Article title: Convergent and/or parallel evolution of RNA-binding proteins in angiosperms after polyploidization
Authors: Liangyu Guo, Shuo Wang, Xi Jiao, Xiaoxue Ye, Deyin Deng, Hua Liu, Yan Li, Yves Van de Peer and Wenwu Wu
Article acceptance date: 20 February 2024

The following Supporting Information is available for this article:

Fig. S1 Identification of RBPs in 21 selected angiosperms.

Fig. S2 Synonymous substitution (*Ks*) ranges associated with the well-documented WGD events in the 21 selected angiosperm species.

Fig. S3 Overview strategy of cold-responsive RNA-seq experiments and expression profiles of leaf development-related genes before and after cold stress.

Fig. S4 Two examples for redefining orthogroups.

Fig. S5 Cold-induced TF duplicates were biased retained after R-WGD.

Fig. S6 Diverse cold-responsive transcriptome datasets consistently demonstrate the cold-upregulation of R-WGD-derived RBP genes.

Fig. S7 Alternative splicing seems to be enriched in genes with longer introns.

Fig. S8 Functional enrichments of cold-induced differentially alternatively spliced genes.

Fig. S9 The GRP7-GRP8 duplicates resulted from β-WGD event (R1-WGD).

Fig. S10 *GRP7* expression was significantly correlated with the average expression of 433 cold-induced GRP7 targets.

Fig. S11 A significant association of 433 cold-induced GRP7 targets with differentially expressed genes in *GRP7*-overexpressing and loss-of-function plants s.

Fig. S12 Overlapping and unique genes between 433 cold-induced GRP7 targets and

differentially alternatively spliced genes.

Fig. S13 Expression profile of the top 30 transcription factors that bind to promoters of cold-induced GRP7 targets.

Fig. S14 Negative correlation between RBP gene number and the WGD age.

Fig. S15 Cold-induced tandem duplicates mainly generated around global cooling periods.

Table S1 Source information of 21 selected angiosperm species.

Table S2 A high-confidence list of RBPs in *Arabidopsis*.

Table S3 Gene accessions, protein domains, and duplication modes of RBPs in 21 selected angiosperms.

Table S4 A statistical summary of duplication modes of RBP genes and non-RBP genes in angiosperm genomes.

Table S5 A total of 4,594 gene families, including 494 gene-rich families, were identified in angiosperm genomes.

Table S6 The well-documented WGD events and their predicted *Ks* ranges for collinear pair genes in 21 selected angiosperms.

Table S7 RBP duplicates retained from different periods of WGDs in 21 selected species.

Table S8 Expression values of RBP genes in eight cold-treated species.

Table S9 Cold-induced differentially alternative splicing genes in eight cold-treated species.

Table S10 A total of 282 RBP orthogroups identified in eight cold-treated species.

Table S11 A total of 857 GRP7 targets, including 433 cold-induced genes.

Table S12 108 cold-induced TFs from a library of 540 TFs with ChIP/DAP-seq data.

Table S13 Summary of the 433 cold-induced genes bound by the 108 cold-induced TFs.

Table S14 Retention status of RBP genes and 494 gene-rich families in angiosperms after WGDs.

Table S15 RNA-seq expression values of leaf development-related genes in eight cold-treated species.

Table S16 A statistical summary of cold-induced RBPs retained from WGD, TSP, TD, DD, and Singleton.

Table S17 Details of cold-induced duplicates from WGDs in eight cold-treated species.

Table S18 Eighteen additional datasets of cold-responsive transcriptomes.

Table S19 29 overlapping RBP orthogroups retained and cold-induced in angiosperms.

Table S20 Timing of WGD events in producing *GRP7/8*-containing collinear gene pairs.

Dataset S1 The programming code used in this study.

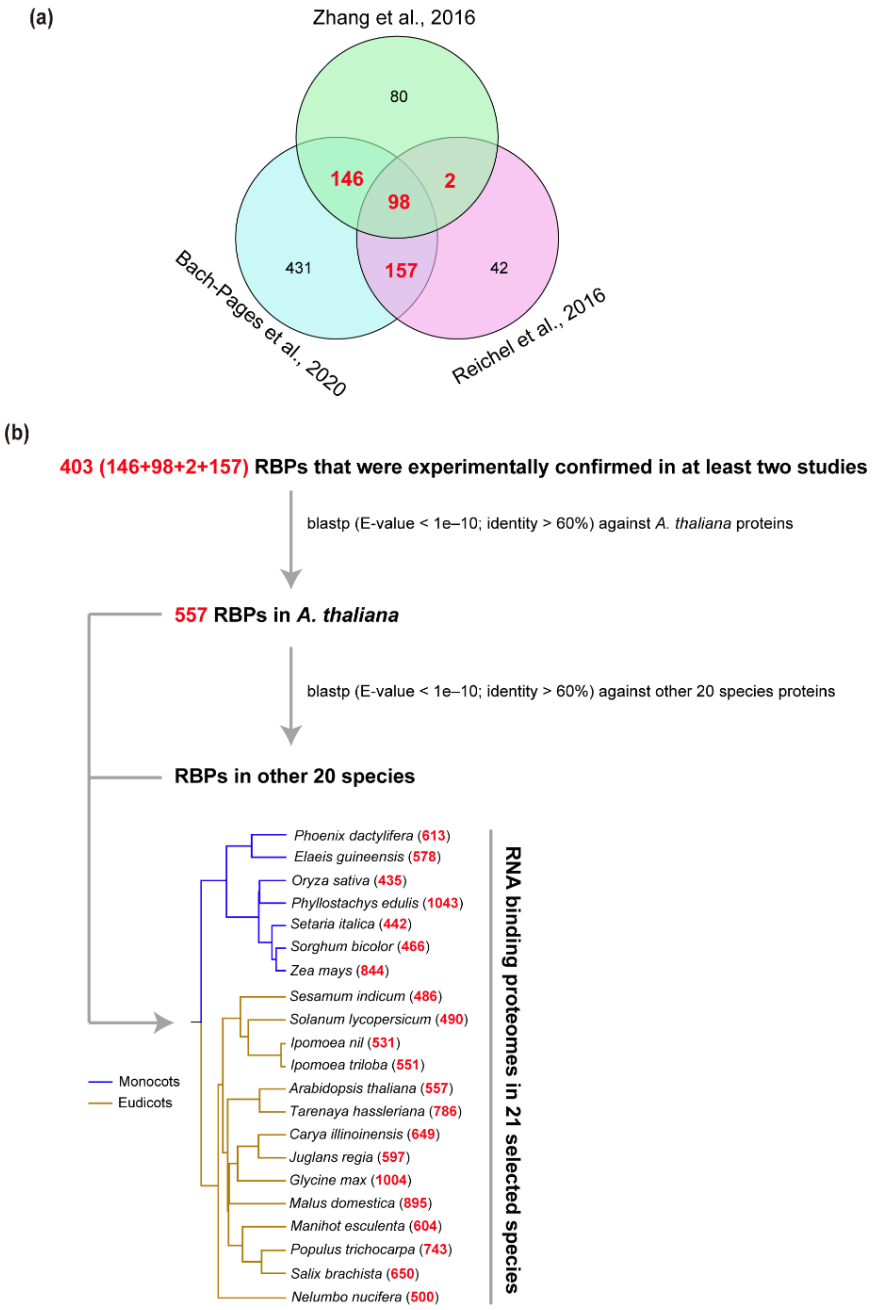

Fig. S1 Identification of RBPs in 21 selected angiosperms. (a) A Venn diagram shows overlapping numbers of RBPs that were experimentally confirmed to be RNA binding in three previous studies (Reichel *et al.*, 2016; Zhang *et al.*, 2016; Bach-Pages *et al.*, 2020). The RBPs identified in at least two studies are highlighted in red and considered to be RNA binding with high confidence (Table S2). (b) A workflow used for the identification of RBPs in 21 selected angiosperms. The number in parentheses after the species indicates the identified number of RBP genes in the corresponding species (Table S3).

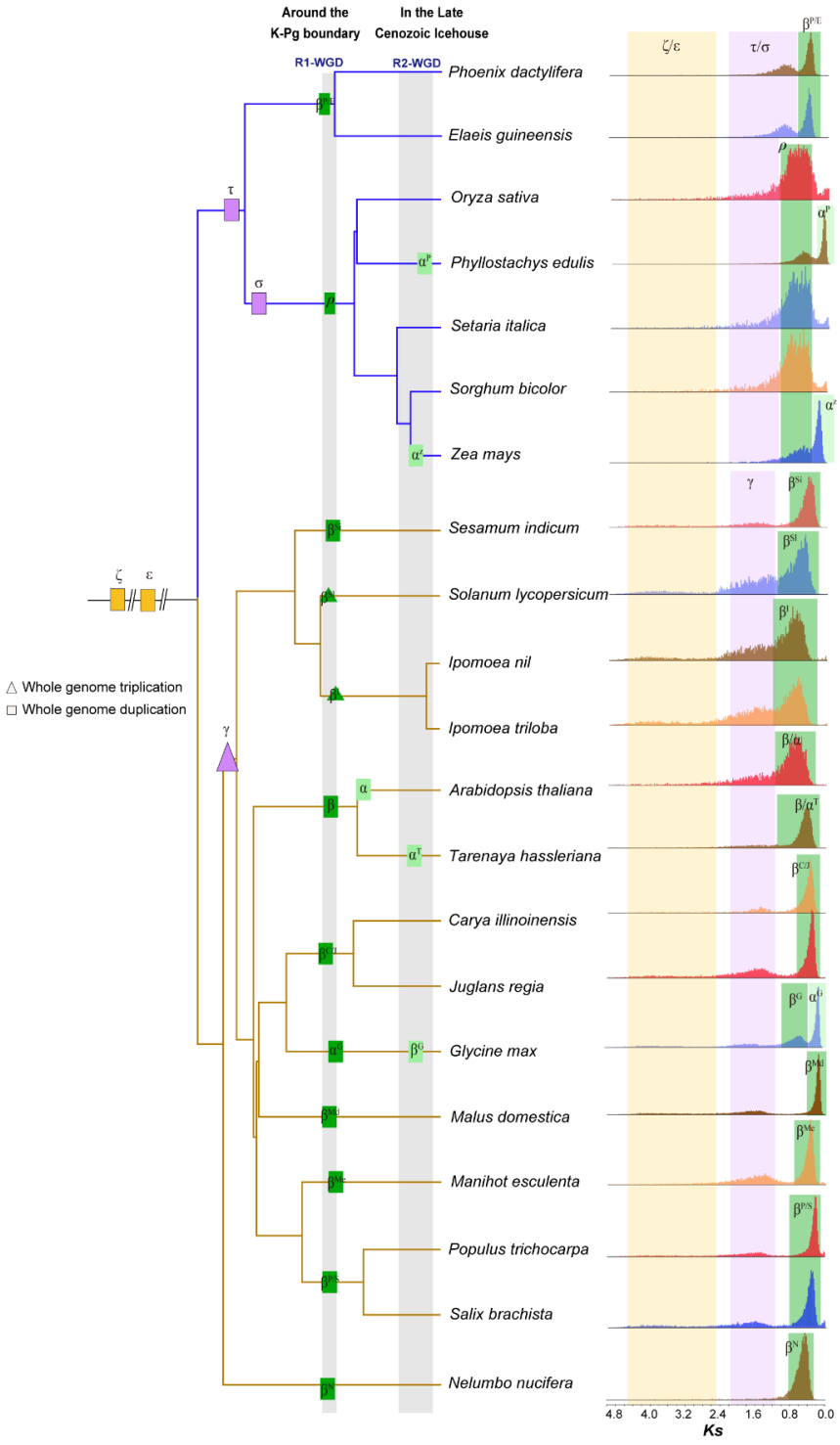

Fig. S2 Synonymous substitution (*Ks*) ranges associated with the well-documented WGD events in the 21 selected angiosperm species. The well-documented WGD events were positioned onto the branches of the phylogeny. Because of synonymous substitution saturation and low gene retention rates from ε/ζ-WGD-derived duplicates, no peaks were detected for ε/ζ-WGD events. We arbitrarily considered collinear gene duplicates with *Ks* > 2.2 as ε/ζ-WGD-derived duplicates. We listed the *Ks* range for each WGD event in Table S6. It is noteworthy that

different periods of WGDs might have similar *Ks* ranges, which was probable resulted from differences in generation times between species.

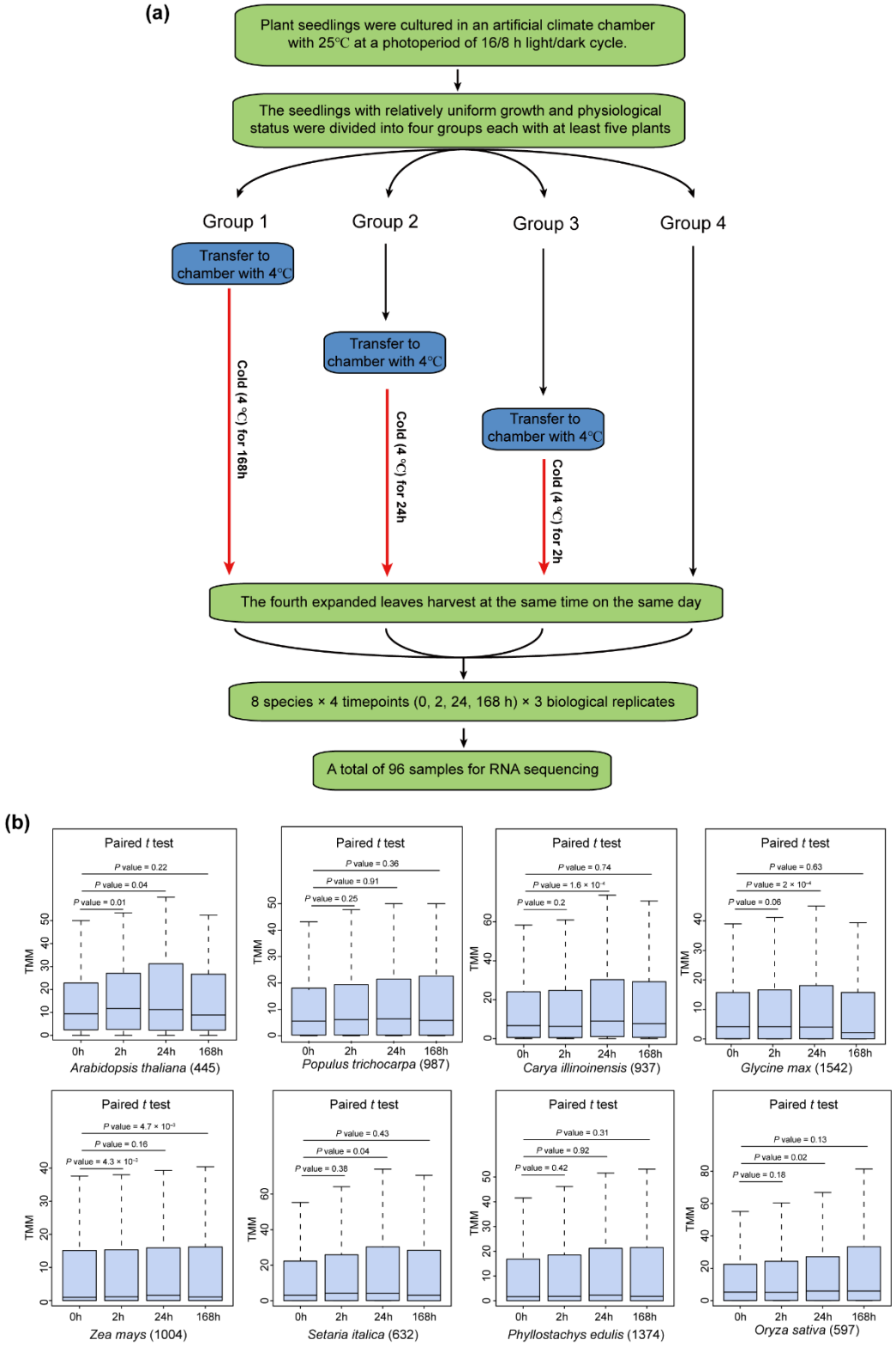

Fig. S3 Overview strategy of cold-responsive RNA-seq experiments and expression profiles of leaf development-related genes before and after cold stress. (a) The overview strategy of cold

treatments on eight selected species in this study. (b) Leaf development-related genes show similar expression in the collected leaves before and after cold stress. The number in parentheses beside each species indicates the number of leaf development-related genes in the species. Based on gene ontology annotation (GO:0048366), we obtained 445 leaf development-related genes from *A. thaliana*. Using the 445 genes, we employed BLASTP to retrieve their homologs in seven other cold-treated species (E-value < 1e–10, identity > 60%) (Altschul *et al.*, 1990). Gene expression was obtained from our RNA-seq analysis of the species at four timepoints (0, 2, 24, and 168 h) of cold stress (Table S15). The expression difference between two timepoints of cold treatments was analyzed by the two-sided paired *t*-test to show roughly similar development stages of the four timepoints of cold-treated samples in each species. Boxplots depict the median, interquartile range (IQR), and 1.5 × IQR (with outliers).

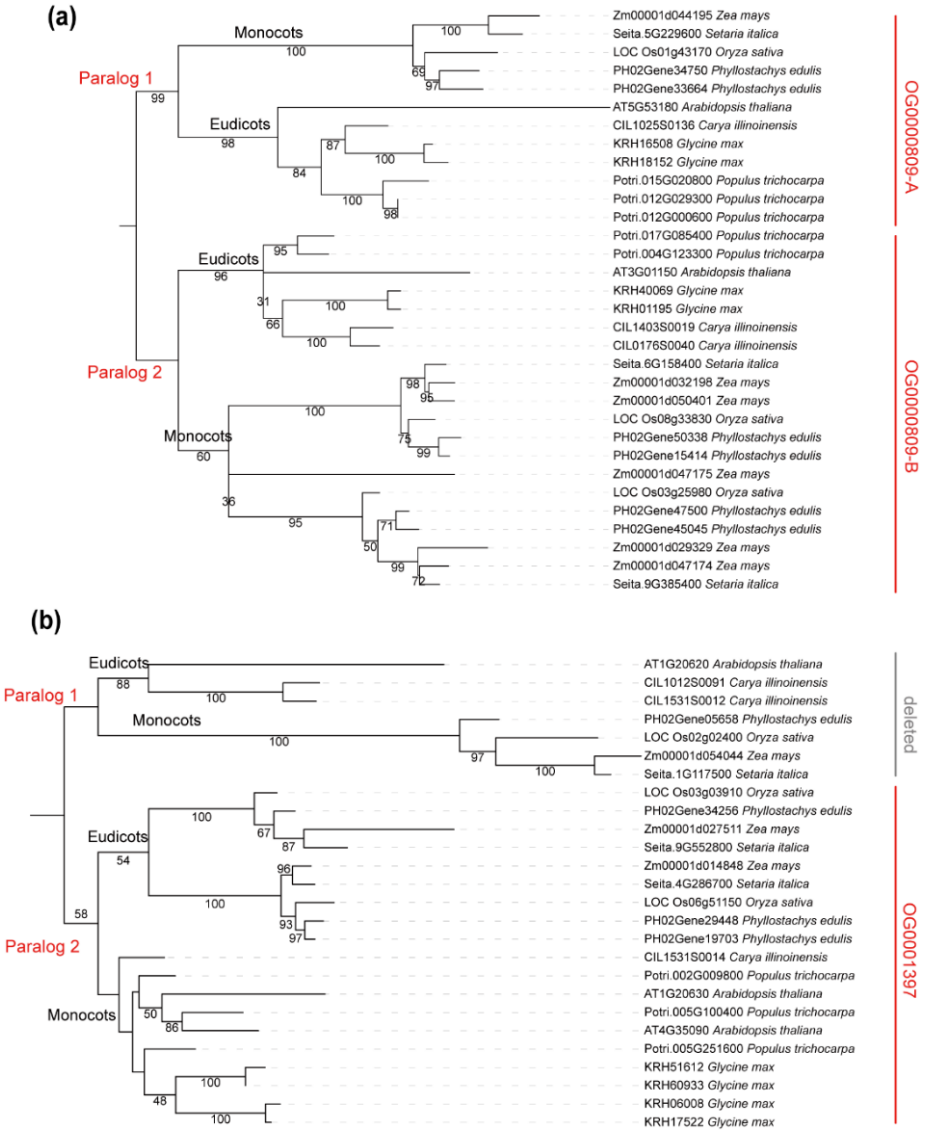

Fig. S4 Two examples for redefining orthogroups. (a) The orthogroup OG0000809 consists of two paralogs, Paralog 1 and Paralog 2, which are generated before the divergence of monocots

and eudicots. Paralog 1 and Paralog 2 are both conserved across eight cold-treated plants. Therefore, this orthogroup was divided into OG0000809-A and OG0000809-B. (b) The orthogroup OG0001397 consists of two paralogs, Paralog 1 and Paralog 2, which also originated before the monocots-eudicots divergence. However, Paralog 1 is lost in *Glcine max* and *Zea mays*, and for consistency, it was excluded from our study. In contrast, Paralog 2 are conserved in eight cold-treated species, and it was retained and still referred to OG0001397.

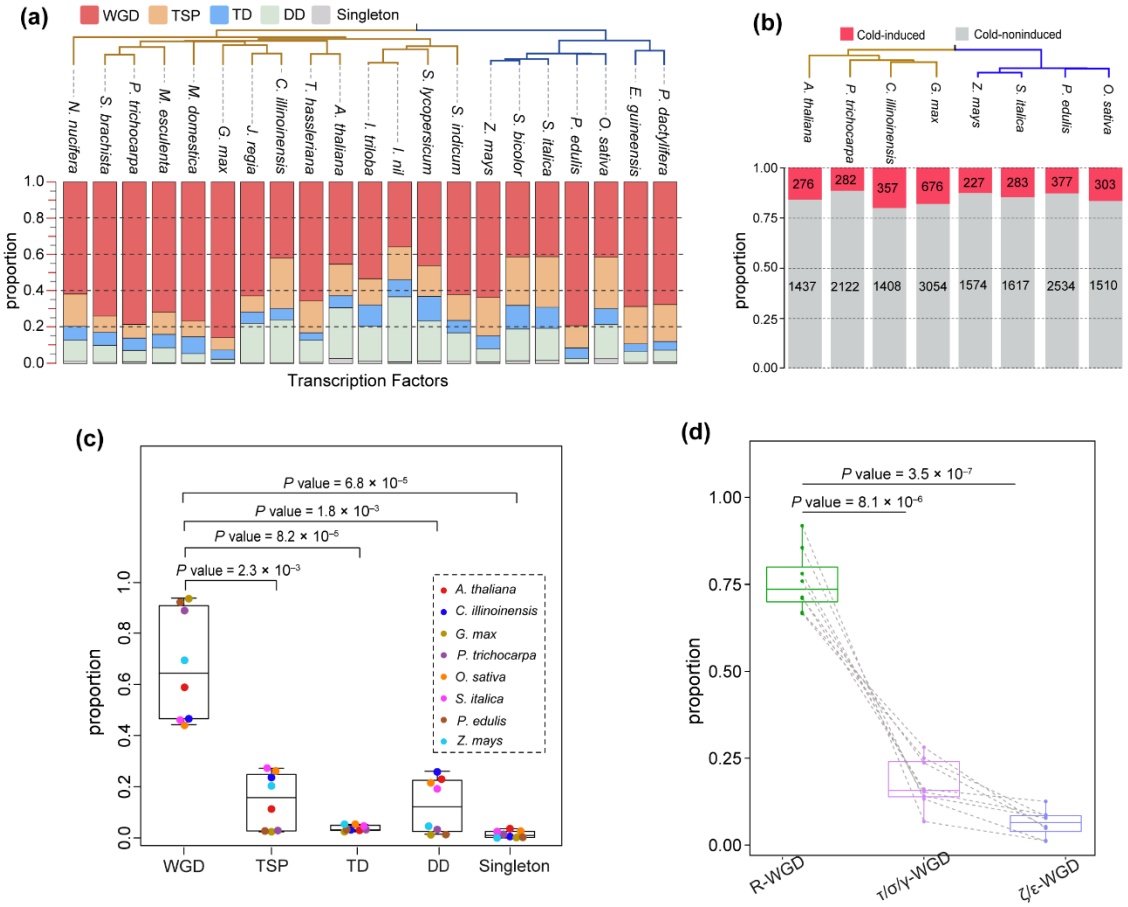

Fig. S5 Cold-induced TF duplicates were biased retained after R-WGD. (a) Proportions of TF genes produced by five duplication modes. We obtained TFs in *A. thaliana*, *P. trichocarpa*, *G. max*, *S. italica*, and *O. sativa* from plantTFDB 5.0 (Jin *et al.*, 2017) and identified TFs in 16 other species according to their respective TF domains. The programs MCScanX (Wang *et al.*, 2012) and DupGen_finder (Qiao *et al.*, 2019) were integrated to infer duplication modes for TF genes in the 21 selected angiosperms. WGD: whole genome duplication; TSP: transposed duplication; TD: tandem duplication; DD: dispersed duplication; and Singleton: single copy gene. (b) Identification of cold-induced TFs in the eight cold-treated species. Cold-induced genes were defined based on a false discovery rate (FDR) < 0.01 and fold change > 1.5 under at least one cold stress (2, 24, or 168 h) with an expression value greater than 20 at least in one condition. (c) Proportion differences of five duplication modes in producing cold-induced TFs. Proportion differences were assessed by the two-sided paired *t*-test between WGD- and other duplication mode-derived cold-induced TFs. (d) Proportion differences of different periods of WGDs in producing cold-induced TF duplicates. Proportion differences were assessed by the two-sided

paired *t*-test between different periods of WGDs in producing cold-induced TF duplicates. The same species is connected by dashed lines between two periods. Boxplots depict the median, interquartile range (IQR), and 1.5 × IQR (with outliers).

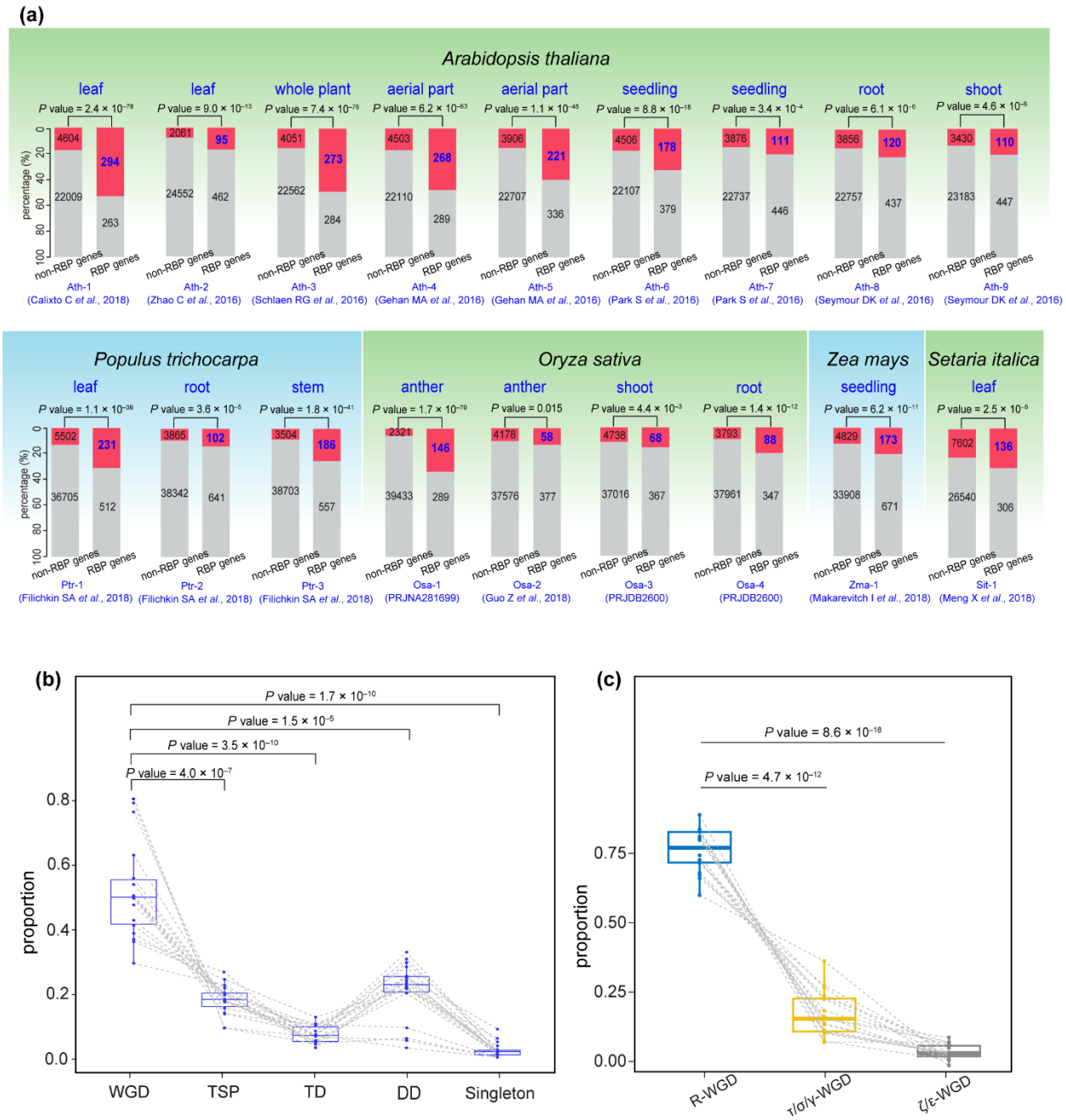

Fig. S6 Diverse cold-responsive transcriptome datasets consistently demonstrate the cold-upregulation of R-WGD-derived RBP genes. (a) A significant proportion of RBP genes are cold-induced in collected cold transcriptome. A total of eighteen sets of cold transcriptomes were collected from previous studies (Table S18), including nine sets for *Arabidopsis thaliana* (Ath-1 to Ath-9), three sets for *Populus trichocarpa* (Ptr-1 to Ptr-3), four sets for *Oryza sativa* (Osa-1 to Osa-4), one set for *Zea mays* (Zma-1) and one set for *Setaria italica* (Sit-1). The significant

difference was assessed by Fisher's exact test on RBP genes and non-RBP genes with and without cold induction. Cold-induced genes were defined based on the same strategy as our transcriptome analysis. The tissues used in each RNA-Seq experiment are labeled above the panel. (b) Proportion differences of five duplication models in producing cold-induced RBP genes that identified in collected cold transcriptomes. The proportion difference was assessed by the two-sided paired *t*-test on cold-induced RBP genes derived from WGD and other duplication models. (c) Proportion differences of different periods of WGDs in producing cold-induced RBP duplicates that identified in collected cold transcriptomes. The difference was assessed by the two-sided paired *t*-test on cold-induced RBP genes derived from different periods of WGDs. Boxplots depict the median, interquartile range (IQR), and 1.5 × IQR (with outliers).

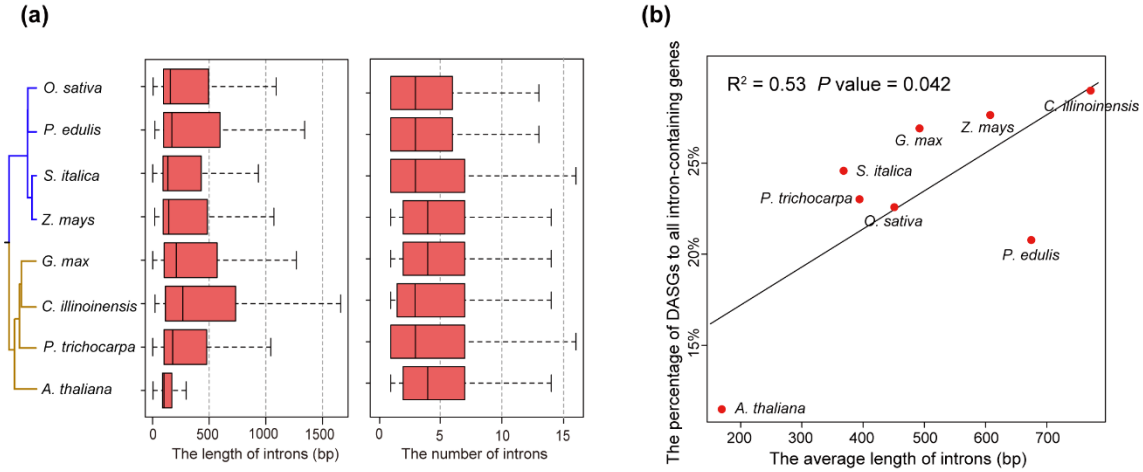

Fig. S7 Alternative splicing seems to be enriched in genes with longer introns. (a) Boxplot diagrams show intron length and number in eight cold-treated species genomes. The number of introns in *A. thaliana* is comparable to that in seven other cold-treated species, but their length in *A. thaliana* is significantly shorter. Boxplots depict the median, interquartile range (IQR), and 1.5 × IQR (with outliers). **(b)** A significant positive Pearson's correlation coefficient of genes being alternatively spliced and intron length. Pearson's correlation coefficient was used to test the correlation between the average length of introns and the percentage of differentially alternatively spliced genes (DASGs) to all intron-containing genes in eight cold-treated species.

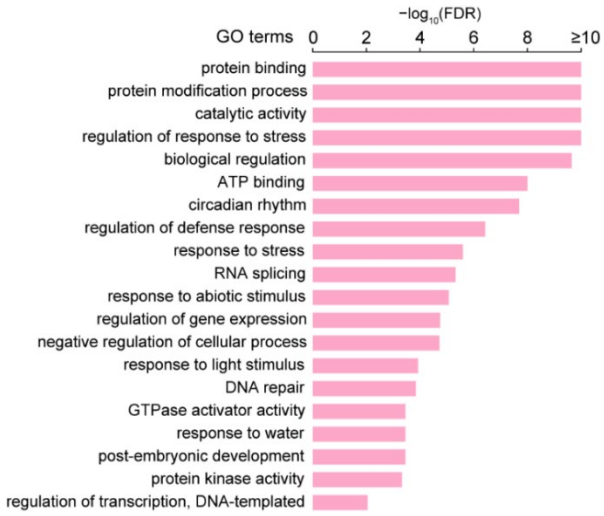

Fig. S8 Functional enrichments of cold-induced differentially alternatively spliced genes. Gene Ontology (GO) term enrichments of *Arabidopsis* DASGs. The top 20 significant GO terms were selected for the representatives of biological processes in which the DASGs are involved. The enrichment levels (*P* value) were corrected by Benjamini–Hochberg false discovery rate (FDR), and scaled as enrichment score −log10 (FDR).

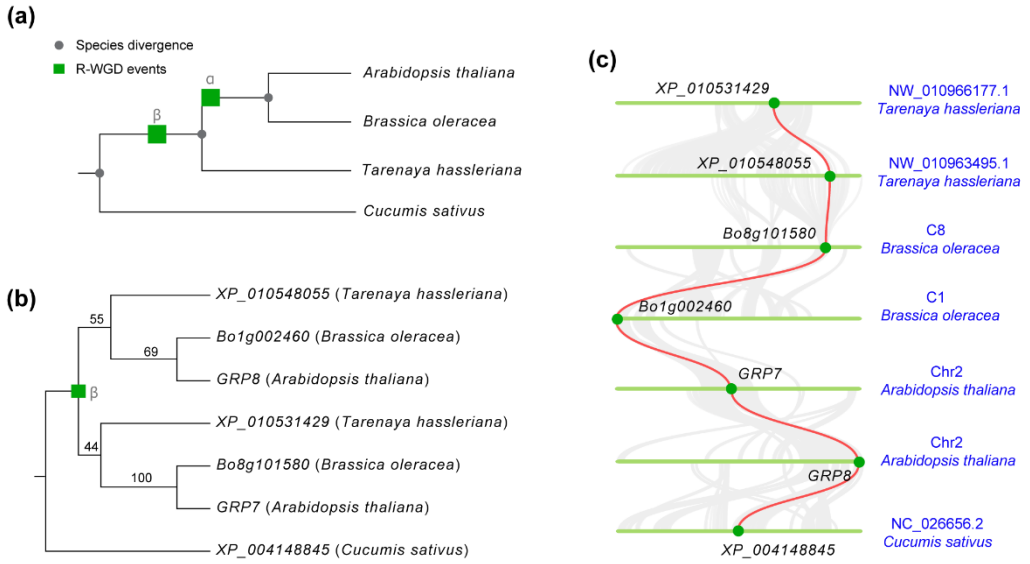

Fig. S9 The GRP7-GRP8 duplicates resulted from β-WGD event (R1-WGD). (a) The occurrence of β-WGD event. The β-WGD event occurred after the divergence of *Cucumis sativus* and before the divergence of *Brassicales* (*Brassica oleracea*, *Tarenaya hassleriana*, and *Arabidopsis thaliana*), and the α-WGD event occurred after the divergence of *T. hassleriana* and before *Brassicaceae* divergence. (b) The evolutionary relationship of *GRP7*, *GRP8* and their orthologous genes in *B. oleracea*, *T. hassleriana* and *C. sativus*. The number on branches indicates the bootstrap value of 100%. (c) Collinearity analysis of *GRP7/8*-containing paired genes in the species. In addition to collinearity analysis, *Ks* analysis supports the production of *GRP7-GRP8* duplicate from β-WGD event (Table S20).

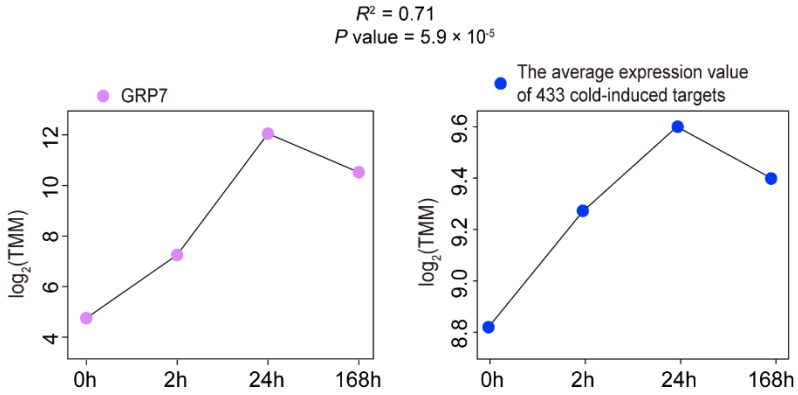

Fig. S10 *GRP7* expression was significantly correlated with the average expression of 433 cold-induced GRP7 targets. Pearson's correlation coefficient was used to test the potential of GRP7 in regulating its downstream cold-induced genes.

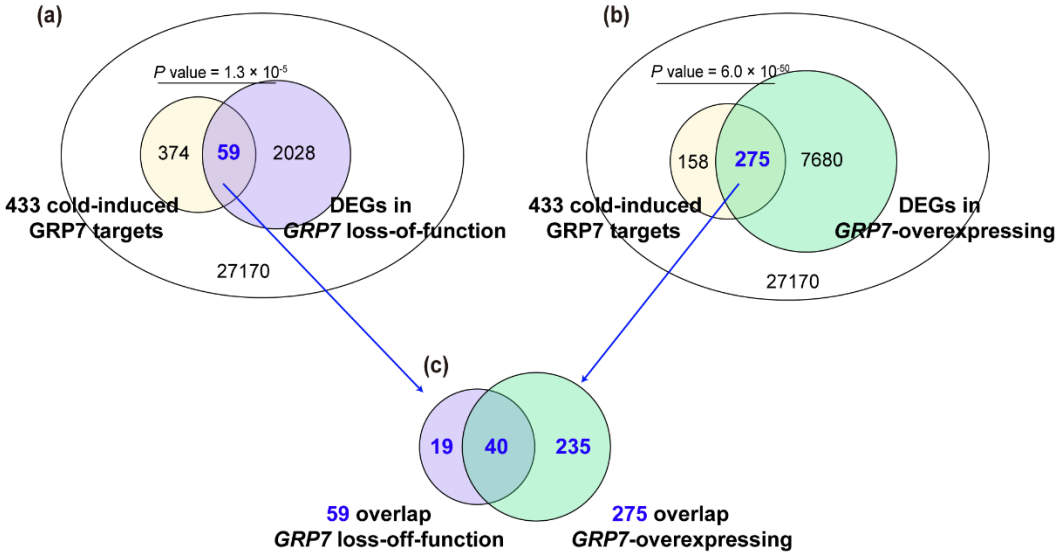

Fig. S11 A significant association of 433 cold-induced GRP7 targets with differentially expressed genes in *GRP7*-overexpressing and loss-of-function plants. Among 433 cold-induced GRP7 targets, many were differentially expressed in *GRP7* loss-of-function plants (a), *GRP7*-overexpressing plants (b), and both plants (c). We obtained a total of 857 GRP7 binding transcripts from iCLIP-seq analysis in *Arabidopsis* (Meyer *et al.*, 2017) and among the transcripts, 433 were further identified to be cold-induced from our RNA-seq analysis on the species at four different timepoints (0, 2, 24, and 168 h) of cold stress (4°C). The DEGs in *GRP7* loss-of-function and overexpressing plants were obtained from (Meyer *et al.*, 2017). The overlapping significance was assessed by Fisher's exact test and the related genes were listed in Table S11.

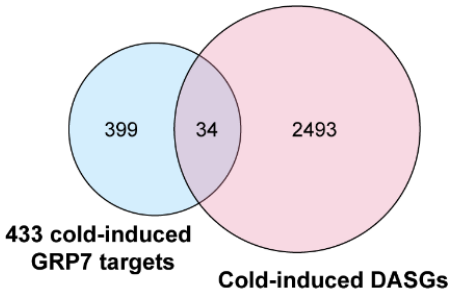

Fig. S12 Overlapping and unique genes between 433 cold-induced GRP7 targets and differentially alternatively spliced genes. A total of 433 cold-induced GRP7 targets were obtained as described in the legend of Fig. S11. The differentially alternatively spliced genes were obtained from our RNA-seq analysis of *Arabidopsis* seedling leaves at four different timepoints (0, 2, 24, and 168 h) of cold stress.

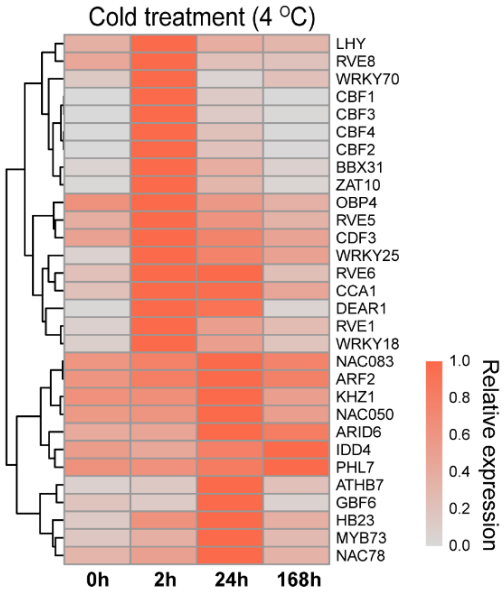

Fig. S13 Expression profile of the top 30 TFs that bind to promoters of cold-induced GRP7 targets. Gene expression was obtained from our RNA-seq analysis of *Arabidopsis* seedling leaves at four timepoints (0, 2, 24, and 168 h) of cold stress. Relative expression indicates the deviation of gene expression at each time point from the highest expression of the gene. This expression profile shows that most of the 30 TFs are induced at an early stage by cold stress (2 or 24 h).

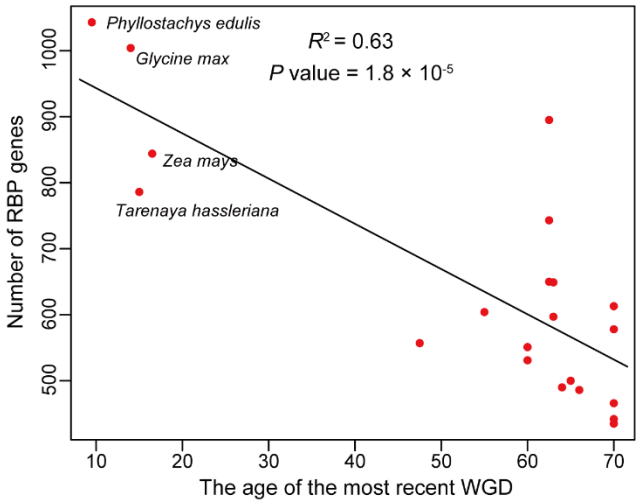

Fig. S14 Negative correlation between RBP gene number and the WGD age. This correlation can be explained by the higher likelihood of gene loss as evolutionary time advances. Four species (*Tarenaya hassleriana*, *Glycine max*, *Zea mays* and *Phyllostachys edulis* marked in the plot) that experienced the R2-WGD event during the Late Cenozoic Icehouse, have relatively higher number of RBPs. Pearson's correlation coefficient was used to test the correlation.

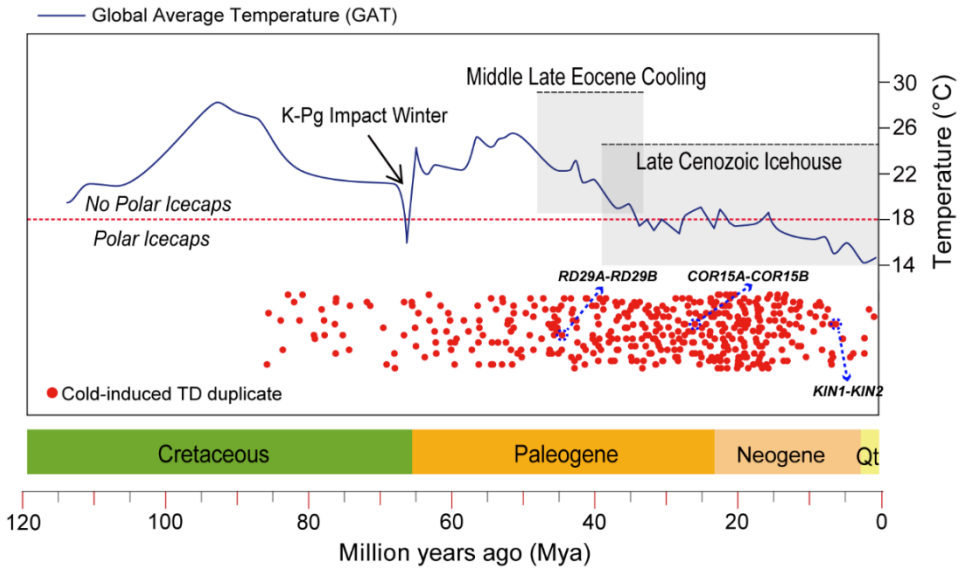

Fig. S15 Cold-induced tandem duplicates mainly generated around global cooling periods. The ages of cold-induced TD duplicates were roughly calculated using *T=Ks/2r*, where *Ks* represents the synonymous substitution values, and *r* represents the mutation rate of substitution per site per year in the lineage. The *r* values (7.0×10^{-9}) were obtained from previous studies (Ossowski *et al.*, 2010). The upper curve graph shows inferred paleotemperature fluctuations in the past 120 Mya (Scotese *et al.*, 2021), including a sharp decline of global average temperature (GAT) around the K-Pg boundary (~66 Mya) and a continuing GAT decline in the Middle Late Eocene Cooling (~48.5–34 Mya) and Late Cenozoic Icehouse (~39.4 Mya to present). When the GAT was below 18°C (red dashed line), large polar icecaps were formed.

Dataset S1 The programming code used in this study.

1. Identification of Pfam domain

```
hmmsearch --domtblout hmm_out --cut_tc Pfam-A.hmm protein.fa -E 1e-5
```

2. Identification of duplication modes

```
MCScanX target_species -k 50 -g -1 -s 5 -e 1e-5 -m 25 -w 5
duplicate_gene_classifier target_species -k 50 -g -1 -s 5 -e 1e-5 -m 25 -w 5
DupGen_finder.pl -i data_directory -t target_species -c outgroup_species -o output_directory -a 1 -d 10
```

3. Calculation of *Ks* value

```
ParaAT.pl -h WGD-derived-duplicates -n cds.fa -a protein.fa -m clustalw2 -p proc -f axt -o out-file
for i in `ls out-file/*.axt`;do KaKs_Calculator -i $i -o out-file/${i}.kaks -m NG;done
```

4. Fisher test in R project

```
# Inputfile is in the following format (tab separated):
# group         num1  num2  num3  num4
# a     100     200     500     3000
# b     200     300     600     4000
# c     300     400     700     5000
#       ...

data <- read.table("inputfile",header = T,row.names = 1)
row_number <- nrow(data )
for (x in 1:row_number) {
  test <- data [x,1:4]
  test <- as.numeric(test)
  test <- matrix(test,nrow=2)
  out <- fisher.test(test,alternative = "greater")
  data[x,5] <- out$p.value
}
names(data)[5]<-"pvalue"
data$fdr <- p.adjust(data$pvalue,method = "BH")
data$negative_log10fdr <- -log10(data$fdr)
write.table (data,"fisher_result.tab", row.names = T,col.names =T, quote =F,sep="\t")
```

5. Relative expression heatmap drawing through pheatmap package in R project.

```
# Inputfile is in the following format (tab separated):
# gene 0h       2h      24h     168h
# gene1         5       10      200     30
# gene2         0       0       6       1
```

```
# gene3          30       400      700      50
#       ...

exp <- read.table("Inputfile", header = T, row.names = 1)
exp[exp<1] = 1
exp$max <- apply(exp, 1, max)
normal_exp <- exp/exp$max
normal_exp <- subset(normal_exp, select = -max)
library(“pheatmap2”)
pheatmap (normal_exp, cluster_row = T, cluster_cols = F)
```

References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403-410.

Bach-Pages M, Homma F, Kourelis J, Kaschani F, Mohammed S, Kaiser M, van der Hoorn R, Castello A, Preston GM. 2020. Discovering the RNA-Binding Proteome of Plant Leaves with an Improved RNA Interactome Capture Method. *Biomolecules* **10**(4).

Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* **45**(D1): D1040-D1045.

Meyer K, Koster T, Nolte C, Weinholdt C, Lewinski M, Grosse I, Staiger D. 2017. Adaptation of iCLIP to plants determines the binding landscape of the clock-regulated RNA-binding protein AtGRP7. *Genome Biology* **18**(1): 204.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**(5961): 92-94.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology* **20**(1): 38.

Reichel M, Liao Y, Rettel M, Ragan C, Evers M, Alleaume AM, Horos R, Hentze MW, Preiss T, Millar AA. 2016. In planta determination of the mRNA-binding proteome of *Arabidopsis* etiolated seedlings. *The Plant Cell* **28**(10): 2435-2452.

Scotese CR, Song H, Mills BJW, van der Meer DG. 2021. Phanerozoic paleotemperatures: The earth's changing climate during the last 540 million years. *Earth-Science Reviews* **215**: 103503.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H *et al*. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**(7): e49.

Zhang Z, Boonen K, Ferrari P, Schoofs L, Janssens E, van Noort V, Rolland F, Geuten K. 2016. UV crosslinked mRNA-binding proteins captured from leaf mesophyll protoplasts. *Plant Methods* **12**: 42.