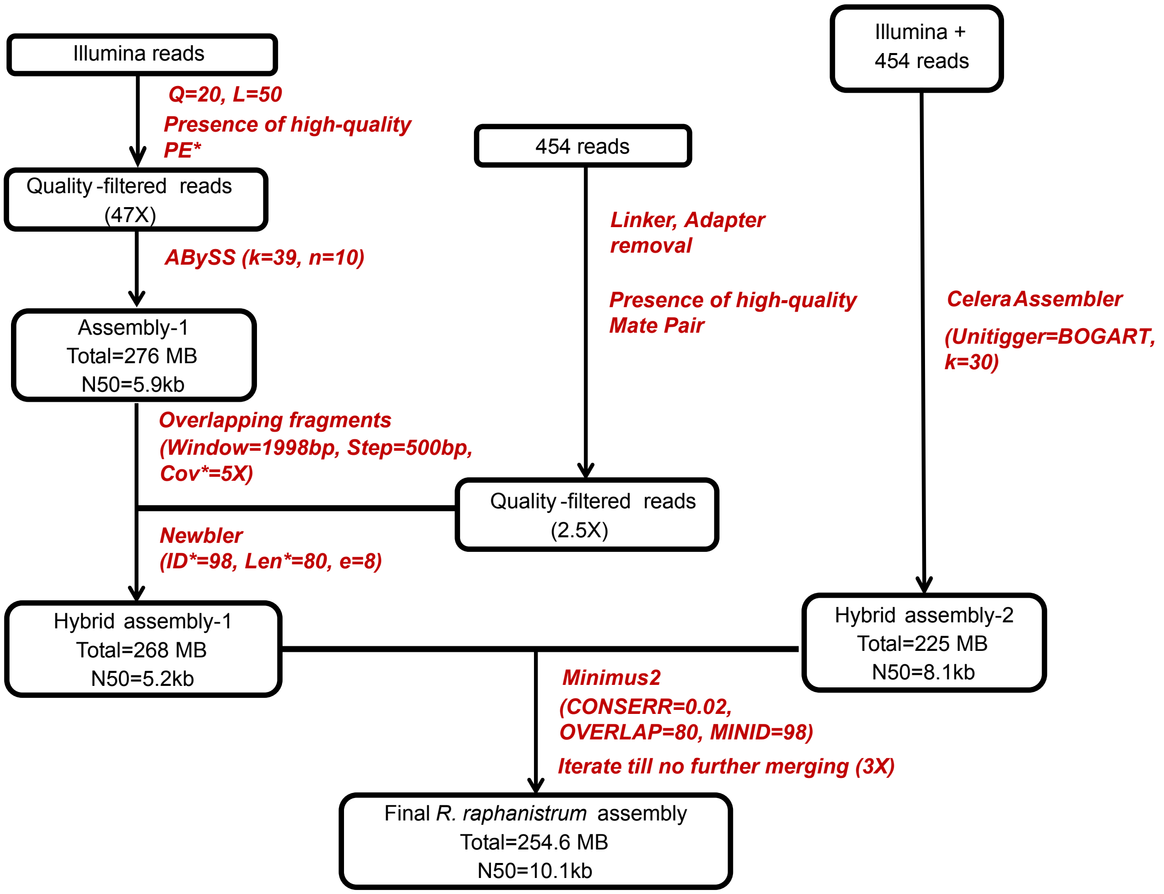
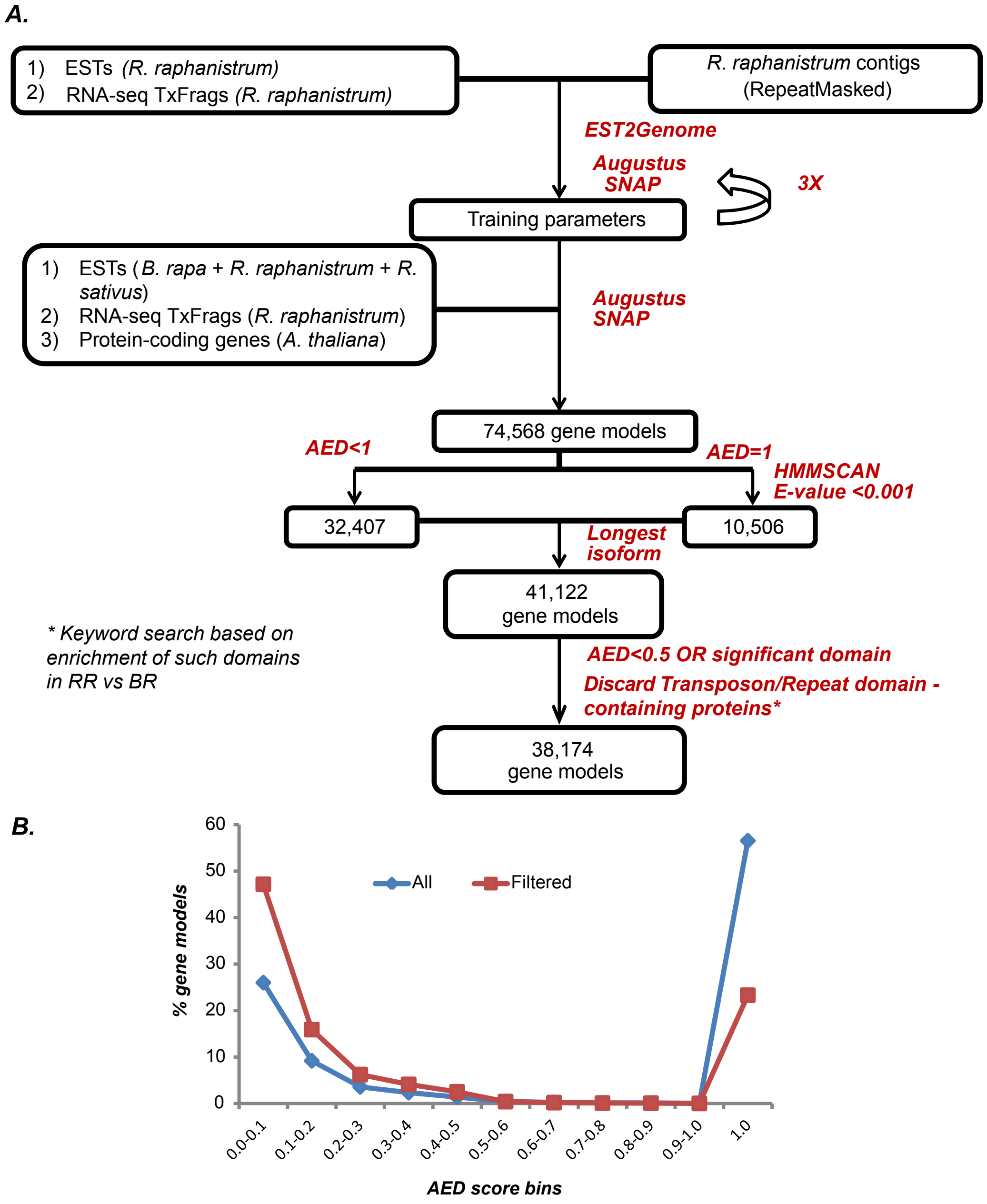
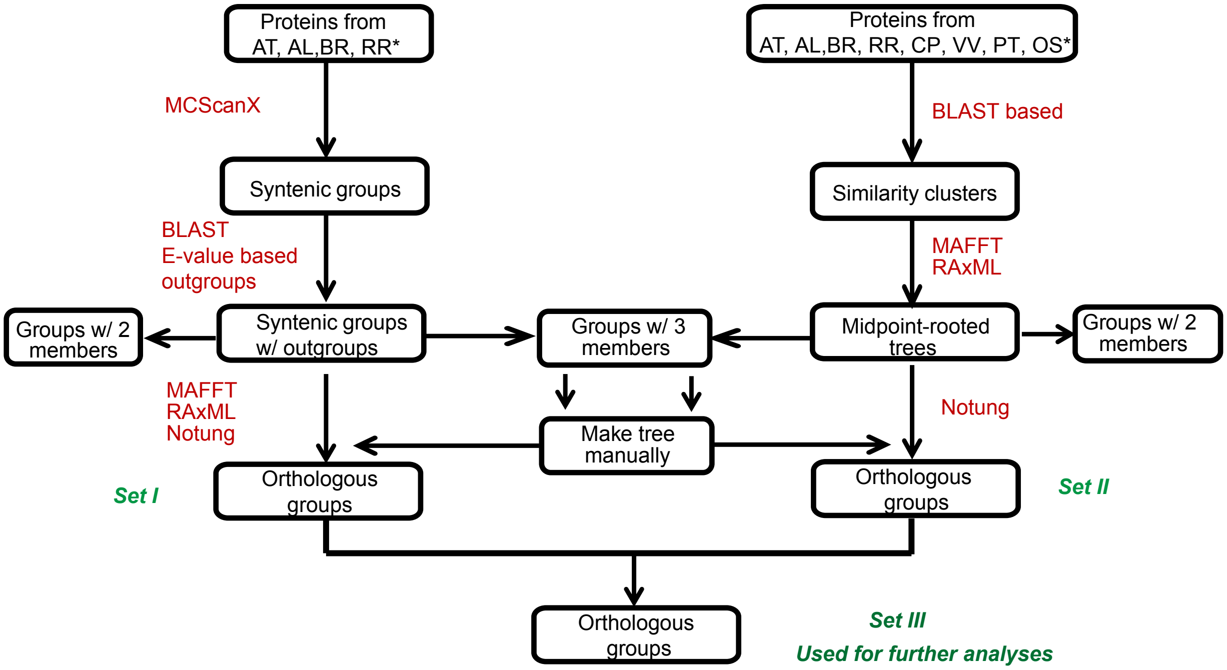

Illumina reads
Q=20, L=50
Presence of high-quality PE*
Quality -filtered reads (47X)
ABySS (k=39, n=10)
Assembly-1
Total=276 MB
N50=5.9kb
454 reads
Linker, Adapter removal
Presence of high-quality Mate Pair
Illumina + 454 reads
Celera Assembler
(Unitigger=BOGART, k=30)
Overlapping fragments (Window=1998bp, Step=500bp, Cov*=5X)
Quality -filtered reads (2.5X)
Newbler (ID*=98, Len*=80, e=8)
Hybrid assembly-1
Total=268 MB
N50=5.2kb
Hybrid assembly-2
Total=225 MB
N50=8.1kb
Minimus2
(CONSERR=0.02, OVERLAP=80, MINID=98)
Iterate till no further merging (3X)
Final *R. raphanistrum* assembly
Total=254.6 MB
N50=10.1kb

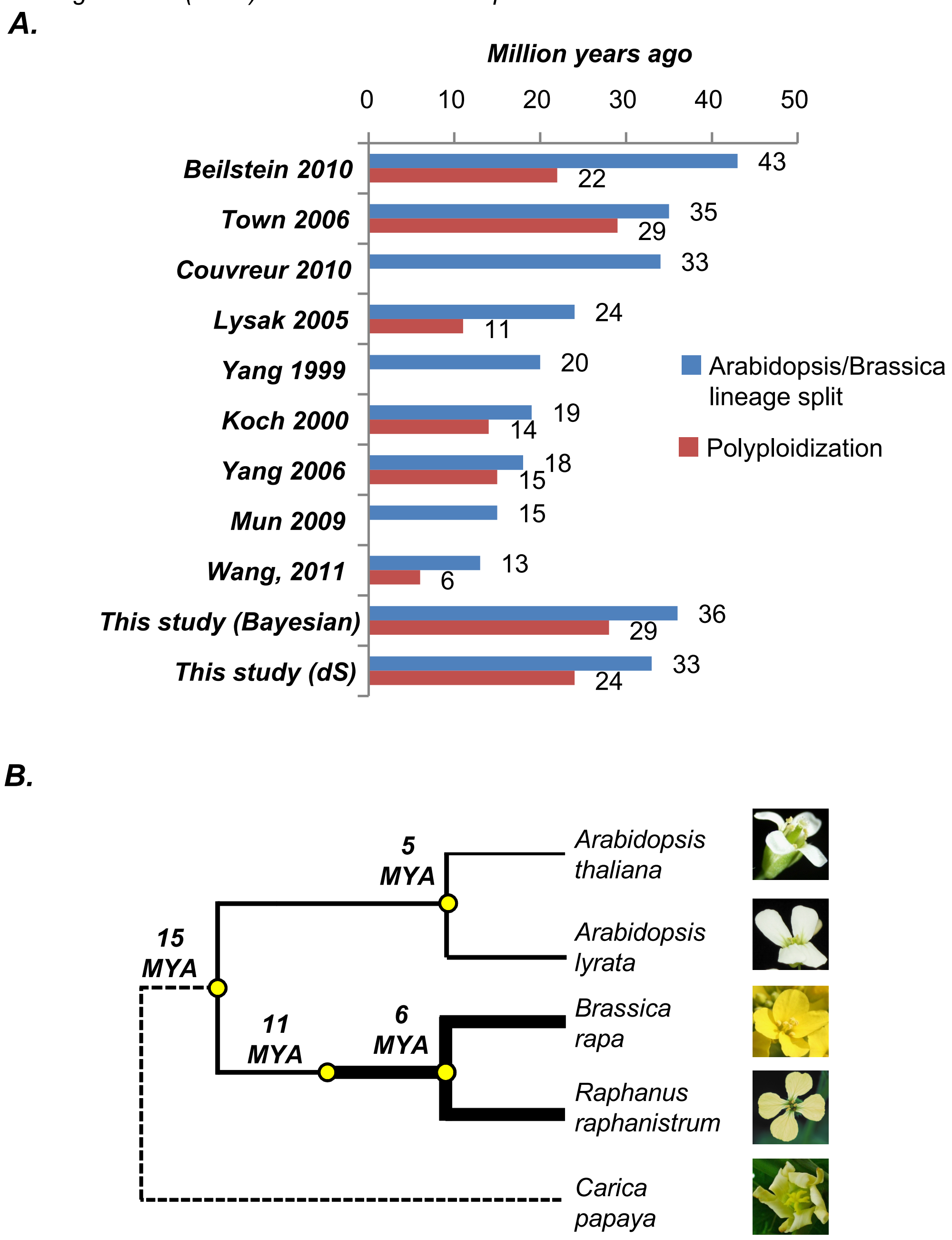
Supplemental Figure 1: Sequencing and assembly of the Raphanus genome. Software and parameters used for each step are noted in red. PE: Paired-End, Cov:Coverage, ID:Identity, Len:Length. See Supplemental Methods section “Genome sequencing, assembly and quality assessment” for more details.


A.
1) ESTs (R. raphanistrum)
2) RNA-seq TxFrags (R. raphanistrum)
R. raphanistrum contigs (RepeatMasked)
EST2Genome
Augustus
SNAP
3X
Training parameters
1) ESTs (B. rapa + R. raphanistrum + R. sativus)
2) RNA-seq TxFrags (R. raphanistrum)
3) Protein-coding genes (A. thaliana)
Augustus
SNAP
74,568 gene models
AED<1
AED=1
HMMSCAN
E-value <0.001
32,407
10,506
Longest isoform
41,122 gene models
* Keyword search based on enrichment of such domains in RR vs BR
AED<0.5 OR significant domain
Discard Transposon/Repeat domain - containing proteins*
38,174 gene models
B.
60
50
40
30
20
10
0
% gene models
All
Filtered
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1.0
1.0
AED score bins

Supplemental Figure 2: Pipeline implemented for annotating the Raphanus genome. (A): Software and parameters used are noted in red. All protein domains related to repetitive elements were discarded in the last step. (B): Distributions of the Annotated Edit Distance (AED) values before and after the penultimate filtering step. RR: Raphanus raphanistrum; BR: Brassica rapa. The 74,568 MAKER-predicted gene models were filtered based on their Annotation Edit Distance (AED) values or the presence of a protein domain as predicted by HMM-PFAM (Eddy, 2008). Two sets of gene models with different levels of accuracy were created: (1) Set I (41,122 models) consisted of gene models with AED≤1, domain E-value<1e-3 and (2) Set II (38,174 models) consisted of models with AED<0.5 or (AED>=0.5 and domain E-value<1e-5). All gene models possessing specific transposon-related domains over-represented in Raphanus vs. Brassica (PF03732.12, PF13975.1, PF03384.9, PF03108.10, PF14392.1, PF14111.1, PF03078.10, PF00075.19, PF13966.1, PF09331.6, PF13456.1) were also discarded from Set II via manual keyword searches. All analyses were performed using Set II gene models given their higher level of agreement with evidence. Functional annotations of gene models were obtained using BLAST2GO (Conesa et al., 2005).

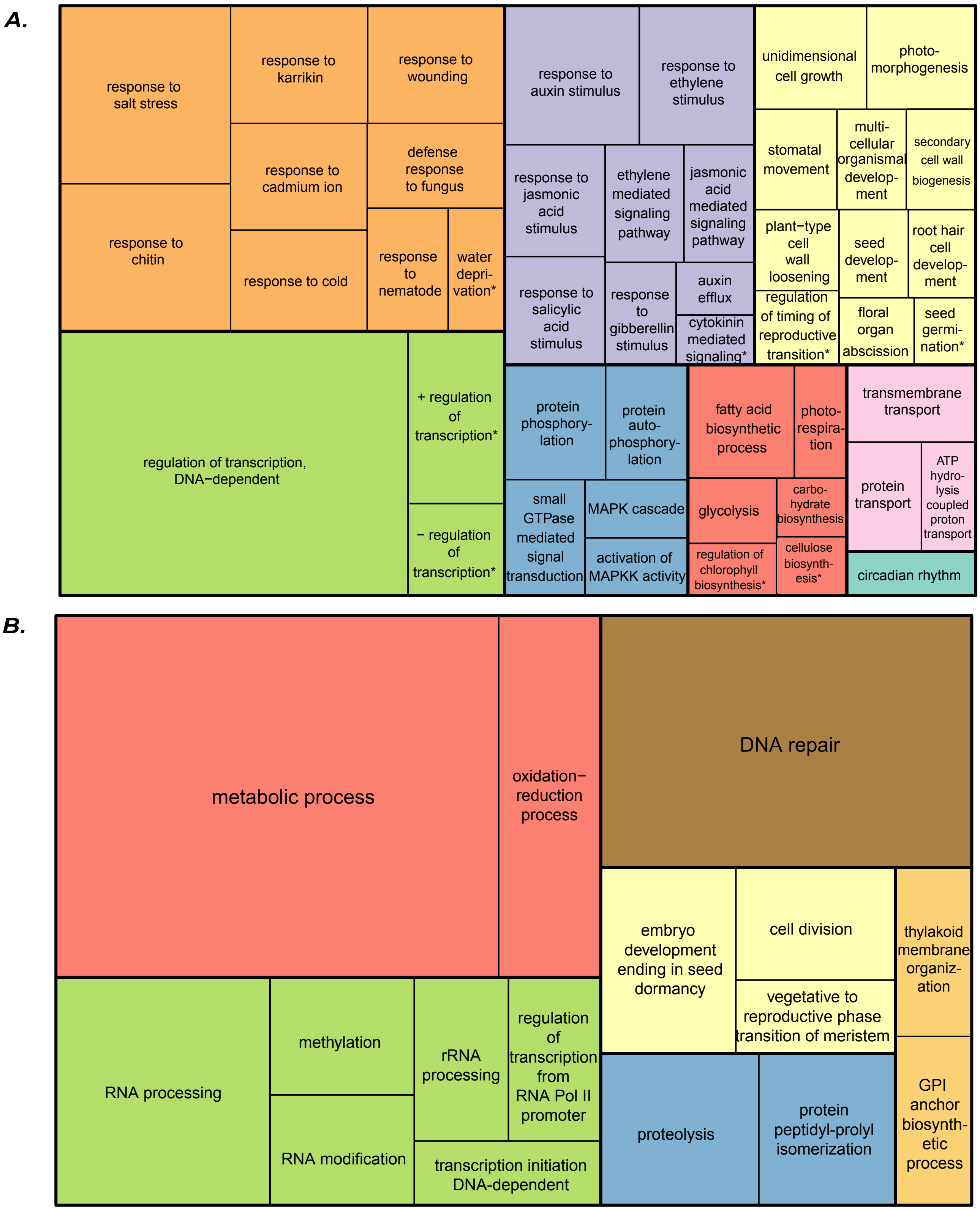


Supplemental Figure 3: Pipeline for defining orthologous groups between A. thaliana, A. lyrata, Brassica, and Raphanus. Software used for each step is noted in red. AT: Arabidopsis thaliana; AL: Arabidopsis lyrata; BR: Brassica rapa; RR: Raphanus raphanistrum; CP: Carica papaya; VV: Vitis vinifera; PT: Populus trichocarpa; OS: Oryza sativa. See Supplemental Methods section “Orthology inference” for more details.

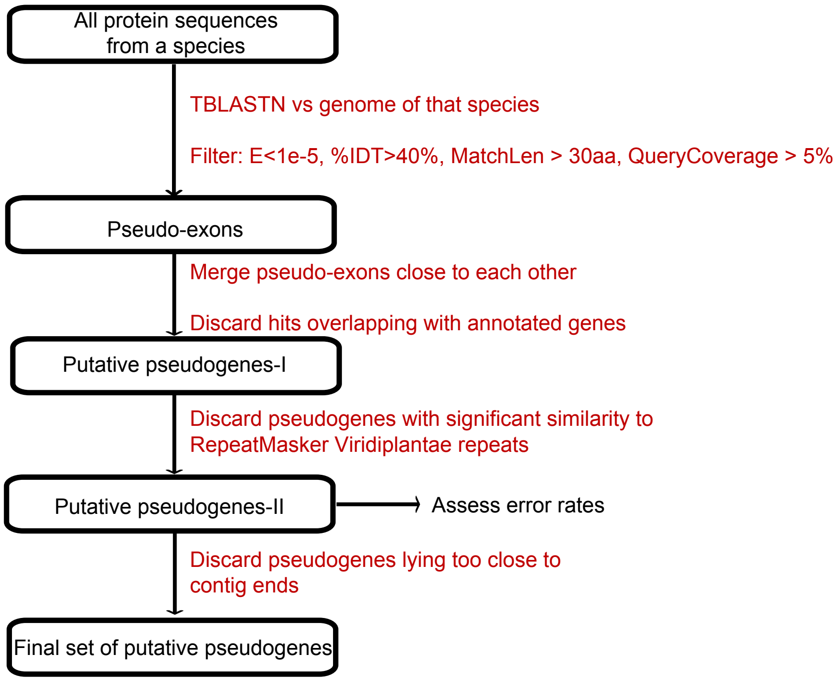


*Timing of events based on dS values, assuming a substitution rate of 15*10e-3 substitutions/site/million years (Koch et al, 2000)*

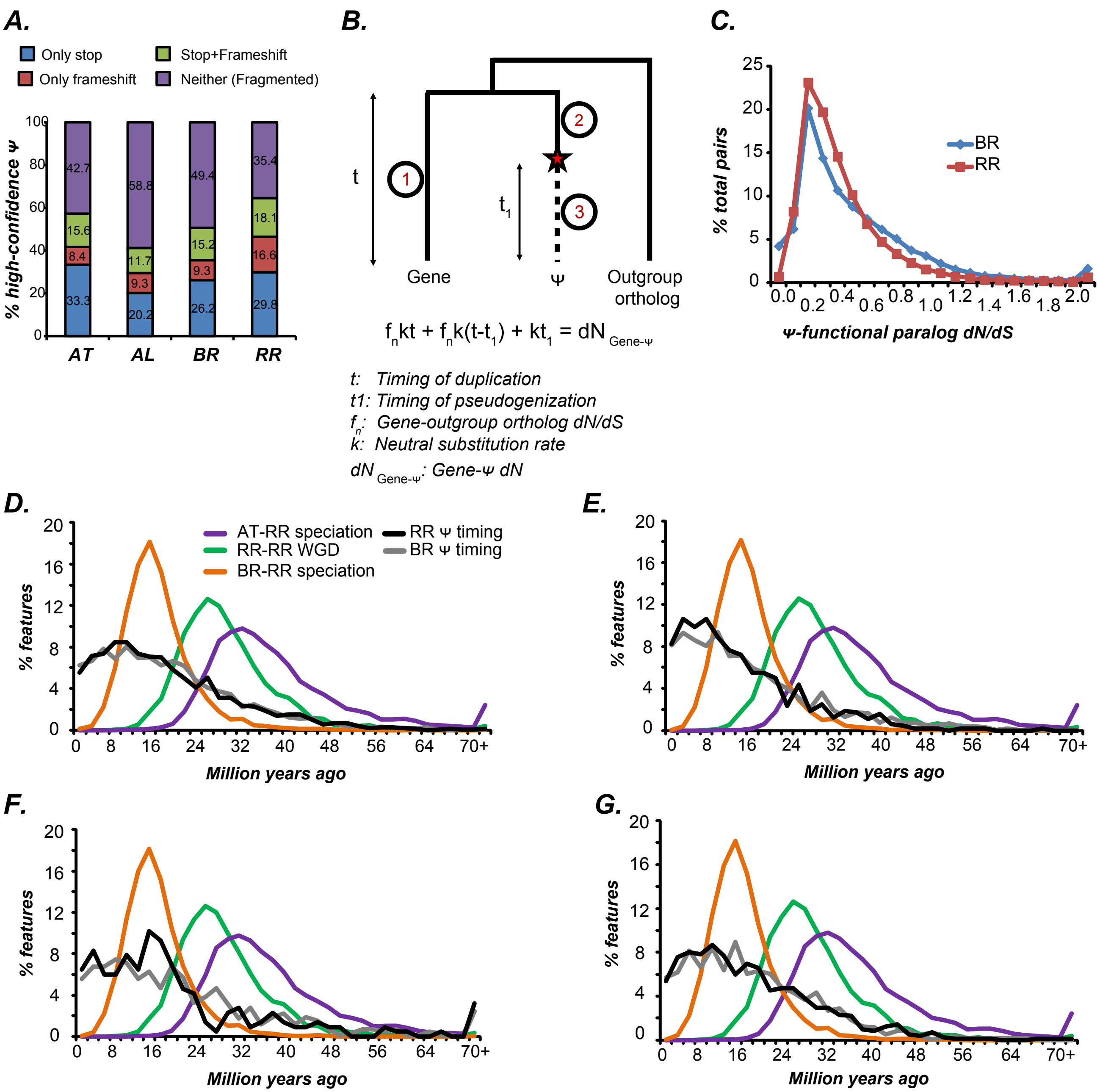
Supplemental Figure 4: Divergence time estimates. (A): Timing of the *A. thaliana-Brassica* split and the triplication event based on previous and this studies. See Supplemental Methods section “Timing of speciation and duplication events” for more details. (B): Timing of speciation and duplicate events calculated with the formula T= dS /(2*n) using a different rate n=15*10-3 substitutions per site per million years (Koch et al., 2000) from that in Figure 1A. MYA: Million years ago.



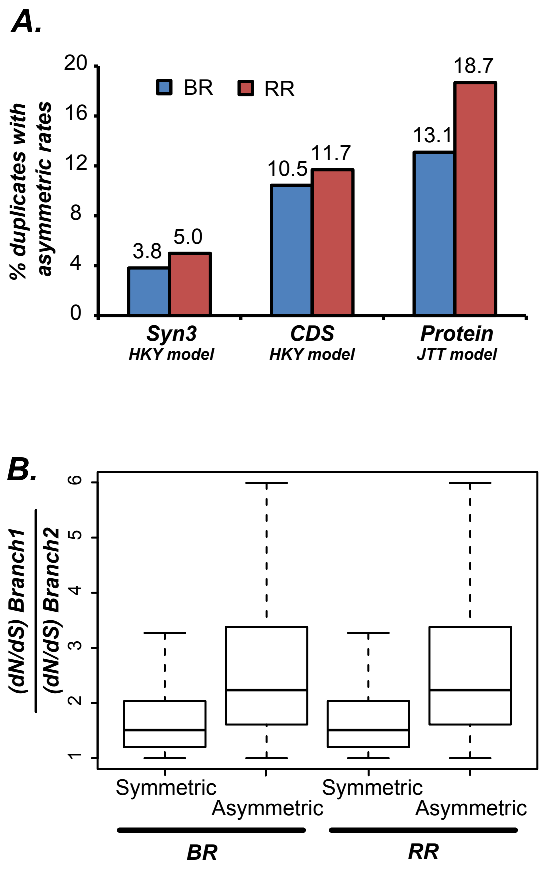
Supplemental Figure 5: GeneOntology categories enriched in (A) retained duplicates and (B) singletons derived from the α' WGT event. Only biological process categories with more than 10 genes and significant test statistics (Fisher Exact Test multiple testing corrected p <0.05 (Storey, 2002)) are shown. The names of the GO categories marked * have been slightly modified to fit into the allotted space. For details regarding GO assignment procedure, see Supplemental Methods: section Gene Ontology and domain enrichment tests.


All protein sequences from a species
TBLASTN vs genome of that species
Filter: E<1e-5, %IDT>40%, MatchLen > 30aa, QueryCoverage > 5%
Pseudo-exons
Merge pseudo-exons close to each other
Discard hits overlapping with annotated genes
Putative pseudogenes-I
Discard pseudogenes with significant similarity to RepeatMasker Viridiplantae repeats
Putative pseudogenes-II
Assess error rates
Discard pseudogenes lying too close to contig ends
Final set of putative pseudogenes

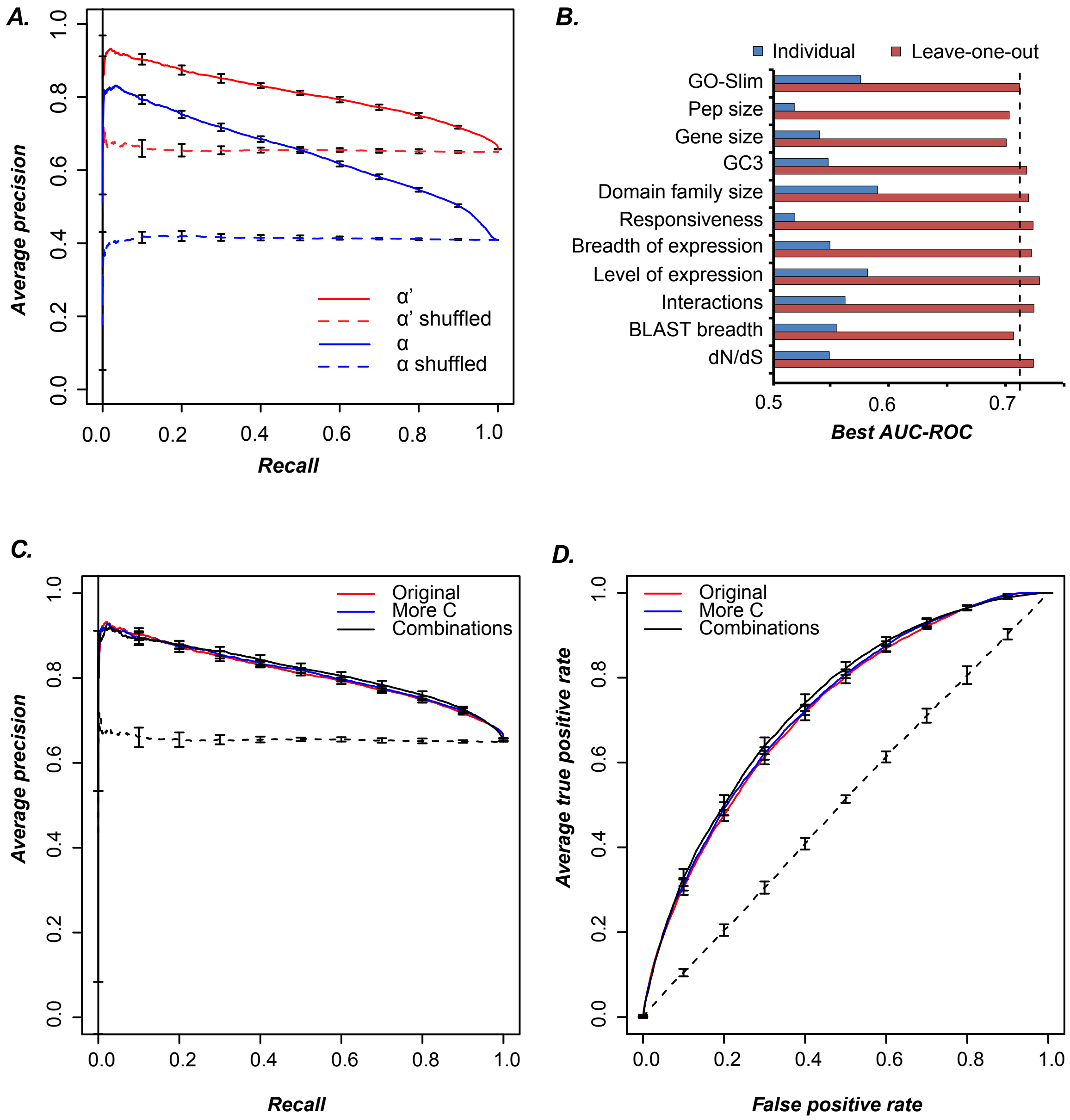
Supplemental Figure 6: Pseudogene prediction pipeline. A modified version of a previously defined pseudogene pipeline (Zou et al., 2009) was used to predict pseudogenes in genomes of all four species under study. See Supplemental Methods: section “Prediction of pseudogenes” for details about the parameters used at each step. Given that the Brassica and Raphanus assemblies have retained ~40,000 genes from the original 90,000 in the neopolyploid after the α' WGT event, we expected to see a large proportion of the ~50,000 lost genes in our predicted pseudogenes. However, we see only 1522-3300, depending on whether we use homeology or synonymous substitution rate as the criteria, respectively (see Methods). We provide additional discussion on the possible reasons for finding such difference in Supplemental Methods section “Evaluating pseudogene predictions”.



Supplemental Figure 7: Patterns of pseudogenization in studied species. (A): Types of pseudogenes identified by the pseudogenization pipeline. (B): Schematic representation of the formula used for estimation of pseudogenization time (Chou et al, 2002), assuming that in (1) and (2) the duplicates experienced selective constraint while in (3) the duplicate evolved neutrally. The red star represents the pseudogenization event. (C): Pseudogene-functional paralog dN/dS for high-confidence pseudogenes. (D-G): To determine whether the timing was robust to the definition of α' pseudogenes, we used four additional methods to estimate pseudogenization timing (1) definition based on dS only, timing using the entire pseudogene sequence (3300 Brassica, 2171 Raphanus pseudogenes) (panel D), (2) definition based on dS only, using only the sequence past the first disabling mutation (Figure 3C), (3) definition based on homeology, using the entire pseudogene sequence (1,522 Brassica, 652 Raphanus pseudogenes) (panel E), and (4) definition based on homeology, using only the sequence past the first disabling mutation (564 Brassica, 215 Raphanus pseudogenes) (panel F). In addition, to identify pseudogenes potentially derived from whole genome duplicates, dS between a pseudogene and its annotated, presumably functional paralog was used. Given the first and third quartiles of the whole genome duplicate dS distribution are 0.2 and 0.6 (Figure 1A), respectively, if 0.2 ≤ pseudogene-paralog dS ≤ 0.6, the pseudogene in question is regarded as derived from whole genome duplication. Changing the range to a more stringent one (0.3 ≤ dS ≤ 0.42) did not influence the estimates significantly (panel G). AT: Arabidopsis thaliana; AL: Arabidopsis lyrata; BR: Brassica rapa; RR: Raphanus raphanistrum



Supplemental Fig. 8: Asymmetric evolution of α' duplicates. (A): Results of relative rates tests between α' duplicates based on HKY (Hasegawa et al. 1985) and JTT (Jones et al. 1992) substitution models. Syn3: synonymous sites at 3rd codon position. CDS: Coding Sequences. Y-axis indicates the % duplicate pairs with significantly different rates of evolution according to Chi-squared test, p≤0.1, after multiple testing correction. See Supplemental Methods section "Relative rates test" for more details. (B): Distributions of the ratio of selective constraints (dN/dS) between the faster evolving branchs and the slower ones. We observed that a statistically significant proportion of orthologous groups (36.9% of the ~450 OGs which had an asymmetrically evolving duplicate) consistently showed asymmetric evolution in both Brassica and Raphanus (z test p<1e-15), possibly a result of their shared ancestry.



Supplemental Figure 9: Performance of duplicate retention machine learning models. (A): Precision/Recall curves for the α and α' model with all features. The dashed lines indicate performance of randomly shuffled datasets. (B): Relative importance of features shown in Figure 5A for predicting α' duplicate retention. The "individual" models were built with only the indicated subset of feature(s), while the "leave-one-out" model was built by including all but the indicated subset of feature(s). Random model has an AUC-ROC (Area Under the Curve-Receiver Operating Characteristics) of 0.5. The averaged AUC-ROC of the full model shown in (A) is indicated with a dotted line. (C): Precision/Recall curves of the α' model after increasing the C parameter in Support Vector Machine) and a model with pairwise combinations of all 60 features (combinations). (D): The AUC-ROC curves of the models in (C). Error bars show the standard deviations over 10 runs.

Supplemental Table 1: Summary statistics of the speciation and WGD time estimates

Species[1]	***Multidivtime***						***Synonymous substitution rate (dS)***			
	Mean	**Median**	**95% CI of mean**	**SD**[2]	**Low 95**[3]	**High 95**[3]	**Mean**	**Median**	**95% CI of mean**	**SD**[2]
AT-AL	11.1	11.3	10.5-11.2	2.3	6.6	14.8	11.3	10.1	10.2-12.5	46.1[1]
AT-BR	36.7	36.5	36.5-36.8	5.4	30.3	49.9	34.3	31.5	34.0-34.6	13.0
AT-RR	36.7	36.5	36.5-36.8	5.4	30.3	49.9	35.2	32.1	34.9-35.6	15.0
BR-RR	19.0	18.8	18.5-19.7	6.5	8.6	34	14.4	13.5	14.3-14.6	7.5
BR-BR	27.4	28.2	27.1-27.7	6.1	16.7	41.5	24.5	23.1	24.2-24.6	7.6
RR-RR	27.8	29.0	27.4-28.2	6.0	17.3	42.1	26.4	24.9	26.1-26.9	11.1

1. AT: *Arabidopsis thaliana*; AL: *A. lyrata*; BR: *Brassica rapa*; RR: *Raphanus raphanistrum.*
2. SD indicates the median standard deviation.
3. Low95 and High95 correspond to the median values of the lower and upper bounds of the 95% Confidence Interval of timing estimates among orthologous groups.

Supplemental Table 2: Datasets used

Feature set	Source	Comments
Functional categories		
GO-Slim categories	TAIR FTP	TAIR v10 annotation. Only biological process categories were used.
Sequence-related features		
Gene and protein sizes, GC3 content	Custom scripts	Values were obtained by analyzing the FASTA and GFF files.
PFAM Domain size	HMMER	Hidden Markov Models were obtained from Pfam HMMER3/b [3.0, March 2010]. Domain designations of AT, BR and RR proteins were obtained by running HMMER with the options –*cut_tc –noali* and further filtering the domains with Evalue<1e-5
Expression-related features		
Breadth and level of expression (NASCarray)	NASCArray	Pearson's Correlation Coefficient was calculated between NASCArray datasets (accessed Sep 2012) using the ATH1 microarray. Of the datasets with > 0.98 PCC, only 1 representative dataset was kept. Breadth and level of expression were calculated for the remaining 1779 datasets, after excluding multigene probes. Low/Medium/High expression levels and breadth were defined as <25th percentile, 25th-75th percentile and >75th percentile of the entire distribution.
Biotic and abiotic responsiveness	ATGenExpress	Genes showing more than 2 fold up or down regulation (*q*-value < 0.05) in at least one conditions were defined as responsive to stress.
RNA-seq	Previously published data	Data from Moghe et al, 2013 was used for this study. Low/Medium/High expression levels and breadth were defined as <25th percentile, 25th-75th percentile and >75th percentile of the entire distribution. Low/Medium/High expression breadth was defined as expression in 0-3, 3-5 and 5-8 datasets respectively.
Network-related features		
Number of interacting partners	Aranet	Number of interactions in the integrated Aranet network inference were used. AraNet is a probabilistic functional gene network i.e. the edge indicates the probability that two nodes (genes) interact. For higher stringency, only those interactions with a log likelihood score > 1 were used.
Conservation-related features		
Breadth of conservation across plants	Phytozome	Data from Phytozome v5 was used. TBLASTN was performed between *A. thaliana* or *Brassica/Raphanus* peptide sequences (Query) and the genome fasta sequence of all Phytozome species (Subject). All hits with E>1e-10 were eliminated. Number of species with significant hits was enumerated.
dN/dS values	PAML, custom script	*dN/dS* was calculated between orthologs using the yn00 function in the PAML package. To obtain one *dN/dS* value for each AT gene, the average *dN/dS* value between *A. thaliana -Brassica* and *A. thaliana -Raphanus* orthologs was computed and used for this analysis.

Supplemental Methods

Genome sequencing, assembly and quality assessment

Raphanus is an obligate out-crosser. To reduce the amount of heterozygosity in the genome, *Raphanus* subspecies *raphanistrum* (weedy) from the Binghampton population in New York was inbred for five generations. Total DNA was extracted from the leaves of the 5th generation inbred plants using Qiagen DNEasy Maxi kit. The extracted DNA was ethanol precipitated and assessed for quality using CHEF gel electrophoresis. For 454 sequencing, DNA was sheared by Covaris sonication, size selected by gel electrophoresis and a 3 kb mate pair library constructed according to manufacturer's instructions (Roche-454). A total of 6 full plates and three half-plates were sequenced using the Titanium chemistry. DNA was further sheared and an Illumina fragment library constructed (average 516 bp). A total of 7 lanes of 100 bp paired-end sequence was generated on an Illumina GAII sequence analyzer.

Before assembly, Illumina reads were trimmed from the 3' end to a Phred quality score ≥20 and length ≥50. The 454 reads were split at linker sequences and only reads with mate pairs were used for assembly. The filtered Illumina and 454 reads represented a 47X and 2.5X coverage of the estimated 573Mb genome. To assemble the *Raphanus* genome, we used three different approaches. We first created an Illumina-only assembly using ABySS 1.2.5 (Simpson et al., 2009) with the optimal kmer length (k=39). We then split the Illumina contigs into overlapping fragments of 1998bp with 1000bp step size at a coverage of 10X per fragment. These split Illumina contig fragments and the quality-filtered 454 reads were used as input to Newbler 2.5.3 (Margulies et al., 2005) to create a hybrid assembly. The following parameters were used for the Newbler assembly*: -large -mi 98 -cpu 1 -ml 80 -ud -rip -m -e 8*. The Newbler assembly showed a marginal improvement in N50 and total assembly size compared to an Illumina-only assembly (Supplemental Figure 1). In the second approach, the *Raphanus* assembly was generated with the Celera Assembler using split 454 reads (sffToCA program, option "*-clear 454 -trim chop*") and Illumina reads trimmed (https://github.com/tanghaibao/trimReads) so the base quality was at least Phred 20. We ran Celera Assembler version 6.1 with unitigger "BOGART" with kmerSize=30 (Miller et al., 2008). Finally, we used the program Minimus2 (Sommer et al., 2007, 2) from the AMOS 3.1.0 package to merge the ABySS/Newbler and the Celera assemblies. The Minimus2 merging step was repeated three times with the merged and the unmerged contigs till convergence. The final Minimus2 assembly was substantially better than the ABySS/Newbler and the Celera assemblies (Supplemental Figure 1) and was used in all subsequent analysis. All the Illumina and 454 reads have been deposited in NCBI SRA (PRJNA209513).

Orthology inference

Using both synteny information as well as gene-species tree reconciliation, we determined orthologous groups (OGs) between 4 Brassiceae species: *A. thaliana*, *A. lyrata*, *Brassica* and *Raphanus* (see Methods, Supplemental Figure 3). A combination of two approaches were used –similarity-based and synteny based (Supplemental Figure 3). In the similarity-based approach, an all-against-all BLAST (Altschul et al., 1997) search was performed between protein sequences from eight species: *A. thaliana*, *A. lyrata*, *Brassica*, *Raphanus, C. papaya, P. trichocarpa, V. vinifera and O. sativa.* The matches with E-value<1e-20, identity>50%, coverage>60%, and match Length>60aa were defined as significant and included in further analysis. Proteins with significant matches were assigned to groups resembling gene families with the single linkage algorithm. Protein sequences in each group were aligned using MAFFT (Katoh et al., 2002), and a phylogenetic tree for each group was generated using RAxML with the PROTGAMMAJTT model and 100 bootstrap replicates (Stamatakis, 2006). These gene trees were midpoint rooted using the retree function in PHYLIP (Felsenstein, 1989) and reconciled with the species tree as defined in Phytozome (Goodstein et al., 2012) using Notung (Chen et al., 2000) to identify orthologous groups.

The N50 of the *Raphanus* assembly is relatively small at 10.1kb, and it is not useful for determining synteny between *Raphanus* and other genomes. Because extensive chromosomal synteny between *R. sativus* and *Brassica* species is known (Li et al., 2011), *Raphanus* gene models were mapped to the *Brassica* scaffolds for establishing a “pseudo-synteny” with GMAP v 2013-03-31. The best matching *Raphanus* sequences were included for further analysis if their coverage and identity was > 70%. To identify syntenic regions between *A. thaliana*, *A. lyrata*, *Brassica* and *Raphanus*, an all-against-all BLAST was performed between these four species and matches filtered with E-value<1e-10, identity>60%, coverage>60%, and match length>60aa. The filtered matches were used as input to MCScanX (Wang et al., 2012), along with locations of *A. thaliana*, *A. lyrata*, *Brassica* (Wang et al., 2011) and *Raphanus* (based on pseudo-synteny) genes. Genes were placed in syntenic blocks with ≥5 genes and with gap≤10 intervening genes. This approach allowed identification of syntenic regions between species as well as associated homeologous blocks derived from whole genome duplications. Significant matches in inferred syntenic blocks between species were regarded as potential orthologs, while matches in inferred homeologous blocks were regarded as potential paralogs derived from whole genome duplications. The α duplication event took place before the divergence between the Brassiceae species analyzed, and we noticed that some of the syntenic blocks contain homeologous regions derived from the earlier α event. Therefore, similar genes in these syntenic blocks may belong to multiple orthologous groups. To further define orthologous relationships among these four species using the synteny information, similar genes in each block were aligned for phylogenetic reconstruction in the same way as noted earlier. The phylogeny was then reconciled with the four species tree to identify putative orthologous groups of genes.

Orthologous pairs obtained using both the above approaches were combined together using single linkage clustering to generate the final set of 23,660 orthologous groups between the Brassiceae species. Our strategy allowed the assignment of orthologous relationships between 21,371 (77.9%) *A. thaliana*, 21,294 (65.2%) *A. lyrata*, 29,564 (72.0%) *Brassica* and 24,567 (64.5%) *Raphanus* genes. Genes in *Raphanus* that could not be assigned to orthologous groups tend to be significantly shorter than those that could be assigned (Kolmogorov–Smirnov test, p<1e-15). For identifying retained duplicates and singletons, we stringently discarded putative tandem duplicates, which were defined as genes with high similarity (E<1e-10) lying within 20 genes on either side of a gene of interest on the chromosome. Only 21,525 *Brassica* and 15,030 *Raphanus* genes lying in either within-genome (homeologous) or between-genome syntenic blocks with *A. thaliana* or *A. lyrata*, were considered to be derived from the α' event and used to distinguish retained duplicates and singletons. A total of 16,557 OGs satisfied these criteria. The numbers reported in the main text are from this set; however, analyses with all 23,660 OGs also produced similar results. *A. thaliana* duplicates derived from the α polyploidization event were obtained from a previous study (Bowers et al., 2003).

Timing of speciation and duplication events

Previous studies have estimated the timings of the speciation and duplication events in Brassicaceae. However, many of these estimates were obtained using a now unavailable fossil pollen as a calibration point or were based on synonymous substitution rate derived from two individual loci (Koch et al., 2000) or 3) or assumed a constant rate of evolution across the Brassicaceae family. These issues have been reviewed in a previous study (Beilstein et al., 2010). Based on the relative rate test (Goldman and Yang, 1994), the synonymous substitution rate (d_S) at the third codon position of singletons did not increase significantly after the polyploidization event, consistent with the molecular clock assumption **(**see Methods**)**.

Therefore, the third codon position d_S can be used for determining the age of the α' WGT event and the *Brassica*-*Raphanus* speciation event.

Two methods were used to determine duplication and speciation time. In the first approach, synonymous substitution rate (d_S) was calculated between pairs of singleton genes and between pairs of retained duplicates using the codeml function in PAML (Yang, 2007). Divergence time was obtained using the formula T=d_S/(2*neutral rate). As expected, if dates are estimated using the previously used substitution rate of 15*10^{-3} substitutions/site/million years (Koch et al., 2000), the median ages of different events are almost halved (Supplemental Figure 4B).

In the second approach, we used multidivtime (Rutschmann F., 2005), a Bayesian dating method that not only considers the rate of evolution but also allows priors to be set. First *C. papaya* genes were assigned to a predicted Brassicaceae orthologous group as an outgroup if a *C. papaya* gene had a significant hit to ≥1 Brassicaceae species analyzed and no hit to any other orthologous group. Although the second criterion is stringent and a number of *C. papaya* genes were not assigned, this consideration reduced the false positive rate in the outgroup selection. Orthologous groups with only one gene from each of the four Brassicaceae species and *C. papaya* were used to determine the timing for speciation. The timing of whole genome duplication in *Brassica* and *Raphanus* was estimated using retained duplicates with the *A. thaliana* orthologs as outgroups. A synonymous site substitution rate of 7*10^{-3} substitutions/site/million years (Ossowski et al., 2010) was used to calculate the speciation and duplication time, with a prior age of 36 million years between the root and the tip of the four Brassiceae species phylogeny. The lower and upper bounds for *A. thaliana*-*A. lyrata* and *A. thaliana*/*A. lyrata*-*Brassica*/*Raphanus* divergence time were set at 5-15 and 30-90 million years ago (MYA), respectively. Multidivtime was run using all default parameters except *bigtime=100*.

Relative rates test

To determine if duplicates differed significantly in their evolutionary rates, the PyCogent implementation of the Relative Rates test was used (Knight et al., 2007). The HKY85 and JTT92 models were used for nucleotide and protein sequences, respectively. Branch-wise d_N/d_S was estimated using the codeml package in PAML (Yang, 2007) after aligning the coding sequences of a *Brassica* or *Raphanus* gene and using *A. thaliana* ortholog as outgroup with PRANK (Kosiol et al., 2007). A free-ratios model, which assumes an independent ratio for each branch, was used for running codeml.

To determine whether the rate of evolution is similar on the *A. thaliana* and *Brassica* branches, a relative rate test was conducted for 1:1 orthologs between *A. thaliana*-*Raphanus* and between *A. thaliana*-*Brassica* with *C. papaya* orthologs as outgroups. Of the 1,177 1:1 orthologous pairs we analyzed, only 8 (0.7%) *A. thaliana*-*Raphanus* pairs and 1 (0.1%) *A. thaliana*-*Brassica* pair departed from a constant rate model after correcting for multiple testing (χ^2 test, *p* ≤0.10). In all comparisons, we corrected for multiple testing using the Q-value package in R (Storey, 2002) and only considered pairs with significantly different rates (χ^2 test, *p*≤0.1) as evolving asymmetrically.

Prediction of pseudogenes

A modified version of a previously defined pseudogene pipeline (Zou et al., 2009) was used to predict pseudogenes in genomes of all four species under study (Supplemental Figure 6). Specifically, we performed TBLASTN using protein coding genes as the query and genomic sequences as the subject using BLAST 2.2.25. We then filtered the output using the thresholds: E-Value < 1e-5, %Identity > 40%, Match Length>30aa and Coverage > 5% of the query sequence to obtain pseudo-exon definitions. Pseudo-exons in close proximity to each other (based on the 95th percentile of the intron length distribution) and having matches to the same protein were then joined together to form putative pseudogenes based on their Smith-Waterman score. Putative pseudogenes overlapping with annotated protein coding regions were removed from the dataset. In addition, pseudogenes with significant similarity to known *Viridiplantae* repeats (Cutoff=300, Divergence=30) as determined by RepeatMasker 3.3.0 were discarded (Supplemental Figure 6).

To assess the error rate of misclassifying a gene as a pseudogene, four analyses were conducted. First, we found that 9.6% and 12.9% of the predicted *Raphanus* pseudogenes with or without a disabling mutation, respectively, have ≥5 reads compared to 61.3% of the protein-coding genes. Second, the median sequencing coverage is 70X for predicted pseudogenes, suggesting that the chance of a sequence being erroneously called a pseudogene due to low read coverage or sequencing errors is low. Third, we analyzed RNA-seq data from *Raphanus* flower and based on presence of ≥5 reads, we found 10.8% and 61.3% of the pseudogenes and the protein-coding genes expressed, respectively. This is similar to our earlier study in *A. thaliana* where we found 10.3% pseudogenes and 79.6% protein-coding genes expressed based on the same criterion (Moghe et al., 2013). Finally, the predicted pseudogenes have significantly higher d_N/d_S values compared to functional ortholog and paralog pairs (KS test p<1e-15, Figure 3B). These findings suggest that the error rate of pseudogene prediction was low enough to not affect our further analyses.

Because of the fragmentary nature of the *Brassica* and *Raphanus* genomes, there was a high false positive rate due to proteins split between contigs being counted as pseudogenes. To reduce the false positive rate, high confidence pseudogenes were determined using a custom python script. Specifically, a pseudogene is considered a high-confidence pseudogene if it contains stop codons or frame-shifts or if it passes a particular test. This test states that a protein is a high confidence pseudogene if $X_U >= Y_U + Z$ and $X_D >= Y_D + Z$, where X_U and X_D are the absolute distances between the pseudogene and the each end of the contig it is on for both sides of the pseudogene, upstream and downstream relative to the orientation of the matching protein, respectively, and where Y_U and Y_D are the absolute distances between the matching region on the protein and the end of the protein for both sides of the protein, upstream (N-terminal side) and downstream (C-terminal side), respectively, and where Z is the 95th percentile intron length for the species being tested.

The number of detectable pseudogenes is higher in post-α'-polyploidization species compared to *A. thaliana*/*A. lyrata*. For each annotated protein-coding gene in *A. thaliana* and *A. lyrata*, there exists 0.15 and 0.34 pseudogene, respectively. In contrast, there is 0.96 and 0.56 pseudogene/annotated gene for *Brassica* and *Raphanus*, respectively (or, after correcting for the fragmentary nature of the *Brassica* and *Raphanus* genomes, 0.82 and 0.35, respectively). The low proportion of pseudogenes/annotated genes in *Raphanus* is likely because of the incomplete *Raphanus* assembly as well as possible overcorrection for fragmentation. The pseudogene numbers obtained for *Brassica* and *Raphanus* are likely to be an underestimate of the actual number of pseudogenes derived from transposition events, given that the repetitive genomic fraction was largely missed in both assemblies. In addition, putative pseudogenes resembling repeats – 5,060 *Brassica* pseudogenes and 518 *Raphanus* pseudogenes – were discarded. There are substantially fewer repeat-related pseudogenes in *Raphanus* most likely because of the lower coverage of the *Raphanus* genome than the *Brassica* genome.

Evaluating pseudogene prediction

Given that the *Brassica* and *Raphanus* assemblies have retained ~40,000 genes from the original 90,000 in the neopolyploid after the α' WGT event, we expected to see a large proportion of the ~50,000 lost genes in our predicted pseudogenes. However, we see only 1522-3300, depending on whether we use homeology or synonymous substitution rate as the criteria, respectively (see Methods). This low number is most likely not a consequence of the partial genome assemblies given that we can detect >90% of the ESTs in both *Brassica* and *Raphanus* assemblies. It is also most likely not due to false negatives of the pseudogene identification pipeline because a similarity search using TBLASTN between 13,720 previously identified homeologous blocks in *Brassica* genome (Wang et al., 2011) could only detect similar sequences for 2124 (15.4%) genes, comparable to what we find using the pseudogene identification pipeline. Specifically, we analyzed 13,720 homeologous blocks in the *Brassica* genome which had at least 1 duplicate gene loss on either of its three subgenomes. If the lost gene was still present as a pseudogene, a TBLASTN search in the block using the retained duplicate gene copies as queries would help identify the pseudogenized copy. However, as noted above, missing genes could be detected in only 15.4% of the homeologous blocks. The thresholds used for filtering the TBLASTN results were E<1e-5, % identity>40%, coverage>5% and a match length>20% of the query.

These observations may be explained by four scenarios – 1) A significant proportion of the duplicate genes were lost via deletion and no longer exist in the genome, 2) The pseudogenes have mutated beyond recognition by BLAST and 3) A significant proportion of pseudogenization has occurred via transposon insertion and subsequent fragmentation – such pseudogenes would be discarded from our analysis in the RepeatMasker step. Under these scenarios, a gene loss event would not be detected by either BLAST or our pseudogene identification pipeline.

Timing of pseudogenization

To estimate the timing of pseudogenization, we used a published approach (Supplemental Figure 7B) (Chou et al., 2002). To determine whether the timing was robust to the definition of α' pseudogenes, we used four additional methods as described in Supplemental Figures 7D-G. To determine whether our findings are robust to different estimates of duplication time in the timing formula, we defined duplication times using three methods: 1) a fixed duplication time of 25 MYA, 2) random sampling from a Gaussian distribution with mean=25 and sd=7 (based on the functional duplicate gene d_S distribution) and 3) calculating the duplication time based on the d_S between pseudogene and the parent gene. In all cases, the distributions obtained for pseudogenization timing were very similar and do not affect our interpretations. All timing estimates ≤0 were discarded.

Gene Ontology and domain enrichment tests

Gene Ontology descriptions were obtained from The Arabidopsis Information Resource (ftp://ftp.arabidopsis.org/Ontologies/Gene_Ontology/). All protein-domain information was obtained using the HMMSCAN software v3.0 (Eddy, 2008) using previously defined thresholds of Pfam Hidden Markov Models (HMMER3/b [3.0, March 2010]. All enrichment tests were performed using Fisher Exact's Test in R and the Q-values for enrichment were determined using the Q-value package in R (Storey, 2002).

Classifying retained duplicates and singletons with machine learning

We used Support Vector Machine (SVM) to generate classifiers that allow distinguishing between retained duplicates and singletons. The feature sets used in this study are detailed in Supplemental Table 2. We should point out that these features can be dependent, e.g. higher GC3 content has been shown to be correlated with stronger purifying selection, greater codon usage bias and higher frequency of DNA methylation (Elhaik and Tatarinova, 2012), and may be associated with expression-related characteristics of retained duplicates. Similarly, higher conservation among retained duplicates may be associated with their biological roles, network connectivity and expression profiles.

For all quantitative features, we binned the values into four quartiles based on the feature value distribution across all genes. All other features (GO-Slim categories and responsiveness to biotic or abiotic stress) were treated as discrete categories. The 4702 retained duplicates and 2533 singletons were assigned roughly equally and randomly to the training and the test dataset. The random split was repeated ten times. SVM-Light (Joachims, 1999) was used to generate classifiers and feature weights. A grid search was performed to determine the optimal SVM parameters. Increasing the C sampled from 1e-06 to 1000, with 10-fold change or using pairwise combinations of all features did not result in any improvement in the AUC and Precision/Recall curves (Supplemental Figure 9C,D). Using a radial basis function with varying gamma values from 1e-06 to 1, with 100-fold change for the next value, also did not result in improved model performance.

Buffering as a means for duplicate gene retention

The buffering model stipulates that duplicate genes may be retained to serve as buffer against disruption of crucial functions (Chapman et al., 2006). However, evolution cannot see into the future and hence, whether buffering can retain duplicate genes to limit the impact of disruption of an important gene is an important question. To explain the mechanism of gene retention due to buffering, Nowak and others proposed four different scenarios under which such retention may occur (Nowak et al., 1997), based upon varying degrees of efficacies of the gene product (eg: activity of an enzyme), mutation rates and pleiotropy between two genes performing the same function. One scenario allowing “redundant” duplicate to be retained is that two genes A and B (with their non-functional alleles *a* and *b*) perform the same function but with different efficacies such that efficacy of A > efficacy of gene B, but the mutation rate of A > mutation rate of B. Under this situation, gene B, although with a lower efficacy, does not mutate as frequently as the higher efficacy gene A. Thus when gene A is mutated into the *a* allele, gene B is under selection and maintained (Nowak et al., 1997).

Although this mechanism and other scenarios presented in Nowak et al. 1997 study provide a theoretical explanation for how buffering can occur, the frequency with which these scenarios occur in nature is not clearly understood. Nevertheless, under the buffering model, the mutation rate of the genes for which there is a selection for maintaining redundancy will be constrained since any mutation that disrupts redundancy will be selected against. Under conditions where the selective advantage conferred by the redundant copy is greater than the frequency with which A→a conversion occurs and where the population size is high, both copies of the gene may be maintained, at least for a few initial generations (Lynch et al., 2001), giving the retained duplicate genes a chance to sub-functionalize or neo-functionalize.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Arumuganathan, K. and Earle, E.D. (1991). Nuclear DNA content of some important plant species. Plant Mol. Biol. Report. **9**: 208–218.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. **107**: 18724–18728.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**: 433–438.

Chapman, B.A., Bowers, J.E., Feltus, F.A., and Paterson, A.H. (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc. Natl. Acad. Sci. U. S. A. **103**: 2730–2735.

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. **7**: 429–447.

Chou, H.-H., Hayakawa, T., Diaz, S., Krings, M., Indriati, E., Leakey, M., Paabo, S., Satta, Y., Takahata, N., and Varki, A. (2002). Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc. Natl. Acad. Sci. U. S. A. **99**: 11736–11741.

Eddy, S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput. Biol. **4**.

Elhaik, E. and Tatarinova, T. (2012). GC3 biology in eukaryotes and prokaryotes. arXiv:1203.3929.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5**: 164–166.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**: 725–736.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40**: D1178–1186.

Joachims, T. (1999). Making Large-Scale Support Vector Machine Learning Practical (MIT Press, Cambridge, MA).

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**: 3059–3066.

Knight, R. et al. (2007). PyCogent: a toolkit for making sense from sequence. Genome Biol. **8**: R171.

Koch, M.A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol. Biol. Evol. **17**: 1483–1498.

Kosiol, C., Holmes, I., and Goldman, N. (2007). An empirical codon model for protein sequence evolution. Mol. Biol. Evol. **24**: 1464–1479.

Li, F., Hasegawa, Y., Saito, M., Shirasawa, S., Fukushima, A., Ito, T., Fujii, H., Kishitani, S., Kitashiba, H., and Nishio, T. (2011). Extensive chromosome homoeology among Brassiceae species were revealed by comparative genetic mapping with high-density EST-based SNP markers in radish (Raphanus sativus L.). DNA Res. **18**: 401–411.

Lynch, M., O'Hely, M., Walsh, B., and Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. Genetics **159**: 1789–1804.

Margulies, M. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**: 376–380.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**: 2818–2824.

Moghe, G.D., Lehti-Shiu, M.D., Seddon, A.E., Yin, S., Chen, Y., Juntawong, P., Brandizzi, F., Bailey-Serres, J., and Shiu, S.-H. (2013). Characteristics and significance of intergenic polyadenylated RNA transcription in Arabidopsis. Plant Physiol. **161**: 210–224.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature **388**: 167–171.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science **327**: 92–94.

Rutschmann F. (2005). Bayesian molecular dating using PAML/multidivtime. A step-by-step manual.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. Genome Res. **19**: 1117–1123.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. BMC Bioinformatics **8**: 64.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688–2690.

Storey, J.D. (2002). A direct approach to false discovery rates. J. R. Stat. Soc. B **64**: 479–498.

Wang, X. et al. (2011). The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. **43**: 1035–1039.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J.C., and Paterson, A.H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**: e49.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**: 1586–1591.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R., and Shiu, S.-H. (2009). Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol. **151**: 3–15.