query-gene shows similarity to most of the genes in the OrthoMCL OGs. On the other hand, a too strict threshold might result in almost none of the genes being assigned to any of the OGs. To decide on what specific value of the cutoff to use we argued that since the single-copy genes are generally

well-conserved, with representatives in almost all of the 17 angiosperm genomes in the OGs, these genes will also most likely have representatives in the three additional angiosperm genomes (therefore not necessarily as a single-copy). For a Jaccard coefficient exceeding 0.5 we observe that the number of single-copy OGs with representatives in any of these 3 genomes start to drop more drastically (Fig. S2, the combined bar of single-copy and not single-copy genes). Hence, we choose this Jaccard coefficient threshold to assign genes to OGs.

SI 2. Phylogenetic approach to verify single-copy status

To validate the single-copy status of the obtained orthologous groups and to assess the possibility that paralogs were erroneously excluded from the OGs which would lead to false positive predictions, we developed a phylogenetic approach. First, we expanded all the single-copy OGs (strictly and mostly) with genes that might have been erroneously excluded: i.e. genes with a high BLAST similarity score to the OG genes that were not included in the OGs themselves. For each of the obtained expanded OGs, gene trees were constructed as follows. Multiple sequence alignments of protein sequences were constructed for each orthologous group using the MUSCLE method (4). Poorly aligned and divergent regions were removed from the multiple sequence alignment using the same customized scripts developed to process the multiple sequence alignments of the PLAZA 2.5 families (5). From these alignments phylogenetic trees were constructed based on the maximum likelihood (ML) approach using the PhyML software (6). A neighbor-joining tree as constructed by BioNJ was used as starting topology. This tree topology was optimized in a maximum likelihood framework using the WAG evolutionary model. The final topology was obtained from 100 bootstrap samples. NOTUNG 2.6 was used to root the trees, using the same species tree as in (7).

The phylogenetic tree of each of these 'expanded OGs' was analyzed to assess whether the expanded gene set formed an outparalogous cluster (i.e. they derived from a duplication which took place before the angiosperm ancestor arose) with respect to the original OG or included some previously missed inparalogs (Fig. S3). In the former case the predicted OG was assumed to be a true strictly or mostly single-copy OG as its single-copy status could be confirmed by the phylogenetic validation step. In case of the latter (missing inparalogs), if the gene copy number in the expanded OG conformed to the initial criteria (no duplicates for more than three species) the OG was assumed to be an mostly single-copy OG, otherwise it was classified as being an invalid single-copy OG and removed from the analysis (see Methods for details). In total, for 2840 (2663 mostly single copy + 177 strictly single-copy) of the 3880 initially observed strictly and mostly single-copy OGs, the single-copy status could be confirmed.

SI 3. Simulations show that the number of identified single-copy OGs exceeds random expectation

To determine whether the observed number of single-copy families could have been obtained by random gene duplication/loss dynamics, we first determined the fraction of gene loss and duplication along each branch of the species tree. The species tree used was the same as the one in

(7). Using this species tree, for each PLAZA OG we performed gene tree-species tree reconciliation using NOTUNG2.6 (8) to derive the predicted number of duplication and loss events along each branch. This procedure was repeated for each PLAZA OG dating back to the angiosperm ancestor and the results were compiled to obtain estimates of the relative number of duplications/losses/no changes that occurred along each branch of the species tree. We ran NOTUNG in the 'rearrange' mode to allow for rearrangements of branches with a low bootstrap support. Incorrect gene trees often result in predictions of duplication events near the root of the species tree, followed by a large number of losses at the tips (9). Since the predictions of these duplication and associated loss events are likely to be due to an inaccurate gene tree rather than presenting a real evolutionary phenomenon, we assessed the 'reliability' of each predicted duplication node using the duplication consistency score (10). Typically, such predicted duplication nodes are very imbalanced: i.e. there is little or no overlap in the species on the daughter branches. The duplication consistency score exactly assesses this imbalance by comparing the overlap in species of the daughter branches with their union. Because these problems often arise due to one misplaced branch, we pruned branches from the gene tree associated with duplication consistency scores < 0.6 and performed tree reconciliation again for the pruned tree. Duplications and loss events obtained for this pruned tree are then further used for the simulation. The end result of this procedure is a species tree with for each branch an estimated distribution of the predicted number of copy number changes a gene would undergo along that branch (Fig. S4A). To obtain the expected number of single-copy families under the assumption of random gene loss we evolved 9513 ancestral angiosperm genes, i.e. the same as the number of OGs that was analyzed, along this species tree according to the estimated copy number change distributions obtained for each branch. Hence, we assume that gene duplication and loss along each branch is independent (assumption of random gene loss). We repeated this procedure 100,000 times and in each round calculated the number of single-copy OGs to obtain an estimate of the expected number of single-copy OGs under the assumption of random gene loss (Fig. S4B). A p-value was calculated as the proportion of simulations in which the number of simulated OGs that is single-copy exceeds or is equal to the number of single-copy OGs that was observed.

Assessing the influence of the number of genomes used on the expected number of single-copy OGs

To assess how the number of expected single-copy OGs evolves if one adds more genomes we used the above simulation strategy to estimate the number of expected single-copy OGs based on respectively 14, 12, 9, 6 and 3 genomes. In this analysis we used the same estimates of the duplication and loss rates along the branches of the species tree as calculated above, but the simulation itself was based on a randomly sampled subset of the 17 genomes. For each of the number of genomes assessed we repeated the entire simulation process 20 times, each time for a different set of sampled genomes (1000 runs for each sample). The results are shown in Fig. S5. We observe that while there is a gradual increase in the expected number of single-copy OGs with a decreasing number of genomes, the specific set of genomes sampled plays a larger role than the

number of genomes sampled. Especially, not including genomes that are still highly duplicated, such as soybean and apple, seems to influence the obtained number of single-copy OGs substantially.

SI 4. Functional and evolutionary characterization single-copy genes

Functional enrichment analysis

The BINGO 2.44 Cytoscape plugin (11) was used to calculate functional enrichment values for the set of single-copy gene. We used a p-value threshold of 0.05 and p-values were corrected for multiple testing using the Benjamini and Hochberg method (12).

Phylogenetic distribution

To assess phylogenetic conservation of the single-copy genes in other organisms we obtained gene homologous relationships between the *A. thaliana* genes and 12 metazoa and 6 fungi from Homologene (13). The species in Homologene were subdivided into fungi and Metazoa. An *A. thaliana* gene was considered to be conserved in either of these both groups if it had homologous genes in 70% of the species in that phylostratum. P-values for phylogenetic conservation were calculated by a hypergeometric test with a multiple testing correction by the Benjamini and Hochberg method (12).

Of the 2986 *A. thaliana* single-copy genes, 2661 are found in the HomoloGene database. Remarkable is the high number of *A. thaliana* single-copy genes with metazoan homologs (Table S3): nearly 20% (675/3514) ($p < 2.22e-16$, hypergeometric test) of the *A. thaliana* genes with metazoan homologs belong to single-copy OGs.

Gene expression analysis

Pre-processed *A. thaliana* gene expression data measuring expression in different organs and different developmental stages was taken from CORNET (14). To obtain absolute gene expression levels, for each gene the geometric mean of its expression level was calculated across all 425 conditions.

To calculate gene expression breadth we first subdivided the arrays into different organ sets using the Plant Ontology associated with each array in CORNET. To ascertain that gene expression patterns in the different organs was comparable, a PCA plot of the expression data was made and showed clustering of the conditions according to the assigned organ. For each condition we fitted a bimodal distribution using the R mixtools package to classify genes as being expressed or not expressed in a certain condition (15). A gene was considered as being expressed in an organ if it was expressed in at least 70% of all the conditions measuring expression in the specific organ. Expression breadth for each gene is then calculated as the number of organs in which the gene is expressed.

Sequence conservation

K_a and K_s values for *A. thaliana* and *A. lyrata* gene pairs were obtained from the PLAZA 2.5 database (4, 13). Briefly, coding sequences were aligned using the CLUSTALW version 1.83 alignment tool (16).

From this alignment ambiguously aligned residues were stripped. The number of synonymous substitutions per synonymous site (K_s) and the number of nonsynonymous substitutions per nonsynonymous site (K_a) were estimated using the codeml package in PAML (17), with the Goldman & Yang (1994) model (18).

For the calculation of the Codon Adaptation Index (CAI), the CodonW package (http://codonw.sourceforge.net/) was used. We first used the correspondence analysis included in this package to select a reference set of *A. thaliana* genes that has highly biased codon usage. This reference set was then used to calculate the CAI for all genes according to (19).

Properties of strictly single-copy genes

We performed the above analyses (functional enrichment calculations, gene expression analysis and sequence conservation analyses) also separately on the set of 177 strictly single-copy genes (Table S4).

References

1. Remm M, Storm CE, & Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314(5):1041-1052.
2. Wang X*, et al.* (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet* 43(10):1035-1039.
3. D'Hont A*, et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488(7410):213-217.
4. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.
5. Proost S*, et al.* (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21(12):3718-3731.
6. Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696-704.
7. Van Bel M*, et al.* (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol* 158(2):590-600.
8. Durand D, Halldorsson BV, & Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol* 13(2):320-335.
9. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 8(7):R141.
10. Vilella AJ*, et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19(2):327-335.
11. Maere S, Heymans K, & Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21(16):3448-3449.
12. Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57(1):289-300.
13. Sayers EW*, et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40(Database issue):D13-25.
14. De Bodt S, Hollunder J, Nelissen H, Meulemeester N, & Inze D (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytol* 195(3):707-720.
15. Hebenstreit D*, et al.* (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* 7:497.
16. Oliver T, Schmidt B, Nathan D, Clemens R, & Maskell D (2005) Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* 21(16):3431-3432.
17. Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* 13(5):555-556.
18. Goldman N & Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725-736.
19. Sharp PM & Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3):1281-1295.

Supplementary figures

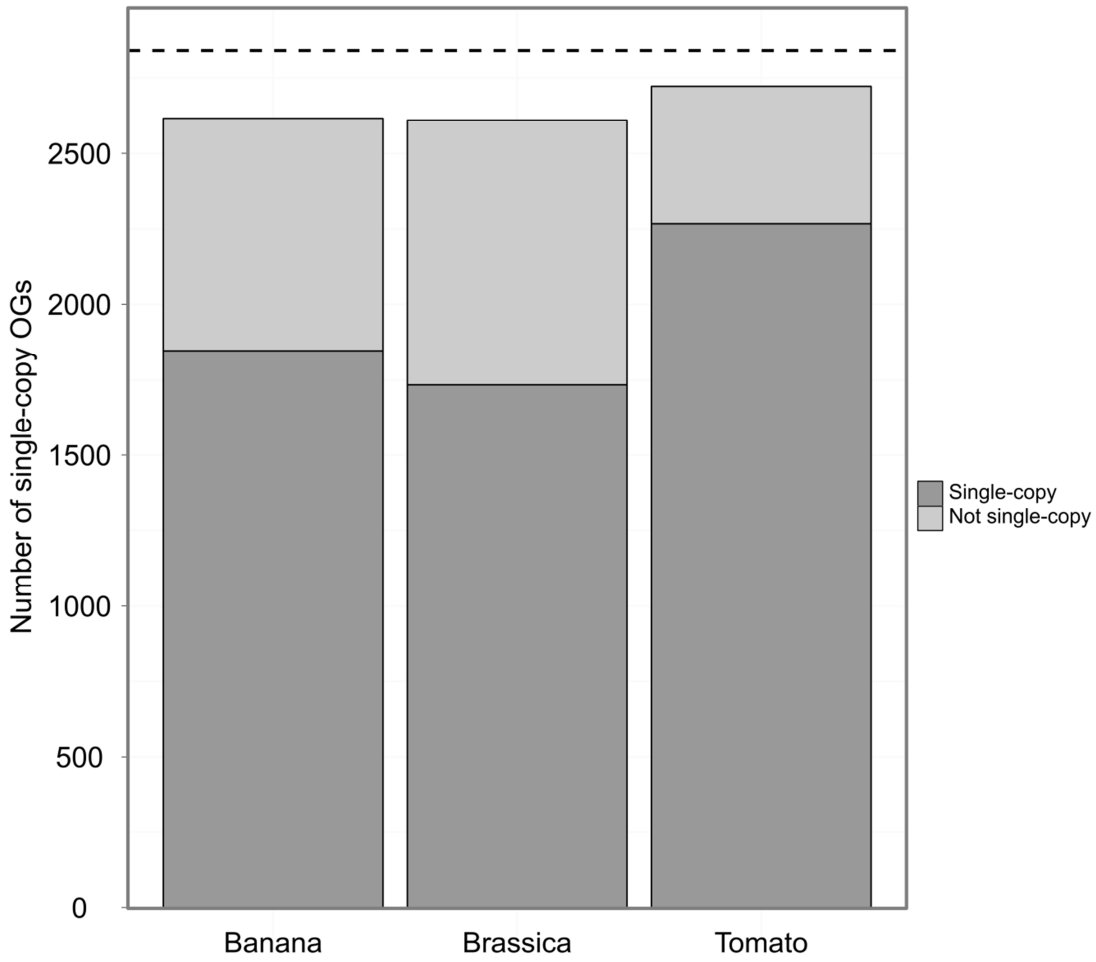

Fig. S1. Single-copy status of identified OGs is confirmed in *Brassica rapa*, tomato and banana. The bar plot represents the number of (mostly) single-copy OGs for which single-copy status could be confirmed (dark grey) in *Brassica,* tomato and banana. The proportion of (mostly) single-copy OGs for which single-copy status could not be confirmed, i.e. inparalogs were detected in the OG, is shown in light grey. The black dotted horizontal line represents the number of (mostly) single-copy OGs retrieved from the PLAZA database.

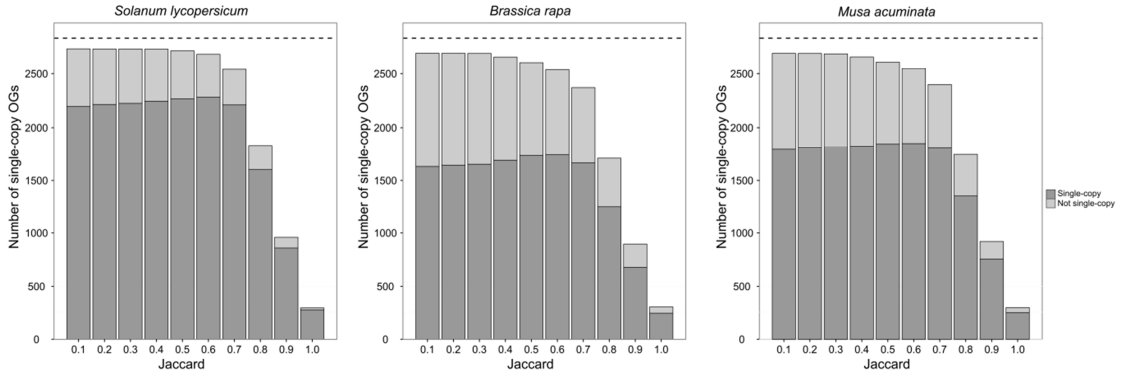

Fig. S2. Number of genes in tomato (*Solanum lycopersicum*), *Brassica rapa* and banana (*Musa acuminata*) that are assigned to the OrthoMCL OGs for different values of the Jaccard coefficient. Genes that are assigned to an OG (y-axis) are separated into genes that are single-copy in the species ('Single-copy') and those that seem to have duplicates for that particular OG ('Not single-copy'). The black dotted horizontal line represents the number of (mostly) single-copy OGs retrieved from the PLAZA database.

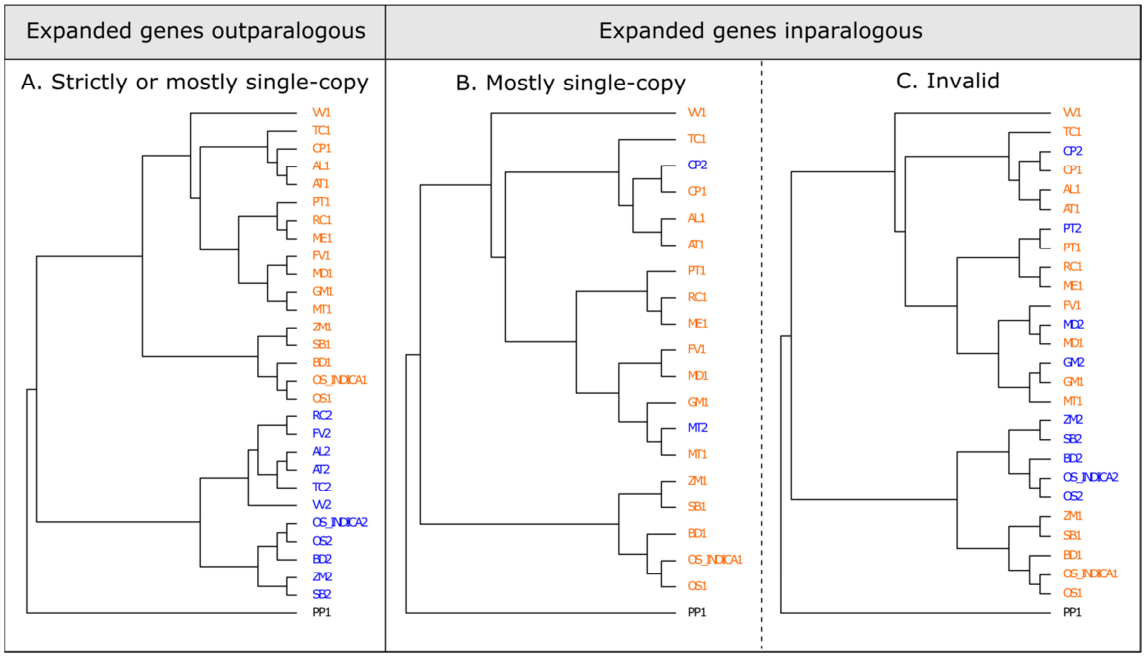

Fig. S3. Classification of expanded orthologous groups. Orthologous groups derived from PLAZA (in orange) were expanded with mutual best BLAST hits (in blue). Outgroup genes are colored in black. Maximum likelihood trees from these expanded orthologous groups were constructed to assess whether the additional genes were outparalogs to the original genes in the PLAZA orthologous groups. Based on this phylogenetic validation step, orthologous groups retained their original classification as strictly or mostly single-copy (A) if all expanded genes were outparalogs, they were classified as mostly single-copy (B) if the presence of inparalogs after expansion did not increase the number of duplicates to more than three species, and invalid (C) if after expansion the orthologous group did contain duplicates for more than three species. See text for details. (Abbreviations: VV = *Vitis vinifera*, TC = *Theobroma cacao*, CP= *Carica papaya*, AL = *Arabidopsis lyrata*, AT = *Arabidopsis thaliana*, PT = *Populus trichocarpa*, RC = *Ricinus communis*, ME = *Manihot esculenta*, FV = *Fragaria vesca*, MD = *Malus domestica*, GM = *Glycine max*, MT = *Medicago trunculata*, ZM = *Zea mays*, SB = *Sorghum bicolor*, BD = *Brachypodium distachyon*, OS_INDICA = *Oryza sativa ssp. indica*, OS = *Oryza sativa ssp. japonica*).

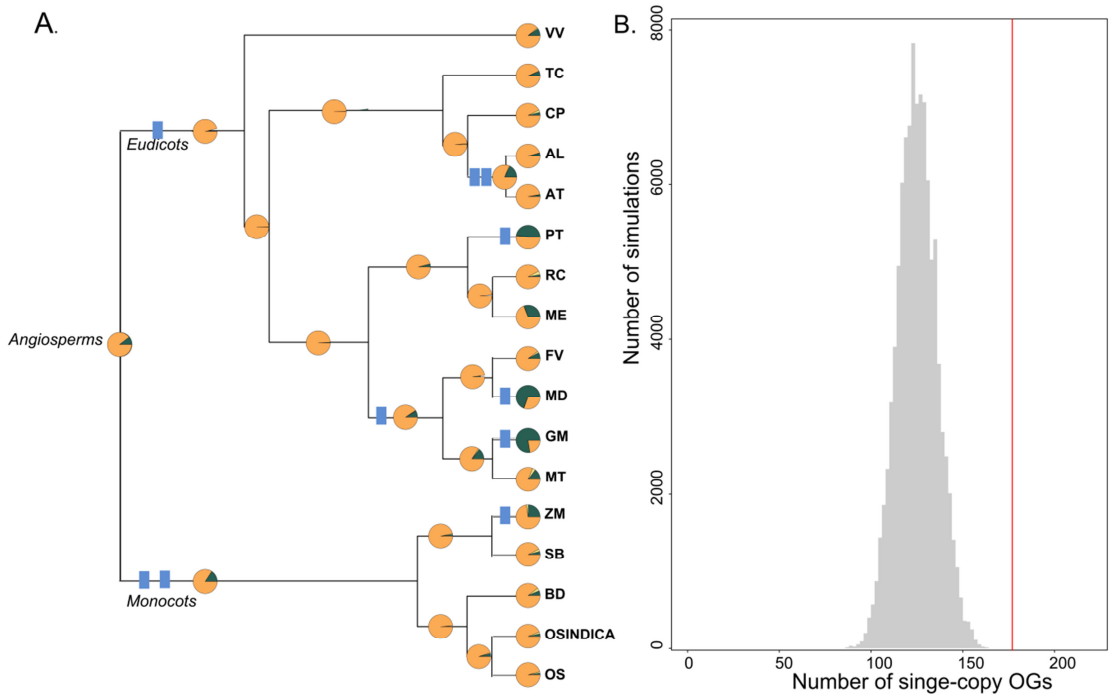

Fig. S4. Simulation of the expected number of single-copy families. (A) Species tree with pie charts on the branches denoting estimated fractions of duplications (green), losses (yellow) and no change (orange), as estimated from the PLAZA OGs using tree reconciliation. Blue rectangles represent known WGD events (B) The number of expected (grey bar) single-copy families in 100,000 simulations of gene family evolution along the phylogenetic tree represented in A. The observed number of single-copy families is represented by the red vertical line. (Abbreviations: VV = *Vitis vinifera*, TC = *Theobroma cacao*, CP= *Carica papaya*, AL = *Arabidopsis lyrata*, AT = *Arabidopsis thaliana*, PT = *Populus trichocarpa*, RC = *Ricinus communis*, ME = *Manihot esculenta*, FV = *Fragaria vesca*, MD = *Malus domestica*, GM = *Glycine max*, MT = *Medicago trunculata*, ZM = *Zea mays*, SB = *Sorghum bicolor*, BD = *Brachypodium distachyon*, OS_INDICA = *Oryza sativa ssp. indica*, OS = *Oryza sativa ssp. japonica*).

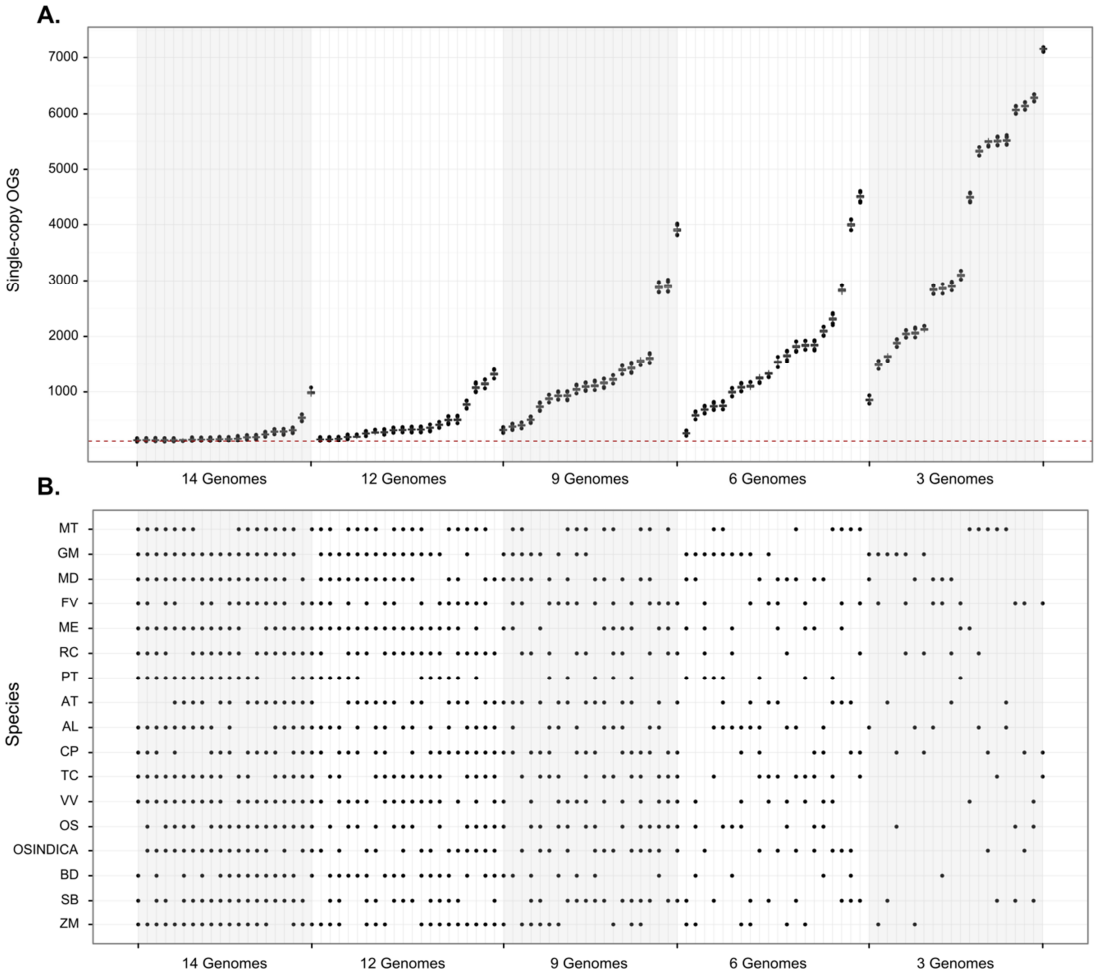

Fig. S5. Assessment of how the expected number of strictly single-copy OGs changes with a decreasing number of genomes. Panel A represents the distributions (boxplots) of the expected number of single-copy OGs for 1000 simulations of gene loss along the species tree in Fig. S4A for respectively 14, 12, 9, 6 and 3 genomes. The y-axis refers to the number of expected single-copy OGs obtained averaged over 1000 simulation, while the x-axis represents the different genome samples to which the simulation was applied (cfr. panel B). Panel B represents the specific set of genomes sampled. The red dotted line in panel A represents the median number of expected strictly single-copy OGs obtained for 17 genomes (Fig. S4B). The same abbreviations as in Fig. S4 were used.

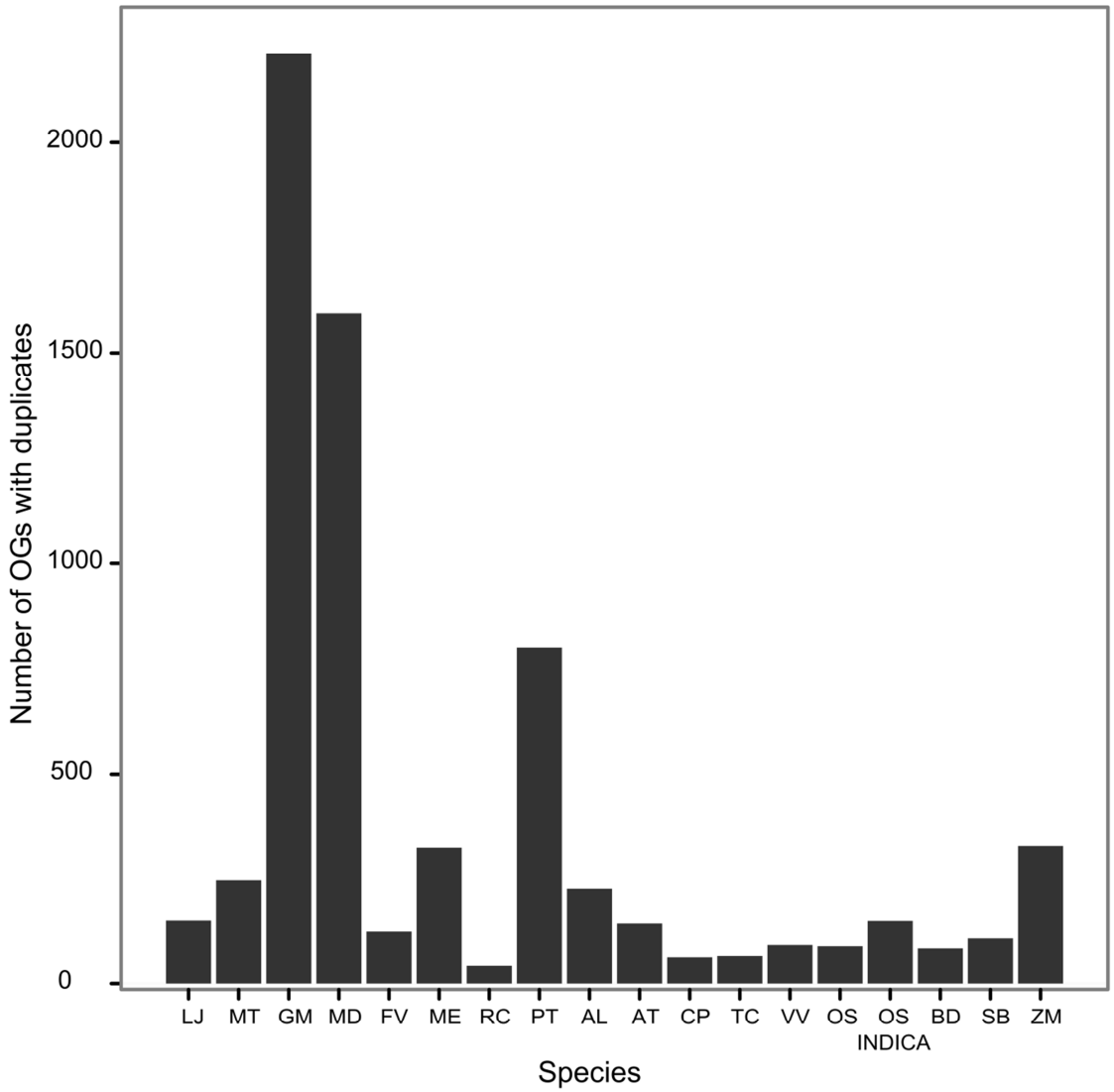

Fig. S6. Bias in duplicate content of strict and majority single-copy OGs. Plant species included in this study (x-axis) are plotted against the number of single-copy orthologous groups (y-axis) that have duplicates for this species. (Abbreviations: VV = *Vitis vinifera*, TC = *Theobroma cacao*, CP= *Carica papaya*, AL = *Arabidopsis lyrata*, AT = *Arabidopsis thaliana*, PT = *Populus trichocarpa*, RC = *Ricinus communis*, ME = *Manihot esculenta*, FV = *Fragaria vesca*, MD = *Malus domestica*, GM = *Glycine max*, MT = *Medicago trunculata*, ZM = *Zea mays*, SB = *Sorghum bicolor*, BD = *Brachypodium distachyon*, OS_INDICA = *Oryza sativa ssp. indica*, OS = *Oryza sativa ssp. japonica*).

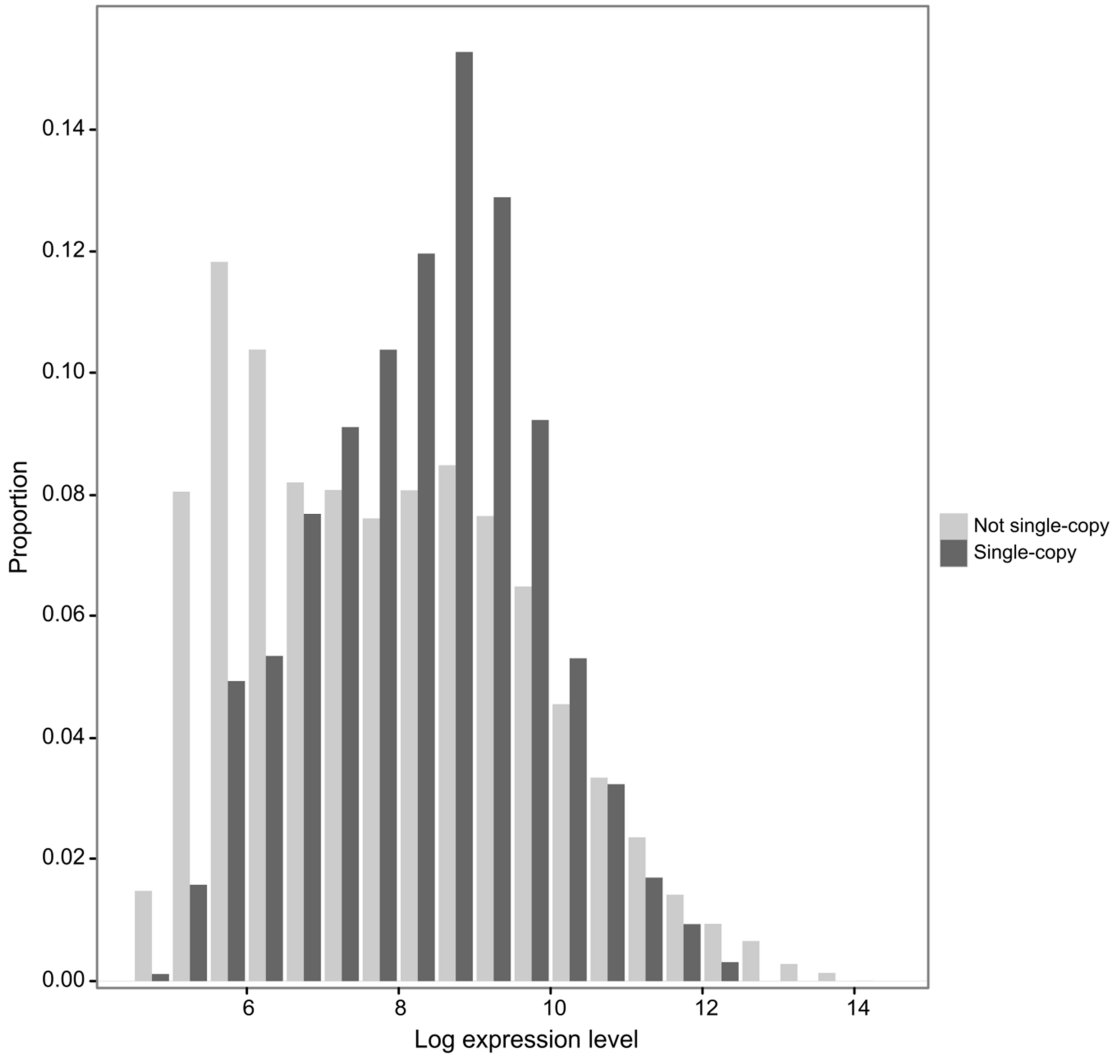

Fig. S7. Absolute expression levels for *A. thaliana* genes in single-copy and non-single-copy OGs, with genes assigned to the photosynthesis GO-category removed from the set. The figure shows the proportion of genes (y-axis) that have a certain absolute expression level (x-axis), calculated as the geometric mean of all expression measurements of a certain gene.

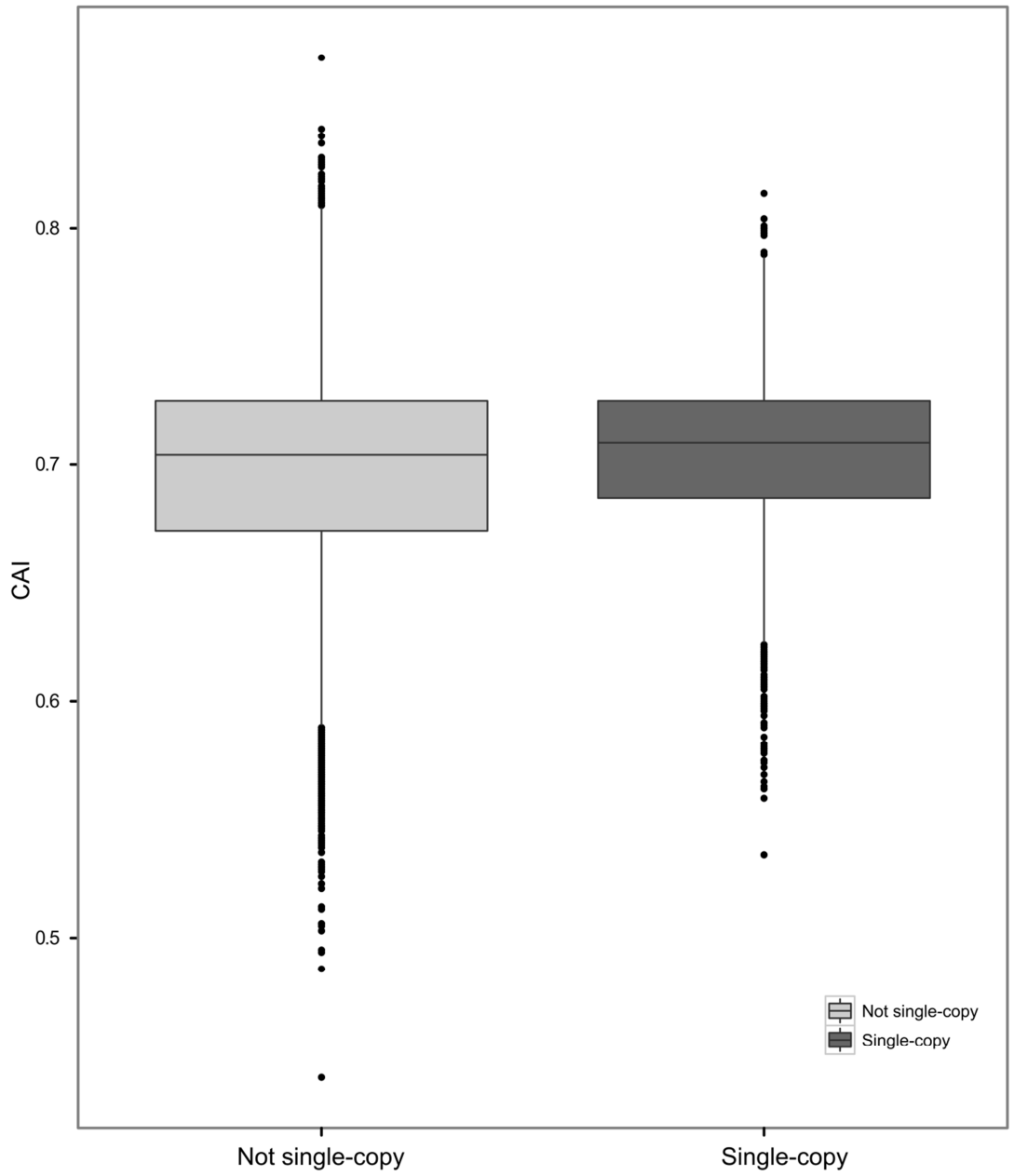

Fig. S8. Codon Adaptation Index (CAI) for single-copy genes and genes that are not single-copy in *A. thaliana*.

Supplementary tables

Table S1. Statistical significance of (adjusted p-value) over- and underrepresentation of GO-categories for 'strictly' single-copy genes. Here all *A. thaliana* genes included to OGs that traced back to the angiosperm common ancestor were used as a background for the hypergeometric test, instead of all *A. thaliana* genes.

	GO-term	Ontology	Adjusted p-value (FDR < 0.05)	# single-copy genes	# of *A. thaliana* genes
Overrepresented	DNA repair	BP	6.74E-11	64	117
	Response to DNA damage stimulus	BP	8.22E-11	66	123
	DNA recombination	BP	2.99E-7	28	41
	DNA metabolic process	BP	5.59E-14	114	293
	DNA replication	BP	1.04E-2	32	74
	Plastid organization	BP	4.16E-4	39	84
	Photosynthesis	BP	3.17E-3	40	94
	Meiosis I	BP	8.36E-3	15	25
	Chloroplast	CC	6.16E-33	538	1544
Under represented	Regulation of transcription	BP	2.31E-11	63	552
	Regulation of gene expression	BP	3.68E-10	82	638
	Phosphorylation	BP	1.89E-17	51	567

Table S2. Statistical significance of (adjusted p-value) over- and underrepresentation of GO-categories for 'strictly' single-copy genes.

	GO-term	Ontology	Adjusted p-value (FDR < 0.05)	# single-copy genes	# of *A. thaliana* genes
Overrepresented	DNA repair	BP	2.7298E-8	64	154
	Response to DNA damage stimulus	BP	5.1921E-9	13	163
	DNA recombination	BP	1.0031E-7	8	52
	DNA metabolic process	BP	3.1052E-10	18	311
	Meiosis I	BP	1.0299E-3	4	30

Table S3. Phylogenetic distribution of the *A. thaliana* single-copy genes according to HomoloGene homology relationships. An *A. thaliana* gene was denoted to belong to a certain phylostratum if it showed homology relationships to at least 70% of the species belonging to that stratum. P-values were calculated by hypergeometric test (FDR adjusted p-value < 0.05) .

Phylostratum	# single-copy with homolog in phylostratum/# single-copy in Homologene	# *A. thaliana genes* with homolog in phylostratum/# A. thaliana genes in Homologene	Adjusted p-value (FDR)
All	231/2661	1290/19935	1.50e-06
Metazoa	675/2661	3514/19935	<2.22e-16
Fungi	301/2661	1807/19935	1.41e-05

Table S4. Gene expression and sequence conservation analysis of the strictly single-copy genes. For each of the properties values for the set of strictly single-copy genes was compared to those of non single-copy genes.

Property	p-value (one-sided Mann-Whitney U-test)
Gene expression level	0.6491
Gene expression breadth	0.0134
K_s	9.99e-05
K_a	0.00017

Table S5. Significantly overrepresented GO categories among *A. thaliana* single-copy OG member genes (hypergeometric test, FDR adjusted p-value < 0.05).

GO-term	Ontology	Adjusted p-value	#single-copy genes	# A. thaliana genes
nitrogen compound metabolic process	BP	3.3944E-49	359	1548
cellular nitrogen compound metabolic process	BP	5.3436E-49	348	1484
nucleic acid metabolic process	BP	1.6422E-42	234	866
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	BP	2.7937E-42	273	1104
DNA metabolic process	BP	6.3775E-33	114	311
DNA repair	BP	1.8137E-21	64	154
response to DNA damage stimulus	BP	1.8137E-21	66	163

metabolic process	BP	2.7354E-16	896	6834
RNA metabolic process	BP	9.8850E-14	115	510
cellular metabolic process	BP	1.2509E-13	721	5407
DNA recombination	BP	1.3949E-12	28	52
plastid organization	BP	1.5284E-11	39	102
organelle organization	BP	6.2101E-11	110	526
photosynthesis	BP	1.2060E-10	40	113
cellular response to stress	BP	1.4844E-10	84	364
cofactor metabolic process	BP	4.0235E-10	59	219
cellular component organization	BP	4.0292E-10	165	935
cofactor biosynthetic process	BP	1.2834E-8	41	135
photosynthesis, light reaction	BP	1.6778E-8	26	63
DNA replication	BP	4.9269E-8	32	94
tetrapyrrole metabolic process	BP	6.1330E-8	25	62
heterocycle metabolic process	BP	9.5825E-8	73	340
porphyrin metabolic process	BP	1.4869E-7	24	60
RNA processing	BP	8.1484E-7	70	337
double-strand break repair	BP	8.3345E-7	14	24
cellular process	BP	8.3345E-7	888	7393
chloroplast organization	BP	8.3345E-7	24	65
embryonic development	BP	1.0020E-6	83	429
DNA-dependent DNA replication	BP	1.2809E-6	19	44
tRNA processing	BP	1.3235E-6	18	40
vitamin biosynthetic process	BP	1.6502E-6	22	58
tRNA metabolic process	BP	1.7376E-6	30	98
heterocycle biosynthetic process	BP	1.8145E-6	34	120
meiosis I	BP	3.0321E-6	15	30
pseudouridine synthesis	BP	5.2235E-6	11	17
tetrapyrrole biosynthetic process	BP	5.6178E-6	19	48

vitamin metabolic process	BP	7.7600E-6	22	63
primary metabolic process	BP	1.2007E-5	695	5719
M phase of meiotic cell cycle	BP	2.4023E-5	17	43
meiosis	BP	2.4023E-5	17	43
porphyrin biosynthetic process	BP	3.4510E-5	17	44
pigment metabolic process	BP	6.0216E-5	26	92
cellular macromolecule metabolic process	BP	7.0722E-5	461	3667
macromolecule metabolic process	BP	8.4860E-5	507	4086
M phase	BP	9.6980E-5	20	62
embryonic development ending in seed dormancy	BP	9.8930E-5	68	372
cellular response to stimulus	BP	1.1586E-4	115	729
chlorophyll metabolic process	BP	1.4052E-4	15	39
protein complex biogenesis	BP	1.8625E-4	32	134
protein complex assembly	BP	1.8625E-4	32	134
chromosome organization involved in meiosis	BP	1.9031E-4	10	19
synapsis	BP	1.9031E-4	10	19
cellular nitrogen compound catabolic process	BP	2.0833E-4	11	23
cellular protein complex assembly	BP	2.1486E-4	23	82
heterocycle catabolic process	BP	2.9991E-4	12	28
carotenoid metabolic process	BP	2.9991E-4	12	28
tetraterpenoid metabolic process	BP	2.9991E-4	12	28
chloroplast fission	BP	3.2750E-4	8	13
chiasma assembly	BP	3.2750E-4	8	13
ncRNA metabolic process	BP	3.6869E-4	41	198
fruit development	BP	4.4823E-4	77	459
cell cycle process	BP	4.4893E-4	29	122
cell cycle phase	BP	4.7084E-4	21	75
seed development	BP	4.7084E-4	74	438
carotenoid biosynthetic process	BP	4.7084E-4	10	21

tetraterpenoid biosynthetic process	BP	4.7084E-4	10	21
small molecule metabolic process	BP	4.8210E-4	175	1248
pigment biosynthetic process	BP	5.4696E-4	21	76
cellular component assembly	BP	5.7005E-4	49	258
chromosome organization	BP	6.0108E-4	42	210
ncRNA processing	BP	6.7313E-4	31	138
water-soluble vitamin biosynthetic process	BP	7.0363E-4	15	45
DNA-dependent DNA replication initiation	BP	7.2688E-4	7	11
meiotic cell cycle	BP	7.3583E-4	18	61
cellular nitrogen compound biosynthetic process	BP	7.5541E-4	65	378
phospholipid biosynthetic process	BP	7.8996E-4	17	56
plastid fission	BP	9.9732E-4	8	15
protein amino acid lipidation	BP	9.9732E-4	8	15
lipoprotein metabolic process	BP	9.9732E-4	8	15
lipoprotein biosynthetic process	BP	9.9732E-4	8	15
reciprocal meiotic recombination	BP	1.7906E-3	8	16
photosynthetic electron transport chain	BP	2.3783E-3	10	25
water-soluble vitamin metabolic process	BP	2.4154E-3	15	50
regulation of DNA repair	BP	2.5207E-3	4	4
fat-soluble vitamin metabolic process	BP	2.6470E-3	7	13
fat-soluble vitamin biosynthetic process	BP	2.6470E-3	7	13
photosynthesis, light harvesting	BP	2.6470E-3	9	21
RNA modification	BP	3.0190E-3	25	111
regulation of DNA metabolic process	BP	3.2310E-3	10	26
cell cycle	BP	3.6064E-3	31	152
Group II intron splicing	BP	4.1344E-3	5	7
DNA catabolic process	BP	4.1344E-3	5	7
carotene biosynthetic process	BP	4.1344E-3	5	7
cytochrome complex assembly	BP	4.4274E-3	7	14

coenzyme metabolic process	BP	6.6150E-3	28	137
base-excision repair	BP	6.6683E-3	8	19
GPI anchor metabolic process	BP	6.8068E-3	6	11
photosynthetic electron transport in photosystem I	BP	7.1051E-3	7	15
protein amino acid deacetylation	BP	7.1051E-3	7	15
mismatch repair	BP	7.1051E-3	7	15
chromatin modification	BP	7.4468E-3	20	86
macromolecular complex subunit organization	BP	7.4468E-3	39	216
protein repair	BP	8.7477E-3	5	8
photosystem II repair	BP	8.7477E-3	5	8
GPI anchor biosynthetic process	BP	8.7477E-3	5	8
mRNA modification	BP	8.7477E-3	4	5
lipid A biosynthetic process	BP	8.7477E-3	4	5
phylloquinone biosynthetic process	BP	8.7477E-3	4	5
phylloquinone metabolic process	BP	8.7477E-3	4	5
meiotic DNA double-strand break formation	BP	8.7477E-3	4	5
lipid A metabolic process	BP	8.7477E-3	4	5
organophosphate metabolic process	BP	1.0123E-2	21	95
reproductive structure development	BP	1.0123E-2	104	735
coenzyme biosynthetic process	BP	1.0123E-2	17	70
histidine family amino acid metabolic process	BP	1.0123E-2	6	12
double-strand break repair via homologous recombination	BP	1.0123E-2	6	12
recombinational repair	BP	1.0123E-2	6	12
histidine metabolic process	BP	1.0123E-2	6	12
carotene metabolic process	BP	1.0123E-2	6	12
phospholipid metabolic process	BP	1.0124E-2	20	89
pyridine nucleotide metabolic process	BP	1.1348E-2	11	36
positive regulation of catalytic activity	BP	1.2126E-2	10	31

chlorophyll biosynthetic process	BP	1.2314E-2	9	26
glycoprotein metabolic process	BP	1.4114E-2	15	60
positive regulation of molecular function	BP	1.5369E-2	10	32
nicotinamide nucleotide metabolic process	BP	1.5369E-2	10	32
histone deacetylation	BP	1.5369E-2	6	13
photoreactive repair	BP	1.5369E-2	3	3
maintenance of fidelity involved in DNA-dependent DNA replication	BP	1.5369E-2	3	3
pyrimidine dimer repair	BP	1.5369E-2	3	3
NADH dehydrogenase complex (plastoquinone) assembly	BP	1.5369E-2	3	3
NADH dehydrogenase complex assembly	BP	1.5369E-2	3	3
transcription from plastid promoter	BP	1.5369E-2	3	3
protein folding	BP	1.6981E-2	33	184
cellular macromolecular complex subunit organization	BP	1.8007E-2	30	163
lipopolysaccharide biosynthetic process	BP	1.9090E-2	4	6
lipopolysaccharide metabolic process	BP	1.9090E-2	4	6
cytochrome b6f complex assembly	BP	1.9090E-2	4	6
vitamin K biosynthetic process	BP	1.9090E-2	4	6
vitamin K metabolic process	BP	1.9090E-2	4	6
cellular ketone metabolic process	BP	2.0616E-2	89	630
covalent chromatin modification	BP	2.1017E-2	15	63
protein amino acid alkylation	BP	2.2745E-2	10	34
protein amino acid methylation	BP	2.2745E-2	10	34
histidine biosynthetic process	BP	2.3840E-2	5	10
histidine family amino acid biosynthetic process	BP	2.3840E-2	5	10
photosystem II assembly	BP	2.3840E-2	5	10
translational termination	BP	2.3840E-2	5	10
oxidoreduction coenzyme metabolic process	BP	2.4044E-2	11	40

methylation	BP	2.4398E-2	17	77
organic acid metabolic process	BP	2.7352E-2	87	621
response to UV	BP	2.7352E-2	15	65
regulation of response to stress	BP	2.7686E-2	17	78
post-embryonic development	BP	2.8889E-2	118	884
small molecule catabolic process	BP	3.1638E-2	29	163
regulation of cellular response to stress	BP	3.1638E-2	6	15
cellular amino acid metabolic process	BP	3.1638E-2	47	300
histone modification	BP	3.2601E-2	14	60
macromolecular complex assembly	BP	3.3405E-2	34	201
organelle fission	BP	3.4303E-2	12	48
cell cycle checkpoint	BP	3.4897E-2	4	7
tRNA modification	BP	3.4897E-2	4	7
nucleobase, nucleoside and nucleotide catabolic process	BP	3.4897E-2	4	7
nucleobase, nucleoside, nucleotide and nucleic acid catabolic process	BP	3.4897E-2	4	7
oxoacid metabolic process	BP	3.4897E-2	86	620
carboxylic acid metabolic process	BP	3.4897E-2	86	620
peroxisomal transport	BP	3.5019E-2	5	11
protein targeting to peroxisome	BP	3.5019E-2	5	11
negative regulation of DNA metabolic process	BP	3.5019E-2	5	11
cellular amine metabolic process	BP	3.6262E-2	50	327
response to ionizing radiation	BP	4.2370E-2	6	16
protein amino acid N-linked glycosylation	BP	4.2697E-2	7	21
electron transport chain	BP	4.2848E-2	13	56
leucine catabolic process	BP	4.2848E-2	3	4
glycoprotein catabolic process	BP	4.2848E-2	3	4
nucleobase catabolic process	BP	4.2848E-2	3	4

cellular lipid metabolic process	BP	4.5662E-2	59	404
generation of precursor metabolites and energy	BP	4.6681E-2	33	199
one-carbon metabolic process	BP	4.8169E-2	19	97
regulation of flower development	BP	4.8169E-2	20	104
peroxisome organization	BP	4.8827E-2	8	27
cellular protein complex disassembly	BP	4.9389E-2	5	12
nucleotide-excision repair	BP	4.9389E-2	5	12
tetrapyrrole catabolic process	BP	4.9389E-2	5	12
porphyrin catabolic process	BP	4.9389E-2	5	12
pteridine and derivative biosynthetic process	BP	4.9389E-2	5	12
xanthophyll metabolic process	BP	4.9389E-2	5	12
cellular macromolecular complex disassembly	BP	4.9389E-2	5	12
plastid	CC	0.0000E-100	551	2139
chloroplast	CC	0.0000E-100	538	2070
cytoplasmic part	CC	3.2208E-84	799	4323
cytoplasm	CC	1.6016E-81	844	4745
intracellular	CC	1.1341E-72	1100	7208
intracellular part	CC	5.3956E-64	1044	6908
intracellular organelle	CC	6.9580E-63	951	6090
organelle	CC	6.9580E-63	951	6091
intracellular membrane-bounded organelle	CC	2.5746E-61	910	5766
membrane-bounded organelle	CC	2.5746E-61	910	5767
plastid part	CC	1.1143E-53	231	782
chloroplast part	CC	2.5780E-49	219	755
thylakoid	CC	1.0171E-34	115	322
intracellular organelle part	CC	1.4528E-32	364	1970
organelle part	CC	1.6880E-32	364	1972

organelle subcompartment	CC	1.7274E-29	94	256
chloroplast thylakoid	CC	4.1071E-29	93	254
plastid thylakoid	CC	4.1071E-29	93	254
thylakoid part	CC	8.8678E-29	95	266
photosynthetic membrane	CC	1.0006E-19	74	227
plastid stroma	CC	4.2381E-19	96	354
thylakoid membrane	CC	7.6313E-19	72	224
plastid thylakoid membrane	CC	1.4477E-18	69	211
chloroplast thylakoid membrane	CC	1.4477E-18	69	211
cell part	CC	7.8143E-18	1353	11708
cell	CC	7.8143E-18	1353	11708
chloroplast stroma	CC	3.9687E-15	85	335
thylakoid lumen	CC	2.2323E-14	34	74
chloroplast envelope	CC	3.8824E-13	85	361
chloroplast thylakoid lumen	CC	1.2761E-12	28	58
plastid thylakoid lumen	CC	1.2761E-12	28	58
plastid envelope	CC	1.3336E-12	87	382
envelope	CC	4.7322E-9	109	601
organelle envelope	CC	4.7322E-9	109	601
nucleoid	CC	3.8088E-8	17	34
plastid inner membrane	CC	1.0878E-7	18	40
chromosome	CC	6.5736E-7	39	158
chloroplast inner membrane	CC	1.2770E-6	16	37
cytoplasmic chromosome	CC	1.2770E-6	11	18
plastid chromosome	CC	1.2770E-6	11	18
plastid membrane	CC	1.4929E-6	24	76
plastid nucleoid	CC	1.8369E-6	12	22
mitochondrion	CC	3.6194E-6	126	817
chloroplast membrane	CC	5.2232E-6	21	65

CUL4 RING ubiquitin ligase complex	CC	1.5258E-5	29	115
NAD(P)H dehydrogenase complex (plastoquinone)	CC	6.2898E-5	7	10
plastoglobule	CC	1.0684E-4	17	55
cullin-RING ubiquitin ligase complex	CC	3.6023E-4	31	148
protein complex	CC	4.7209E-4	147	1084
replication fork	CC	2.7463E-3	5	8
extrinsic to membrane	CC	2.8124E-3	19	83
photosystem	CC	5.6117E-3	11	38
nuclear replisome	CC	5.7897E-3	3	3
nuclear replication fork	CC	5.7897E-3	3	3
replisome	CC	5.7897E-3	3	3
organelle membrane	CC	6.2118E-3	64	435
chloroplast stromal thylakoid	CC	7.0928E-3	4	6
chromosomal part	CC	1.1858E-2	22	115
ubiquitin ligase complex	CC	1.2623E-2	36	221
organelle inner membrane	CC	1.3429E-2	32	191
origin recognition complex	CC	1.4081E-2	4	7
endoplasmic reticulum part	CC	1.4081E-2	14	62
subsynaptic reticulum	CC	1.4081E-2	14	62
heterotrimeric G-protein complex	CC	2.1892E-2	14	65
nuclear chromosome	CC	2.4196E-2	8	28
prefoldin complex	CC	2.4237E-2	4	8
photosystem II	CC	2.7373E-2	7	23
extrinsic to plasma membrane	CC	2.7373E-2	14	67
chromosome, centromeric region	CC	2.9799E-2	5	13
voltage-gated potassium channel complex	CC	3.7342E-2	3	5
potassium channel complex	CC	3.7342E-2	3	5
cation channel complex	CC	3.7342E-2	3	5
ion channel complex	CC	3.7342E-2	3	5

CAAX-protein geranylgeranyltransferase complex	CC	4.0308E-2	2	2
nitrite reductase complex [NAD(P)H]	CC	4.0308E-2	2	2
amyloplast	CC	4.0308E-2	2	2
integral to chloroplast inner membrane	CC	4.0308E-2	2	2
integral to plastid inner membrane	CC	4.0308E-2	2	2
starch grain	CC	4.0308E-2	2	2
catalytic activity	MF	2.5987E-15	975	7553
nuclease activity	MF	5.8483E-14	52	144
hydrolase activity	MF	4.6432E-13	395	2632
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	MF	8.1729E-12	40	104
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	MF	5.2792E-11	28	58
helicase activity	MF	3.8648E-8	45	160
RNA methyltransferase activity	MF	4.2909E-8	14	20
endonuclease activity	MF	1.3794E-7	26	68
purine NTP-dependent helicase activity	MF	6.4994E-7	31	98
ATP-dependent helicase activity	MF	6.4994E-7	31	98
isomerase activity	MF	3.3223E-6	47	198
transferase activity, transferring one-carbon groups	MF	3.7077E-6	46	193
methyltransferase activity	MF	6.8496E-6	45	191
hydrolase activity, acting on ester bonds	MF	2.6105E-5	141	904
deacetylase activity	MF	4.0361E-5	12	23
hydrolase activity, acting on acid anhydrides	MF	1.3496E-4	110	688
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	MF	2.2812E-4	108	681
pyrophosphatase activity	MF	3.1805E-4	107	679
peptidyl-prolyl cis-trans isomerase activity	MF	3.5601E-4	18	56
nucleoside-triphosphatase activity	MF	3.5601E-4	103	651
cis-trans isomerase activity	MF	4.3767E-4	18	57

ATPase activity, coupled	MF	4.3767E-4	52	273
iron-sulfur cluster binding	MF	4.4626E-4	13	33
metal cluster binding	MF	4.4626E-4	13	33
structure-specific DNA binding	MF	9.2885E-4	15	45
deoxyribonuclease activity	MF	9.2885E-4	6	8
alpha-amylase activity	MF	9.2885E-4	6	8
mismatched DNA binding	MF	9.2885E-4	7	11
DNA-dependent ATPase activity	MF	9.8769E-4	12	31
ribonuclease activity	MF	1.0790E-3	18	62
acetyltransferase activity	MF	1.0790E-3	18	62
double-stranded DNA binding	MF	1.8480E-3	12	33
phosphogluconate dehydrogenase (decarboxylating) activity	MF	2.1488E-3	6	9
translation termination factor activity	MF	3.2024E-3	7	13
translation release factor activity	MF	3.2024E-3	7	13
exonuclease activity	MF	3.8354E-3	14	46
histone deacetylase activity	MF	5.2144E-3	8	18
protein deacetylase activity	MF	5.2144E-3	8	18
ATPase activity	MF	5.2144E-3	55	328
endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters	MF	5.3079E-3	11	32
DNA helicase activity	MF	7.6470E-3	8	19
damaged DNA binding	MF	8.9392E-3	9	24
nucleotidyltransferase activity	MF	1.0352E-2	30	154
aminoacyl-tRNA hydrolase activity	MF	1.0417E-2	5	8
UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase activity	MF	1.0417E-2	4	5
glycogen debranching enzyme activity	MF	1.0417E-2	4	5
N-acetyltransferase activity	MF	1.4028E-2	13	47

endoribonuclease activity, producing 5'-phosphomonoesters	MF	1.5058E-2	10	31
amylase activity	MF	1.7843E-2	7	17
nucleotide binding	MF	1.8385E-2	258	2085
macrolide binding	MF	1.8385E-2	8	22
FK506 binding	MF	1.8385E-2	8	22
5'-nucleotidase activity	MF	1.8385E-2	3	3
isoamylase activity	MF	1.8385E-2	3	3
DNA primase activity	MF	1.8385E-2	3	3
endoribonuclease activity	MF	2.6433E-2	12	45
N-acyltransferase activity	MF	2.6454E-2	13	51
drug binding	MF	3.2259E-2	8	24
DNA-directed DNA polymerase activity	MF	3.2259E-2	8	24
3'-5' exonuclease activity	MF	4.2467E-2	8	25
selenium binding	MF	4.6324E-2	4	7
ribonuclease III activity	MF	4.7043E-2	5	11