The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences

Shuaibin Lian (shuai_lian@xynu.edu.cn)
Xinyang Normal University https://orcid.org/0000-0002-0211-1842
Yongjie Zhou
Xinyang Normal University
Zixiao Liu
Xinyang Normal University
Andong Gong
Xinyang Normal University
Lin Cheng
Xinyang Normal University

Research article

Keywords: Paralogous gene pair, Differentially expressed gene, Whole genome duplication

Posted Date: April 21st, 2020

DOI: https://doi.org/10.21203/rs.2.16022/v2

Version of Record: A version of this preprint was published on June 16th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02460-x.

Abstract

Background Theoretically, paralogous genes generated through whole genome duplications should share identical expression levels due to their identical sequences and chromatin environments. However, functional divergences and expression differences have arisen due to selective pressures throughout evolution. A comprehensive investigation of the expression patterns of paralogous gene pairs in response to various stresses and a study of correlations between the expression levels and sequence divergences of the paralogs are needed.

Results In this study, we analyzed the expression patterns of paralogous genes under different types of stress and investigated the correlations between the expression levels and sequence divergences of the paralogs. We analyzed the differential expression patterns of the paralogs under four different types of stress (drought, cold, infection, and herbivory) and classified them into three main types according to their expression patterns. We then further analyzed the differential expression patterns under various degrees of stress and constructed corresponding co-expression networks of differentially expressed paralogs and transcription factors. Finally, we investigated the correlations between the expression levels and sequence divergences of the paralogs and identified positive correlations between expression level and sequence divergence. With regard to sequence divergence, we identified correlations between selective pressures and phylogenetic relationships.

Conclusions These results shed light on differential expression patterns of paralogs in response to environmental stresses and are helpful for understanding the relationships between expression levels and sequences divergences.

Background

Several studies have found that most plants have undergone multiple rounds of whole genome duplication (WGD) [1-3], which has long been recognized as an important evolutionary force. At least one ancient WGD occurred before the divergence of monocots and eudicots in angiosperm evolution. For example, *Arabidopsis thaliana* has undergone two recent WGD events, with the most recent one occurring at approximately 23 million years ago (Mya) [4]. Soybean (Glycine max) has also experienced two WGDs [5], which occurred at approximately 59 Mya and then 13 Mya. WGDs can duplicate entire chromosomes, thereby resulting in a large number of duplicate genes. These duplicate genes are considered to play important roles in enhancing organisms' adaptation to the environment and promoting species diversification [6-9]. The functions of the duplicate genes have diverged remarkably throughout evolution, although most duplicate genes have been lost [10,11].

Although many mechanisms can explain the functional divergences of the duplicate genes, the paralogous genes generated through WGDs should initially share identical sequences and chromatin environments and possess stronger expression correlations than would be found among other duplication types [12]. Theoretically, paralogs should share identical expression levels in the absence of

selective pressures and stress [13], because they share identical sequences. Functional divergences and expression differences have arisen due to selective pressures and harsh environments after hundreds of millions of years of evolution [14]. The divergences in the regulatory regions of genes may have changed their expression patterns, whereas changes in the coding regions may have resulted in the acquisition of new functions [15-17]. Therefore, gene expression divergence is an important evolutionary driving force for paralogs.

Several studies have examined the relationship between the sequence and expression divergence of duplicates [17-21]. Warnefors and Kaessmann investigated the correlations between the divergence of gene and protein expression in mammals and identified several positive correlations [22]. However, a study in sunflower has indicated that there are no correlations. This study instead described decoupling between gene expression and sequence divergence [23], with similar results reported in flycatcher species [24]. Furthermore, many studies have confirmed that genes with high expression levels evolve more slowly than those with low expression levels [25], and correlations between expression divergence and selective pressure have also been reported. For example, studies in *Drosophila* indicated that positive selection is closely related to expression divergence [26], whereas others have reported that purifying selection is the primary driving force of the divergences in expression and sequence [27]. Consequently, it is important to know whether there are correlations between expression divergences of paralogs that may have resulted from selective pressure in plants.

We investigated the differential expression patterns and expression divergences of paralogs under four different types of stress (two biotics stresses and two abiotics stresses) in *Arabidopsis thaliana*. Furthermore, we identified correlations between sequence divergences and selective pressures. Lastly, we constructed co-expression networks of paralogs with different expression patterns and associated transcription factors. A workflow chart showing the different steps presented in this study can be found in Figure **S1**.

Results

Homolog identification and paralog expression classification

We identified 6,481 paralogs (paralogous gene pairs) in the model plant species *Arabidopsis thaliana* based on a homology analysis which involved 20 other species using the InParanoid 8 Software (see the **Methods** section for details) [28]. The list of 6,481 paralogs is show in **Table S1**. The phylogenetic relationships of the 21 species were obtained from Lian et al. and Ren et al. [2, 29]. Thereafter, we analyzed the interactions and distributions of the paralogs and repeats in the chromosomes, respectively (**Fig. 1A**). The corresponding interaction information is presented in **Table S2** and **Table S3**. The repeats of *Arabidopsis thaliana* were identified using the RepeatMasker and HashRepeatFinder tools (described in the **Methods** section). These results indicated that the paralogs and repeats were highly coincident with regard to their locations and interactions, and the corresponding coincidence rate was 82.4%. This

which further confirmed that the paralogous gene pairs were mostly generated through genome duplications, including WGDs and small-scale duplications (SSDs) [2, 30].

Next, we classified the paralogs into three types (FF, FP or PP) (see definitions in **Methods**) according to their expression patterns under four different types of stress, including two biotic stresses (infection by the necrotrophic fungus *Botrytis cinerea*, *Bc*; and herbivory by the chewing larvae of *Pieris rapae*, *Pr*) and two abiotic stresses (drought [*Dr*] and cold [*Cd*]). We identified 382, 1510, and 4589 differentially expressed pairs of FF, FP, and PP paralogs in *Dr* stress; 402, 1611, and 4468 differentially expressed pairs of FF, FP, and PP paralogs in *Cd* stress; 649, 1710, and 4122 differentially expressed pairs of FF, FP, and PP paralogs in *Bc* stress; and 143, 723, and 5615 differentially expressed pairs of FF, FP, and PP paralogs in *Pr* stress, respectively (**Fig. 1B**). The list of FF, FP and PP paralogs under four different stresses is shown in **Table. S4**. The statistic significances of differences in expression of FF and FP genes under the four different types of stress were examined by using Mann-Whitney *U*-test (**Fig. 1B**). Differences were considered significant when their *P*-value was less than 0.05. The log2|FC| values of FF and FP paralogs under four different stresses are shown in **Table S5**. We also investigated co-expressed FF paralogs under the four different types of stress by computing the *Pearson* coefficient *r*. The proportions of the co-expressed FF paralogs were 77.4% in *Dr* stress, 84.3% in *Cd* stress, 79% in *Bc* stress and 93% in *Pr* stress (**Fig. 1B**). The threshold of *Pearson* coefficient was $r>0.5$.

These results showed that (1) most paralogous genes were not expressed or differentially expressed, and only a small proportion of the paralogous genes were both differentially expressed, which suggests that most paralogous genes are not involved in stress response mechanisms; and (2) the expression patterns of paralogs involved in stress response were significantly different, especially for FF and FP paralogs, which suggests that these differentially expressed paralogs (DEPs) are significantly differentially expressed in stress response; (3) most paralogs with FF expression patterns under four different types of environmental stress tend to show similar expression patterns.

Differential expression patterns of paralogs under biotic and abiotic stress

To investigate the differential expression patterns of FF and FP paralogs under the four different types of stress, we generated a Venny diagram of their overlaps(**Fig. 2** for FF, **Fig. S2** for FP). We first clustered all 1,052 FF paralogs and 2,703 FP paralogs into seven expression modules according to their differential expression patterns under different types of stress. The log2|FC| values of seven FF and FP expression clusters are shown in **Table S6** and **Table S7**, respectively. The corresponding heatmaps and the specific functions of the FF and FP paralogs are shownin **Figure 2** and **Figgure S2**, respectively. Furthermore, we identified the transcription factors (TFs) in each cluster. The FF and FP paralogs belonging to the first three clusters were differentially expressed during allfour different types of stress. The paralogs belonging to the last four clusters were differentially expressed during only one type of stress. We also performed function enrichment and KEGG analysis for the FF paralogs to assign functional categories to each module (**Fig. 2D**).

Cluster 1 contained four DEPs, two of which were TFs and were shared by all four types of stress (**Fig. 2A-C**). Functional enrichment analysis indicated that these four paralogs were mainly involved in galactose metabolism, and two of the TFs were *bHLH* transcription factors. These results indicate that plants require more energy to deal with harsh environments, which has been confirmed by a recent study [31]. Cluster 2 contained 90 differentially expressed paralogs, eighteen of which were TFs and were shared by abiotic stresses (*Dr* and *Cd*) (**Fig. 2A-C**). The functions of DEPs in cluster 2 were mainly involved in the response to various abiotic stresses, such as water deprivation, temperature fluctuations, and karrikin (**Fig. 2D**). Studies have confirmed that these genes are involved in the biosynthesis of abscisic acid, and they improved the abiotic stress tolerance in *Arabidopsis thaliana* when overexpressed[32, 33]. Karrikin, a signaling molecule, is found in smoke from burning vegetation, and it triggers seed germination for many angiosperms [34]. This may be a protective mechanism used by plants for seed development in response to harsh environmental conditions, such as drought, cold, and high salinity [35]. Cluster 3 contained 33 differentially expressed paralogs, six of which were TFs that were shared by biotic stresses (**Fig. 2A-C**). The corresponding functions were mainly involved in the response to various biotic stresses, such as protection from attacks by fungi, bacteria and oomycetes, as well as immunological processes. We identified five differentially expressed *WRKY* TFs (*WRKY6*, *WRKY40*, *WRKY54*, *WRKY70* and *WRKY18*), reflecting the important roles of *WRKY* TFs in the response to biotic stress. For example, *WRKY70* and *WRK54* are involved in basal defense mechanisms against *Hyaloperonospora parasitica* and disease resistance in Arabidopsis [36]. On the other hand, *WRKY6* and *WRKY40* play important roles in transducing *E-2*-hexenal perception, which is a green leaf volatile (GLV) that is produced upon wounding, herbivory or infection by pathogens [37].

With regard to clusters 4 through 7, we identified 179 (containing 24 TFs), 242 (containing 30 TFs), 456 (containing 74 TFs) and 56 paralogs (containing 21 TFs) that were differentially expressed under *Dr*, *Cd*, *Bc*, and *Pr* stress, respectively. The proportions of the co-expressed paralogs were 6.9%, 6.6%, 8.6% and 20.4% under *Dr*, *Cd*, *Bc* and *Pr* stress, respectively (**Fig. 2A-C**). The functional enrichment of cluster 4 indicated that the 179 paralogs were mainly enriched in carbohydrate biosynthesis, photosynthesis and drought recovery. Furthermore, *bHLH* negatively regulates jasmonate signaling and improves tolerance to drought stress [38]. The functions of cluster 5 were mainly enriched in the response to cold and ultraviolet light. As previously reported, these genes are involved in diurnal oscillation and beta-amylase biosynthesis, which increases the sensitivity of the PSII photochemical reaction to freezing and ambient stress in Arabidopsis [39, 40]. The functions of clusters 6 and 7 were mainly enriched in systemic resistance, toxin metabolism, immune response and protection from insects (**Fig. 2D**).

These results indicate that (1) paralogs with different expression clusters participate in different biological processes and have different biological functions; (2) the paralogous genes with functional redundancy were differentially expressed during the exposure to different types of stress, and (3) the expression patterns of the paralogous genes can change under different stress conditions.

Differential expression patterns of paralogs under different degrees of the same type of stress

We next investigated the effects of different degrees of stress on the expression patterns of the paralogs and classified the paralogs into two types according to expression level, which we defined as the enhancing expression pattern () and decreasing expression pattern () (**Fig. 3**). We identified 1521 and 10, 1773 and 8, 1985 and 13, and 364 and 26, enhancing and decreasing paralogs in *Dr*, *Cd*, *Bc*, and *Pr* stress, respectively (**Fig. 3A, B**). The log2|FC| values of paralogs with enhancing and decreasing patterns under four stresses are shown in **Table S8** and **Table S9**.

For the enhancing expression pattern, the paralogs were not expressed or differentially expressed at the onset of different stress. With prolonged or increased stress, more paralogs became differentially expressed (**Fig. 3 A, C**).At the strongest phase of *Dr*, *Cd*, *Bc* and *Pr* stress, the proportions of DEPs all reached 100%. The functional enrichment of the paralogs indicated that those responsive to the *Dr* stress were mainly involved in processes related to water deprivation and photosynthesis [41], those responsive to the *Cd* stress were mainly involved in processes related to temperature fluctuations and cold [42], those responsive to the *Bc* stress were mainly involved in processes related to protection from bacterial infection [43], and those responsive to the *Pr* stress were mainly involved in processes related to the defense response and immunological events [44]. Furthermore, we found that some enhancing paralogs were differentially expressed in at least two different types of stress simultaneously, and the proportions of the up-regulated paralogs in *Dr*, *Cd*, *Bc* and *Pr* co-enhanced with another type of stress were 22.1%, 22.6%, 14.5%, and 20.1%, respectively (**Fig. 3C**). These results indicate that most paralogs can respond to or be activated by several types of stress. Functional enrichment analysis of the 255 paralogs that responded to both *Dr* and *Cd* stress confirmed the functional redundancy with regard to water deprivation and temperature fluctuations. The functions of the 11 paralogs (**Fig. 3A**) shared by the four types of stress were mainly enriched in ion homeostasis and auxin transport [45], which have been reported to be involved in a wide array of stress responses [46, 47].

For the decreasing expression pattern, the paralogs were significantly differentially expressed at the onset of different types of stress. With prolonged stress, more paralogs were not expressed or differentially expressed (**Fig. 3B, D**). The functional enrichment of the paralogs indicated that those responsive to *Dr* and *Cd* stress were mainly involved in processes related to monocarboxylic acid and carboxylic acid biosynthesis. Recent studies have reported that these small molecules can help plants to adapt to extreme stress conditions [48, 49].

These results indicate that the expression patterns of the paralogs vary under different types of stress as well as with different degrees of stress, suggesting that the expression levels of paralogs are not only related to the type but also the severity of stress. These results also reveal that most paralogs are differentially expressed in response to multiple stresses, suggesting that the functional redundancy of paralogs is a protective mechanism for the adaptation of plants to different stress environments throughout evolution.

Co-expression networks of DE paralogs and transcriptional factors under different types of stress

To understand how transcription factors (TFs) regulate the expressed of DEPs in response to stress, we constructed co-expression networks for *Dr*, *Cd*, *Bc* and *Pr* stresses (**Fig. 4**).

The co-expression networks revealed several important insights. Firstly, among the enhancing and decreasing expression patterns of down-regulated DEPs under *Dr*, *Cd*, *Bc* and *Pr* stress, DEPs with both enhancing and decreasing patterns showed low expression, except for DEPs with a decreasing pattern under *Dr* stress. Secondly, the top three TFs co-expressed with DEPs were *MYB*, *ERF* and *bHLH* under *Dr* stress (**Fig. 4A**); *ERF*, *bHLH* and *NAC* under *Cd* stress (**Fig. 4B**); *ERF*, *MYB* and *WRKY* under *Bc* stress (**Fig. 4C**); and *ERF*, *NAC* and *MYB* under *Pr* stress (**Fig. 4D**). Previous studies have reported that *ERF* plays important roles in responses to both biotic and abiotic stresses [50-52]. For example, *ERF9* protects Arabidopsis from necrotrophic fungi, and post-anaerobic reoxygenation—the main defense mechanism in plants [53]—is regulated by *ERF96* [54]. A study has also confirmed that *bHLH* can mediate the trade-off between abiotic and biotic molecular pattern-triggered immunity in Arabidopsis [55, 56]. However, *MYB* is mainly involved in response to biotic stress [57, 58]. Thirdly, we identified specific TFs under different types of stress. For example, *NIN-LIKE* is a master regulator of the response of Arabidopsis to *Dr* stress [59]. *E2FD/DEL2* controls cell proliferation in Arabidopsis during exposure to *Cd* stress [60]. *BES1* promotes brassinosteroid signaling and development in *Arabidopsis thaliana* during exposure to *Bc* stress [61]. Finally, there were more interactions between DEPs and TFs with an enhancing expression pattern than those with a decreasing expression pattern (**Fig. 4**). The increased number of interactions indicated that more TFs regulated the responses of the paralogs to the enhancing severity of stress. These results are very helpful for understanding the regulatory mechanisms of TFs with regard to the responses of paralogs to stress.

Expression divergences positively correlate with sequence divergences

We continued our study by investigating whether there were positive or negative correlations between expression divergences and sequence divergences [62]. First, the paralogs with FF and FP expression patterns were investigated. To estimate the sequence divergence between paralogs, we computed the synonymous (*Ks*) substitution rate, which is recognized as a proxy of the sequence divergence time. According to previous studies [21, 62], we used the rescaled Pearson's correlation coefficient to perform linear regression analysis (see the **Methods** section for details). The regression results of the expression levels of FF and FP paralogs and the *Ks* rates are shown in **Figure 5**.

We found a significant negative correlation between the rescaled and *Ks* values for FF and FP gene pairs ($P < 0.001$, *U*-test, **Fig. 5A**). The negative correlation between the and *Ks* values was indicative of a positive correlation between expression divergence and sequence divergence. These results indicate that the expression divergences of both FF and FP gene pairs were positively correlated with sequence divergences. Furthermore, we investigated the distribution of *Ks* values for FF and FP paralogs and identified one peak with a value of 1.8 in the density plot (**Fig. 5B**). These results indicate that the gene pairs originating at a value of 1.8 experienced a large amount of synonymous substitution. More than 80% of FF and FP paralogs had *Ks* values larger than 1.0, suggesting that they have persisted for a

relatively long evolutionary duration time and are highly divergent. In addition, the gene pairs near the *Ks* peak probably experienced larger expression divergences [63].

We also investigated the correlations of DEPs with enhancing and decreasing expression patterns under *Dr, Cd, Bc* and *Pr* stress. We identified a negative correlation between the expression divergences and *Ks* value for all four types of stress ($P < 0.001$, *U*-test, **Fig. 5C**). These results indicate that the expression divergences of DEPs in response to stress were positively correlated with sequence divergences. Furthermore, a density plot of the corresponding *Ka* and *Ks* values had a *Ks* peak value of 1.8 (**Fig. 5D**), indicating that these genes have persisted for a relatively long evolutionary duration and are highly divergent.

In summary, this study reveals new correlations between the expression divergences and sequence divergences of paralogous genes, which adds to the current understanding of the evolutionary mechanisms behind stress adaptation in plants.

Selective pressures are correlated with the expression divergences of paralogs

We next investigated whether there were correlations between expression divergences and selective pressures of the paralogous genes. To infer selective pressures, we used FF and FP DEPs under *Dr*, *Cd*, *Bc* and *Pr* stress to compute their non-synonymous/synonymous substitutions rate ratios (*Ka*/*Ks*). The boxplot of *Ka* and *Ks* values, as well as the *Ka/Ks* ratios, of FF and FP DEPs under the four types of stress is shown in **Figure 6**.

These results revealed two important insights. First, the median value of the *Ka/Ks* ratio for FP was consistently larger than 1.0, but that of FF was smaller than 1.0 for all four types of stress, indicating that the FP gene pairs underwent positive selection but the FF gene pairs underwent purifying/negative selection. Secondly, the *Ka* and *Ks* values of FP for all four types of stress were consistently larger than those of FF, revealing that the FP gene pairs experienced more non-synonymous/synonymous substitutions and were evolutionarily older than the FF gene pairs. To ensure that the phenomena we observed were not due to chance, we compared our results with a randomized experiment containing an equal number of randomized gene pairs (**Fig. S3**, **Methods**), and found that the *Ka/Ks* ratio of FF was consistently smaller than 1.0 and that of the randomized experiment [29], but the *Ka/Ks* ratio of FP was consistently larger than 1.0 and that of the randomized experiment (*P*<). Statistical significance was determined by 10,000 randomized comparisons.

These results indicate that FF paralogous pairs experienced relaxed selection constraints and retained functional redundancy, but FP paralogous pairs experienced strong positive selection and more sequence divergence, which led to functional divergence. These findings suggest that paralogs with different expression patterns likely experienced different selection constraints.

Discussion

Sequence divergences of the paralogs support the phylogenetic relationships among species

To investigate the correlations between sequence divergences and phylogenetic relationships, we examined the synonymous substitution rate (*Ks*) of paralogs between *Arabidopsis thaliana* and 20 other species (**Fig. 7A**). The corresponding boxplot of *Ks* values is shown in **Figure 7B**. Generally, smaller *Ks* values indicated less synonymous substitutions and divergences as well as stronger phylogenetic relationships. The results in **Figure 7** show that three species, *Arabidopsis lyrata*, *Boechera stricta*, and *Brassica rapa*, had much smaller *Ks* values (0.3707, 0.878, and 0.905, respectively) for *Arabidopsis thaliana*, as compared with 17 other species (all larger than 1.0). This indicates that the genomes of these three species display less divergence and closer phylogenetic relationships with *Arabidopsis thaliana*, which is consistent with the phylogenetic results of angiosperms [64]. Furthermore, we identified an inversely proportional correlation between species conservation and family size (**Fig. S4**). The family size of the paralogs significantly decreased as the occurrence of the species increased. A recent study has proposed a model of exponential decrease of duplicate genes over time [2]. Further studies are needed to investigate whether the relationship between species conservation and family size of the paralogs fits the exponential decay model, as these results may improve our understanding of the evolution of the duplicate genes.

Conserved domains and cis-elements

A recent report has confirmed that the expression divergence of the duplicated genes is primarily attributed to alterations in cis-elements [65], which have been proposed to mediate the expression divergence of genes in rice [66]. To further assess the impacts of cis-elements on expression divergence, we investigated the conserved domains and cis-elements of the paralogs in all 21 species.

We identified one paralogous gene family with seven genes in all 21 species and used the CDD Database to identify their conserved domains. The most highly conserved protein domains were the catalytic domain of the serine/threonine kinases (STKs), interleukin-1 receptor associated kinases and related STKs (STKc-IRAK) (**Fig. 8**). The STKs catalyze the transfer of the gamma-phosphoryl group from ATP to serine/threonine residues on the protein substrates. IRAKs are involved in the Toll-like receptor (TLR) and interleukin-1 (IL-1) signaling pathways. Thus, they regulate innate immune responses and inflammation [67, 68]. Using the MEME software, we identified 15 conserved motifs of STKc-IRAK, and found that most motifs were widespread in TFs, such as *LBD*, *ARF*, *SAP*, *Whirly*, *SRS*, *Dof* and *GRAS* (**Fig. 9A, B**). Furthermore, the seven genes in all 21 species shared similar motif structures and gene lengths.

We used PlantCARE to predict cis-element variations of the STKc-IRAK gene family and identified 13 cis-elements related to stress in the 2000-bp promoter sequence of the paralogous gene family (**Fig. 10**). The top ten components are shown in **Figure 10A**, and they include a low temperature response component (*LTR*), *MYB* binding site involved in the drought induction (*MBS*), MeJA reaction component (*CGTCA-motif*), salicylic acid reaction component (*TCA-element*), gibberellin reaction component (*GARE-motif* and *P-box*), auxin response element (*TGA-element*), abscisic acid reaction component (*ABRE*), MeJA element (*TGACG-motif*), stress response element (*TC-rich repeats*) and optical response elements (*3-AF1 binding*

site, GT1-motif, and *Sp1*). The number of cis-elements identified in each gene is shown in **Figure 10B**. Among them, the top two elements were the *CGTCA-motif* and *TGACG-motif*, accounting for 25% for all elements. These cis-elements are all related to stress, which suggests that they may be involved in the transcriptional control of abiotic stresses and hormonal responses [69].

Conclusions

In this study, we analyzed the expression patterns of paralogous genes under different types of stress and investigated the correlations between the divergences in expression and sequence of the paralogs. Firstly, we analyzed the differential expression patterns of the paralogs under four different stresses (*Dr, Cd, Bc* and *Pr*) and classified them into three types according to their expression patterns. Secondly, we analyzed their differential expression patterns under different degrees of stress and constructed corresponding co-expression networks of differentially expressed paralogs and TFs. Thirdly, we investigated the correlations between the divergences in expression and sequence and identified positive correlations between the expression divergences and sequence divergences. Lastly, we found that paralogs with different expression patterns likely experienced different selection constraints. FF paralogous pairs likely experienced relaxed selection constraints, while FP paralogous pairs experienced strong positive selection. These results suggest that paralogs which experienced relaxed selection tend to be functionally redundant while those which experienced strong positive selection tended to show more sequence divergence, Overall, these results provide new insightsinto the differential expression patterns of paralogs in response to environmental stresses and how those expression patterns relate to sequence divergences.

Methods

Homolog identification and paralog classification

We used the homolog analysis software InParanoid 8 with default parameters to identify paralogous gene pairs between *Arabidopsis thaliana* and 20 other species according to their phylogenetic relationships (**Fig. 7A**) [28]. The genomes and annotation files of Arabidopsis and the 20 other species were all downloaded from the EnsemblPlant (http://plants.ensembl.org) and UniProt (https://www.uniprot.org/) database. For detailed version information, For detailed version information, please refer to the attached **Table S10**. Among the 20 orthologs homology comparison results, multiple Arabidopsis genes corresponding to one ortholog gene were screened as candidate paralogs. In order to analyze the expression differences between a pair of genes, we selected the first two paralogs gene pairs with the highest similarity in each family as a preliminary identification and removed redundant duplicates. The originally identified homolog pairs were verified by BLAST alignment using the full-length amino acid. According to the e-value and similarity, homologous gene pairs with e-value< 10e-5 and similarity ≥50% were selected. The screening results were further verified with paralogs in the EnsemblPlant databases (Table. S1). After removing the identical gene pairs, 6,481 paralogous gene pairs (paralogs) remained. Thereafter, we classified each paralogous gene pair into one of three types (FF,

FP or PP) according to whether it was differentially expressed under different stress conditions. FF paralogs refer to paralogous gene pairs in which both genes in a pair were differentially expressed. FP paralogs refer to paralogous gene pairs in which one gene in a pair was differentially expressed and the other was not expressed or differentially expressed. PP paralogs refer to paralogous gene pairs in which both genes in a pair were not expressed or differentially expressed.

Transcriptome analysis

The transcriptome data of *Arabidopsis thaliana* under drought stress, cold stress, infection by the necrotrophic fungus *Botrytis cinerea,* and herbivory by the chewing larvae of *Pieris rapae* were obtained from the Chinese Academy of Sciences with Bio-Project Accession No. PRJNA525452 (https://www.ncbi.nlm.nih.gov/bioproject/525452) [70]. Three time points were selected for each stress condition, with a separate control for each. See **Table S11** for transcriptome data. At each time point, the transcriptional response to each single and sequential stress was compared with an untreated control or a mock-treated control. We first used Trimmomatic-0.36 software to remove the low-quality RNA-sequencing reads, and then used HISAT(Hierarchical Indexing for Spliced Alignment of Transcripts) 2-2.0.4 to map clean reads to reference genomes with default parameters for bam file generation. The expression levels of all mapped reads were normalized by FPKM (Fragments Per Kilobase of transcript per Million mapped reads) methods. Cufflinks (V2.2.0) software was then used to generate FPKM values for each gene. EdgeR was used to identify differentially expressed genes (DEGs) under four different types of stress with parameters padj<0.05 and |log2FC|>1 [71]. For determining the maximum dynamic range of stress response, the response to each of the four stresses was monitored in a different time frame of three time points, depending on how quickly the stress response developed . At each time point, the transcriptional response to each single and sequential stress was compared with an untreated control (for treatments not involving B. cinerea) or a mock-treated control (100% relative humidity conditions, as was uesd in *B. cinerea* treatments) for comparison. Control plants were sampled at the same time as stress-treated plants[70]. For differential expression pattern, we used transcriptome data at 7_d for *Dr* stress, 24_h for *Cd* stress, 18_h for *Bc* stress and 24_h for *Pr* stress. For enhancing and decreasing expression pattern analysis, we used transcriptome data at 5_d, 6_d and 7_d for *Dr* stress, 0_h, 3_h and 24_h for *Cd* stress, 6_h, 12_h and 18_h for *Bc* stress, and 6_h, 12_h and 24_h for *Pr* stress.

Interactions and distribution analysis

We used the RepeatMasker and HashRepeatFinder tools to identify repetitive sequences in *Arabidopsis thaliana*. The threshold of similar repetitive sequences was set to 85%, and repeats shorter than 150 nucleotides were removed. We determined the locations of the repeats and paralogs on the chromosomes using annotation data and used the R packages GlobalOptions and Circlize to identify interactions and distributions on the chromosomes.

Weighted gene co-expression network analysis

The weighted gene co-expression network analysis (WGCNA) package within R summarizes and standardizes the methods and functions for co-expression network analysis [72]. The WGCNA network construction tool was used to generate the nodes and edges of the genes by computing the correlations of the expression values. The nodes corresponded to genes, and the edges were determined by pairwise correlations between gene expression levels. The corresponding calling function within the R package was 'blockwiseModules'. The parameters were set as follows: powers = 10, minModuleSize = 30 and mergeCutHeight = 0.25. Other parameters were kept at their default settings. The nodes with a correlation of r<0.5 and edges with a weighted threshold of <0.3 were removed. Afterwards, the Cytoscape tool (https://cytoscape.org/) was used to plot the interactions using the nodes and edges of conserved genes.

Expression and sequence divergence analysis

The non-synonymous (*Ka*) and synonymous (*Ks*) substitutions of each paralog were computed using the '*dnds*' function within MATLAB. *Ka/Ks*>1 indicates that the gene experienced positive selection, *Ka/Ks*<1 indicates that the gene experienced negative selection, and *Ka/Ks*=1 indicates that the gene experienced selection[73]. The boxplots of *Ka* and *Ks* values were generated using the '*ggplot2*' function within R. The Pearson coefficient of the expression level of each paralogous gene pair was computed using the '*corr*' function within MATLAB using the following equation:

[Please see the supplementary files section to view the equations.]

Randomized experiments

We simulated randomized experiments to test the statistical significance of *Ka* and *Ks* for the FF and FP paralogs [29]. When the selective pressure was not characteristic of the FF or FP gene pairs, the results of the randomized experiment and real data were similar. To achieve this, we randomly generated an equal number of FF and FP gene pairs for each stress condition from 6,481 paralogs. We repeated the randomized experiment 10,000 times to evaluate the intrachromosomal colocalization of these random pairs. For example, to test the significance of the *Ks* value for 382 FF paralogs under *Dr* stress, we randomly generated 382 gene pairs from the 6,481 paralogs, and computed their *Ks* values, with 10,000 replications. The frequency distributions of the *Ka* and *Ks* rates, as well as the *Ka/Ks* ratio, with 0.1 steps are shown in **Fig. S3**.

Statistical methods

The Mann-Whitney *U*-test (function 'ranksum' in software'MATLAB' version R2016b) was used to examine the statistical significance between two samples, with a default significance level of 0.05. The Mann-Whitney U-test is a nonparametric test for equality of population medians of two independent samples. The main advantage of this test is that it makes no assumption that the samples are from normal distributions.

Cis-element and conserved domain analysis

The online platform PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html) was used for cis-element analysis utilizing the 2,000 bp promoter regions of the seven paralogs [74]. The Multiple Em for Motif Elicitation (MEME) software (http://meme-suite.org/tools/meme) was used for motif discovery. The motif number was 15, and the motif width was 50 amino acids. The Conserved Domain Database (CCD, https://www.ncbi.nlm.nih.gov/cdd/) was used to analyze the conserved domain sequences [75]. Functional enrichment was performed by using Metascape tools [76], and the resulting P values were adjusted to Q values by the Benjamini–Hochberg correction with a false discovery rate of 5%.

Availability of data and materials

The genetic data of the 21 species are listed in **Figure 7A**, including the CDS sequences and annotation data, which were downloaded from the EnsemblPlants(http://plants.ensembl.org/) and UniProt (https://www.uniprot.org/) database. In addition, 2,296 transcription factors (1,717 loci) of *Arabidopsis thaliana* were downloaded from the Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/index.php).

Abbreviations

WGD: whole genome duplication; SSDs: small-scale duplications; FF: FF paralogs refer to paralogous gene pairs in which both genes in a pair were differentially expressed; FP: FP paralogs refer to paralogous gene pairs in which one gene in a pair was differentially expressed and the other was not expressed or differentially expressed; PP: PP paralogs refer to paralogous gene pairs in which both genes in a pair were not expressed or differentially expressed; *Dr*: *drought*; *Cd*: *cold*; *Bc*: *botrytis cinerea*; *Pr*: *pieris rapae*; DEP: differentially expressed paralog; KEGG: kyoto encyclopedia of genes and genomes; TF: transcription factor; Ks: synonymous; Ka: non-synonymous; DEG: differentially expressed gene; FC: fold change; WGCNA: weighted gene co-expression network analysis.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that no competing interests exist.

Funding

This work was supported by the National Natural Science Foundation of China (Grant. 61501392, U1604112, 31701740), Nanhu Scholars Program for Young Scholars of XYNU. The National Natural Science Foundation of China played a role in the design of the study and collection, analysis, and interpretation of data. Nanhu Scholars Program for Young Scholars of XYNU provided the financial support for writing the manuscript.

Authors' contributions

SL and YZ implemented the algorithms and carried out the experiments. SL and LC drafted the manuscript. SL, YZ, ZZ and LC designed the study and analyzed the results. LC, and AG participated in the analysis and discussion. SL and TL contributed equally. All authors read and approved the final manuscript.

Acknowledgments

Authors thank anonymous reviewers for their comments on the manuscript. The linguistic editing and proofreading provided by TopEdit LLC during the preparation of this manuscript are acknowledged.

Authors' information

[1] College of Physics and Electronic Engineering, Xinyang Normal University, Xinyang, China.

[2] College of Life Sciences, Xinyang Normal University, Xinyang, China.

References

[1] Zhikai Liang, James C. Schnable. Functional Divergence between Subgenomes and Gene Pairs after Whole Genome Duplications. Molecular Plant,2018,11(3). DOI: 10.1016/j.molp.2017.12.010.

[2] Ren Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, Ji Qi. Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Molecular Plant,2018,11(3): 414-428. DOI: 10.1016/j.molp.2018.01.002.

[3] Christenhusz M J M, Byng J W. The number of known plants species in the world and its annual increase. Phytotaxa, 2016, 261(3):201. DOI: 10.11646/phytotaxa.261.3.1.

[4] Barker Michael S, Vogel Heiko, Schranz M Eric. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biology and Evolution,2009,1(1):391-399. DOI: 10.1093/gbe/evp040.

[5] Anne Roulin, Paul L. Auer, Marc Libault, Jessica Schlueter, Andrew Farmer, Greg May, Gary Stacey, Rebecca W. Doerge, Scott A. Jackson. The fate of duplicated genes in a polyploid plant genome. The Plant Journal,2013,73(1):143-153. DOI: 10.1111/tpj.12026.

[6] Matthew J. Hegarty, Simon J. Hiscock. Genomic Clues to the Evolutionary Success of Polyploid Plants. Current Biology,2008,18(10). DOI: 10.1016/j.cub.2008.03.043.

[7] Marie Sémon, Kenneth H. Wolfe. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends in Genetics,2007,23(3):108-112. DOI: 10.1016/j.tig.2007.01.003.

[8] Jiao Yuannian, Wickett Norman J, Ayyampalayam Saravanaraj, Chanderbali André S, Landherr Lena, Ralph Paula E, Tomsho Lynn P, Hu Yi, Liang Haiying, Soltis Pamela S, Soltis Douglas E, Clifton Sandra W, Schlarbaum Scott E, Schuster Stephan C, Ma Hong, Leebens-Mack Jim, dePamphilis Claude W. Ancestral polyploidy in seed plants and angiosperms. Nature,2011,473(7345):97-100. DOI: 10.1038/nature09916.

[9] Yang Zhenzhen, Wafula Eric K, Honaas Loren A, Zhang Huiting, Das Malay, Fernandez-Aparicio Monica, Huang Kan, Bandaranayake Pradeepa C G, Wu Biao, Der Joshua P, Clarke Christopher R, Ralph Paula E, Landherr Lena, Altman Naomi S, Timko Michael P, Yoder John I, Westwood James H, dePamphilis Claude W. Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. Molecular biology and evolution,2015,32(3):767-790. DOI: 10.1093/molbev/msu343.

[10] Ning Zhang, Liping Zeng, Hongyan Shan, Hong Ma. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytologist,2012,195(4):923-937. DOI: 10.1111/j.1469-8137.2012.04212.x.

[11] Lynch M, Conery J S. The evolutionary fate and consequences of duplicate genes. Science,2000,290(5494):1151-1155. DOI: 10.1126/science.290.5494.1151.

[12] Yupeng W, Xiyin W, Haibao T, Xu T, Ficklin S. P, Alex F. F. Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms. PLoS ONE, 2011, 6(12): e28150-. DOI: 10.1371/journal.pone.0028150.

[13] Gout Jean-Francois, Lynch Michael. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. Molecular biology and evolution,2015,32(8):2141-2148. DOI: 10.1093/molbev/msv095.

[14] Contrasted patterns of selective pressure in three recent paralogous gene pairs in theMedicagogenus (L.). BMC Evolutionary Biology, 2012, 12(1). DOI: 10.1186/1471-2148-12-195.

[15] Li W H, Yang J, Gu X. Expression divergence between duplicate genes. Trends in Genetics, 2005, 21(11):602-607. DOI: 10.1016/j.tig.2005.08.006.

[16] Innan Hideki, Kondrashov Fyodor. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews. Genetics,2010,11(2). DOI: 10.1038/nrg2689.

[17] Yupeng Wang, Xiyin Wang, Andrew H Paterson. Genome and gene duplications and gene expression divergence: a view from plants. Annals of the New York Academy of Sciences,2012,1256(1):1-14. DOI: 10.1111/j.1749-6632.2011.06384.x.

[18] Khan Nadeem, Hu Chun-Mei, Amjad Khan Waleed, Naseri Emal, Ke Han, Huijie Dong, Hou Xilin. Evolution and Expression Divergence of E2 Gene Family under Multiple Abiotic and Phytohormones Stresses in Brassica rapa. BioMed research international,2018,2018. DOI: 10.1155/2018/5206758.

[19] Hodgins Kathryn A, Yeaman Sam, Nurkowski Kristin A, Rieseberg Loren H, Aitken Sally N. Expression Divergence Is Correlated with Sequence Evolution but Not Positive Selection in Conifers. Molecular biology and evolution,2016,33(6):1502-1516. DOI: 10.1093/molbev/msw032.

[20] Echave Julian, Wilke Claus O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. Annual review of biophysics,2017,46. DOI: 10.1146/annurev-biophys-070816-033819.

[21] Dahai Gao, Dennis C. Ko, Xinmin Tian, Guang Yang, Liuyang Wang. Expression Divergence of Duplicate Genes in the Protein Kinase Superfamily in Pacific Oyster. Evolutionary Bioinformatics,2015,2015(Suppl. 1):57-65. DOI: 10.4137/EBO.S30230.

[22] Warnefors Maria, Kaessmann Henrik. Evolution of the correlation between expression divergence and protein divergence in mammals. Genome biology and evolution,2013,5(7):1324-1335. DOI: 10.1093/gbe/evt093.

[23] Moyers B T, Rieseberg L H. Divergence in Gene Expression Is Uncoupled from Divergence in Coding Sequence in a Secondarily Woody Sunflower. International Journal of Plant Sciences, 2013, 174(7):1079-1089. DOI: 10.1086/671197.

[24] Severin Uebbing, Axel Künstner, Hannu Mäkinen, Niclas Backström, Paulina Bolivar, Reto Burri, Ludovic Dutoit, Carina F. Mugal, Alexander Nater, Bronwen Aken, Paul Flicek, Fergal J. Martin, Stephen M. J. Searle, Hans Ellegren. Divergence in gene expression within and between two closely related flycatcher species. Molecular Ecology,2016,25(9):2015-2028. DOI: 10.1111/mec.13596.

[25] Eduardo P.C. Rocha. The quest for the universals of protein evolution. Trends in Genetics,2006,22(8):412-416. DOI: 10.1016/j.tig.2006.06.004.

[26] Nuzhdin, S. V. Common Pattern of Evolution of Gene Expression Level and Protein Sequence in Drosophila. Molecular Biology and Evolution, 2004, 21(7):1308-1317. DOI: 10.1093/molbev/msh128.

[27] Liao, B.-Y. Low Rates of Expression Profile Divergence in Highly Expressed Genes and Tissue-Specific Genes During Mammalian Evolution. Molecular Biology and Evolution, 2006, 23(6):1119-1128. DOI: 10.1093/molbev/msj119.

[28] Ostlund Gabriel, Schmitt Thomas, Forslund Kristoffer, Köstler Tina, Messina David N, Roopra Sanjit, Frings Oliver, Sonnhammer Erik L L. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research,2009,38(Database i): D196-D203. DOI: 10.1093/nar/gkp931.

[29] Shuaibin Lian, Tianliang Liu, Shengli Jing, Hongyu Yuan, Zaibao Zhang and Lin Cheng. Intrachromosomal colocalization strengthens co-expression, co-modification and evolutionary conservation of neighboring genes. BMC Genomics, (2018), 19:455-. DOI: 10.1186/s12864-018-4844-1.

[30] Clark, R. M., Schweikert, G., Toomajian, C. Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. Science, 2007, 317(5836):338-342. DOI: 10.1126/science.1138632.

[31] Ji-Hye Jang, Yun Shang, Hyun Kyung Kang, Sun Young Kim, Beg Hab Kim, Kyoung Hee Nam. Arabidopsis galactinol synthases 1 (AtGOLS1) negatively regulates seed germination. Plant Science,2018,267:94-101. DOI: 10.1016/j.plantsci.2017.11.010.

[32] Chen Jui-Hung, Jiang Han-Wei, Hsieh En-Jung, Chen Hsing-Yu, Chien Ching-Te, Hsieh Hsu-Liang, Lin Tsan-Piao. Drought and Salt Stress Tolerance of an Arabidopsis Glutathione S-Transferase U17 Knockout Mutant Are Attributed to the Combined Effect of Glutathione and Abscisic Acid1. Plant Physiology,2012,158(1):340-351. DOI: 10.1104/pp.111.181875.

[33] Lee S Y, Boon N J, Webb A. A. R, Tanaka R. J. Synergistic Activation of RD29A via Integration of Salinity Stress and Abscisic Acid in Arabidopsis thaliana. Plant & Cell Physiology, 2016, 57(10):2147-2160. DOI: 10.1093/pcp/pcw132.

[34] Mark T. Waters, Adrian Scaffidi, Yueming K. Sun, Gavin R. Flematti, Steven M. Smith. The karrikin response system of A rabidopsis. The Plant Journal,2014,79(4). DOI: 10.1111/tpj.12430.

[35] Sangmin Lee, Pil Joon Seo, Hyo-Jun Lee, Chung-Mo Park. A NAC transcription factor NTL4 promotes reactive oxygen species production during drought-induced leaf senescence in Arabidopsis. The Plant Journal,2012,70(5). DOI: 10.1111/j.1365-313X.2012.04932.x.

[36] Li J, Zhong R, Palva E T. WRKY70 and its homolog WRKY54 negatively modulate the cell wall-associated defenses to necrotrophic pathogens in Arabidopsis. Plos One, 2017, 12(8): e0183731. DOI: 10.1371/journal.pone.0183731.

[37] Mirabella Rossana, Rauwerda Han, Allmann Silke, Scala Alessandra, Spyropoulou Eleni A, de Vries Michel, Boersma Maaike R, Breit Timo M, Haring Michel A, Schuurink Robert C. WRKY40 and WRKY6 act downstream of the green leaf volatile E-2-hexenal in Arabidopsis. The Plant journal: for cell and molecular biology,2015,83(6):1082-1096. DOI: 10.1111/tpj.12953.

[38] Nakata Masaru, Mitsuda Nobutaka, Herde Marco, Koo Abraham J K, Moreno Javier E, Suzuki Kaoru, Howe Gregg A, Ohme-Takagi Masaru. A bHLH-type transcription factor, ABA-INDUCIBLE BHLH-TYPE

TRANSCRIPTION FACTOR/JA-ASSOCIATED MYC2-LIKE1, acts as a repressor to negatively regulate jasmonate signaling in arabidopsis. The Plant cell,2013,25(5):1641-1656. DOI: 10.1105/tpc.113.111112.

[39] Mizuno Takeshi, Yamashino Takafumi. Comparative transcriptome of diurnally oscillating genes and hormone-responsive genes in Arabidopsis thaliana: insight into circadian clock-controlled daily responses to common ambient stresses in plants. Plant and Cell Physiology,2008,49(3):481-487. DOI: 10.1093/pcp/pcn008.

[40] Fatma Kaplan, Charles L. Guy. RNA interference of Arabidopsis beta-amylase8 prevents maltose accumulation upon cold shock and increases sensitivity of PSII photochemical efficiency to freezing stress. The Plant Journal,2005,44(5):14. DOI: 10.1111/j.1365-313x.2005.02565.x.

[41] Huang Junli, Gu Min, Lai Zhibing, Fan Baofang, Shi Kai, Zhou Yan-Hong, Yu Jing-Quan, Chen Zhixiang. Functional Analysis of the Arabidopsis PAL Gene Family in Plant Growth, Development, and Response to Environmental Stress. Plant Physiology,2010,153(4):1526-1538. DOI: 10.1104/pp.110.157370.

[42] Cuevas, Juan C, López-Cobollo, Rosa, Alcázar, Rubén, Zarza, Xavier, Koncz, Csaba, Altabella, Teresa,Salinas, Julio, Tiburcio, Antonio F, Ferrando, Alejandro. Putrescine Is Involved in Arabidopsis Freezing Tolerance and Cold Acclimation by Regulating Abscisic Acid Levels in Response to Low Temperature. Plant Physiology,2008,148(3):1094-1105. DOI: 10.4161/psb.4.3.7861.

[43] Maekawa Shugo, Inada Noriko, Yasuda Shigetaka, Fukao Yoichiro, Fujiwara Masayuki, Sato Takeo, Yamaguchi Junji. The carbon/nitrogen regulator ARABIDOPSIS TOXICOS EN LEVADURA31 controls papilla formation in response to powdery mildew fungi penetration by interacting with SYNTAXIN OF PLANTS121 in Arabidopsis. Plant physiology,2014,164(2):879-887. DOI: 10.1104/pp.113.230995.

[44] Raksha S, Seonghee L, Laura O, Elison B. Two chloroplast-localized proteins: AtNHR2A and AtNHR2B, contribute to callose deposition during nonhost disease resistance in Arabidopsis. Molecular Plant-Microbe Interactions, 2018: MPMI-04-18-0094-R-. DOI: 10.1094/MPMI-04-18-0094-R.

[45] Abdel-Ghany Salah Esmat. Contribution of plastocyanin isoforms to photosynthesis and copper homeostasis in Arabidopsis thaliana grown at different copper regimes. Planta,2009,229(4):767-779. DOI: 10.2307/23390386.

[46] Gao Huiling, Xie Wenxiang, Yang Changhong, Xu Jingyi, Li Jingjun, Wang Hua, Chen Xi, Huang Chao-Feng. NRAMP2, a trans-Golgi network-localized manganese transporter, is required for Arabidopsis root growth under manganese deficiency. The New phytologist,2018,217(1):179. DOI: 10.1111/nph.14783.

[47] Remy Estelle, Cabrito Tânia R, Baster Pawel, Batista Rita A, Teixeira Miguel C, Friml Jiri, Sá-Correia Isabel, Duque Paula. A major facilitator superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in Arabidopsis. The Plant cell,2013,25(3):901-926. DOI: 10.1105/tpc.113.110353.

[48] Consonni, Chiara, Bednarek, Pawel, Humphry, Matt, Francocci, Fedra, Ferrari, Simone, Harzen, Anne,van Themaat, Emiel Ver Loren, Panstruga, Ralph. Tryptophan-Derived Metabolites Are Required for Antifungal Defense in the Arabidopsis mlo2 Mutant. Plant Physiology,2010,152(3):1544-1561. DOI: 10.2307/25680756.

[49] Michael Hartmann, Tatyana Zeier, Friederike Bernsdorff, Vanessa Reichel-Deland, Denis Kim, Michele Hohmann, Nicola Scholten, Stefan Schuck, Andrea Bräutigam, Torsten Hölzel, Christian Ganter, Jürgen Zeier. Flavin Monooxygenase-Generated N-Hydroxypipecolic Acid Is a Critical Element of Plant Systemic Immunity. Cell,2018,173(2). DOI: 10.1016/j.cell.2018.02.049.

[50] Yosuke Maruyama, Natsuko Yamoto, Yuya Suzuki, Yukako Chiba, Ken-ichi Yamazaki, Takeo Sato, Junji Yamaguchi. The Arabidopsis transcriptional repressor ERF9 participates in resistance against necrotrophic fungi. Plant Science,2013,213:79-87. DOI: 10.1016/j.plantsci.2013.08.008.

[51] Jeon Jin, Cho Chuloh, Lee Mi Rha, Van Binh Nguyen, Kim Jungmook. CYTOKININ RESPONSE FACTOR2 (CRF2) and CRF3 Regulate Lateral Root Development in Response to Cold Stress in Arabidopsis. The Plant cell,2016,28(8):1828. DOI: 10.1105/tpc.15.00909.

[52] Zwack Paul J, Compton Margaret A, Adams Cami I, Rashotte Aaron M. Cytokinin response factor 4 (CRF4) is induced by cold and involved in freezing tolerance. Plant cell reports,2016,35(3):573-584. DOI: 10.1007/s00299-015-1904-8.

[53] Tsai K J, Chou S J, Shih M C. Ethylene plays an essential role in the recovery of Arabidopsis during post-anaerobiosis reoxygenation. Plant, Cell & Environment, 2014, 37(10). DOI: 10.1111/pce.12292.

[54] Wang Xiaoping, Liu Shanda, Tian Hainan, Wang Shucai, Chen Jin-Gui. The Small Ethylene Response Factor ERF96 is Involved in the Regulation of the Abscisic Acid Response in Arabidopsis. Frontiers in plant science,2015,6. DOI: 10.3389/fpls.2015.01064.

[55] Fan Min, Bai Ming-Yi, Kim Jung-Gun, Wang Tina, Oh Eunkyoo, Chen Lawrence, Park Chan Ho, Son Seung-Hyun, Kim Seong-Ki, Mudgett Mary Beth, Wang Zhi-Yong. The bHLH transcription factor HBI1 mediates the trade-off between growth and pathogen-associated molecular pattern-triggered immunity in Arabidopsis. The Plant cell,2014,26(2):828-841. DOI: 10.1105/tpc.113.121111.

[56] Eunkyoo O, Jia-Ying Z, Ming-Yi B, Augusto, A. R, Yu S, Zhi-Yong. Cell elongation is regulated through a central circuit of interacting transcription factors in the Arabidopsis hypocotyl. eLife, 2014, 3. DOI: 10.7554/eLife.03031.

[57] Frerigmann Henning, Berger Bettina, Gigolashvili Tamara. bHLH05 is an interaction partner of MYB51 and a novel regulator of glucosinolate biosynthesis in Arabidopsis. Plant physiology,2014,166(1):349-369. DOI: 10.1104/pp.114.240887.

[58] Ming Yang. The FOUR LIPS (FLP) and MYB88 genes conditionally suppress the production of nonstomatal epidermal cells in Arabidopsis cotyledons. American Journal of Botany,2016,103(9):1559-1566. DOI: 10.3732/ajb.1600238.

[59] Yan D, Easwaran V, Chau V, Okamoto M, Ierullo M, Kimura M. NIN-like protein 8 is a master regulator of nitrate-promoted seed germination in Arabidopsis. Nature Communications, 2016, 7:13179. DOI: 10.1038/ncomms13179.

[60] Sozzani Rosangela, Maggio Caterina, Giordo Roberta, Umana Elisabetta, Ascencio-Ibañez Jose Trinidad, Hanley-Bowdoin Linda, Bergounioux Catherine, Cella Rino, Albani Diego. The E2FD/DEL2 factor is a component of a regulatory network controlling cell proliferation and development in Arabidopsis. Plant Molecular Biology,2010,72(4-5):381-395. DOI: 10.1007/s11103-009-9577-8.

[61] Jiang Jianjun, Zhang Chi, Wang Xuelu. A recently evolved isoform of the transcription factor BES1 promotes brassinosteroid signaling and development in Arabidopsis thaliana. The Plant cell,2015,27(2). DOI: 10.1105/tpc.114.133678.

[62] Liao X, Bao H, Meng Y, Plastow G, Moore S. Sequence, Structural and Expression Divergence of Duplicate Genes in the Bovine Genome. Plos One, 2014, 9(7): e102868. DOI: 10.1371/journal.pone.0102868.

[63] Wen-Hsiung Li, Jing Yang, Xun Gu. Expression divergence between duplicate genes. Trends in Genetics,2005,21(11):602-607. DOI: 10.1016/j.tig.2005.08.006.

[64] Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nature Communications, 2014, 5:4956. DOI: 10.1038/ncomms5956.

[65] Jiménez-Delgado Senda, Pascual-Anaya Juan, Garcia-Fernàndez Jordi. Implications of duplicated cis-regulatory elements in the evolution of metazoans: the DDI model or how simplicity begets novelty. Briefings in functional genomics & proteomics,2009,8(4):266-275. DOI: 10.1093/bfgp/elp029.

[66] Zhong Zhenhui, Lin Lianyu, Chen Meilian, Lin Lili, Chen Xiaofeng, Lin Yahong, Chen Xi, Wang Zonghua, Norvienyeku Justice, Zheng Huakun. Expression Divergence as an Evolutionary Alternative Mechanism Adopted by Two Rice Subspecies Against Rice Blast Infection. Rice (New York, N.Y.),2019,12(1). DOI: 10.1186/s12284-019-0270-5.

[67] Vijayakumar G, Shaherin B, Prasannavenkatesh D, Sangdun C, Uversky V. N. Molecular Evolution and Structural Features of IRAK Family Members. PLoS ONE, 2012, 7(11): e49771-. DOI: 10.1371/journal.pone.0049771.

[68] Lehti-Shiu, Melissa D, Zou, Cheng, Hanada, Kousuke, Shiu, Shin-Han. Evolutionary History and Stress Regulation of Plant Receptor-Like Kinase/Pelle Genes. Plant Physiology,2009,150(1):12-26. DOI:

10.1104/pp.108.134353.

[69] Hanada K, Kuromori T, Myouga F, Toyoda T, Shinozaki K, Walsh B. Increased Expression and Protein Divergence in Duplicate Genes Is Associated with Morphological Diversification. PLoS Genetics, 2009, 5(12): e1000781. DOI: 10.1371/journal.pgen.1000781.

[70] Silvia Coolen, Silvia Proietti, Richard Hickman, Nelson H. Davila Olivas, Ping-Ping Huang, Marcel C. Van Verk, Johan A. Van Pelt, Alexander H.J. Wittenberg, Martin De Vos, Marcel Prins, Joop J.A. Van Loon, Mark G.M. Aarts, Marcel Dicke, Corné M.J. Pieterse, Saskia C.M. Van Wees. Transcriptome dynamics of Arabidopsis during sequential biotic and abiotic stresses. The Plant Journal, 2016, 86(3):249-267. DOI: 10.1111/tpj.13167.

[71] Anders, S. Differential gene expression analysis based on the negative binomial distribution. Journal of Marine Technology & Environment, 2009, 2(2).

[72] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008, 9(1):559. DOI: 10.1186/1471-2105-9-559.

[73] Yang Z H, Nielsen R, Goldman N, Pedersen A. M. K. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics, 2000, 155(1):431-449. DOI: 10.1002/1526-968X(200005)27:1<32::AID-GENE50>3.0.CO;2-T.

[74] Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Peer Y.V.D, Rouzé P, Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002, 30, 325–327. DOI: 10.1093/nar/30.1.325.

[75] Crooks G.E, Hon G, Chandonia J.M, Brenner S.E. WebLogo: A sequence logo generator. Genome Res. 2004, 14, 1188-1190. DOI: 10.1101/gr.849004.

[76] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 10(1):1523. Doi: 10.1038/s41467-019-09234-6.

Supplementary File Information

Supplementary data

Table. S1. The gene list of 6,481 paralogs.

Table. S2. The interaction information of paralogs.

Table. S3. The interaction information of repeats.

Table. S4. The list of FF, FP and PP paralogs under four different stresses.

Table. S5. The log2FC values of FF and FP paralogs under four different stresses.

Table. S6. The log2FC values of seven FF paralogs expression clusters.

Table. S7. The log2FC values of seven FP paralogs expression clusters.

Table. S8. The log2FC values of paralogs in enhancing patterns under four stresses.

Table. S9. The log2FC values of paralogs in decreasing patterns under four stresses.

Table. S10. Version information for 21 species.

Table. S11. The transcriptome information of four stresses.

Supplementary Figures

Figure S1. Workflow chart showing the different steps undertaken in this study.

Figure S2. The differential expression patterns of the FP paralogs under four different types of stress.

(A). A Venn diagram of the FP paralogs under four different types of stress.

(B). The number of transcription factors in each cluster of FP paralogs.

(C). A heatmap of seven expression modules of the FP paralogs under four different types of stress.

Figure S3. The frequency distributions of 10,000 repetitions of the randomized experiment for determining the *Ka* and *Ks* values, as well as the *Ka/Ks* ratio of FF and FP DEPs under four different types of stress. Navy blue corresponds to the randomized experiments, whereas the red dashed line corresponds to the real values.

Figure S4. The inversely proportional correlations between species conservation and family size of the paralogous gene pairs.

Figures

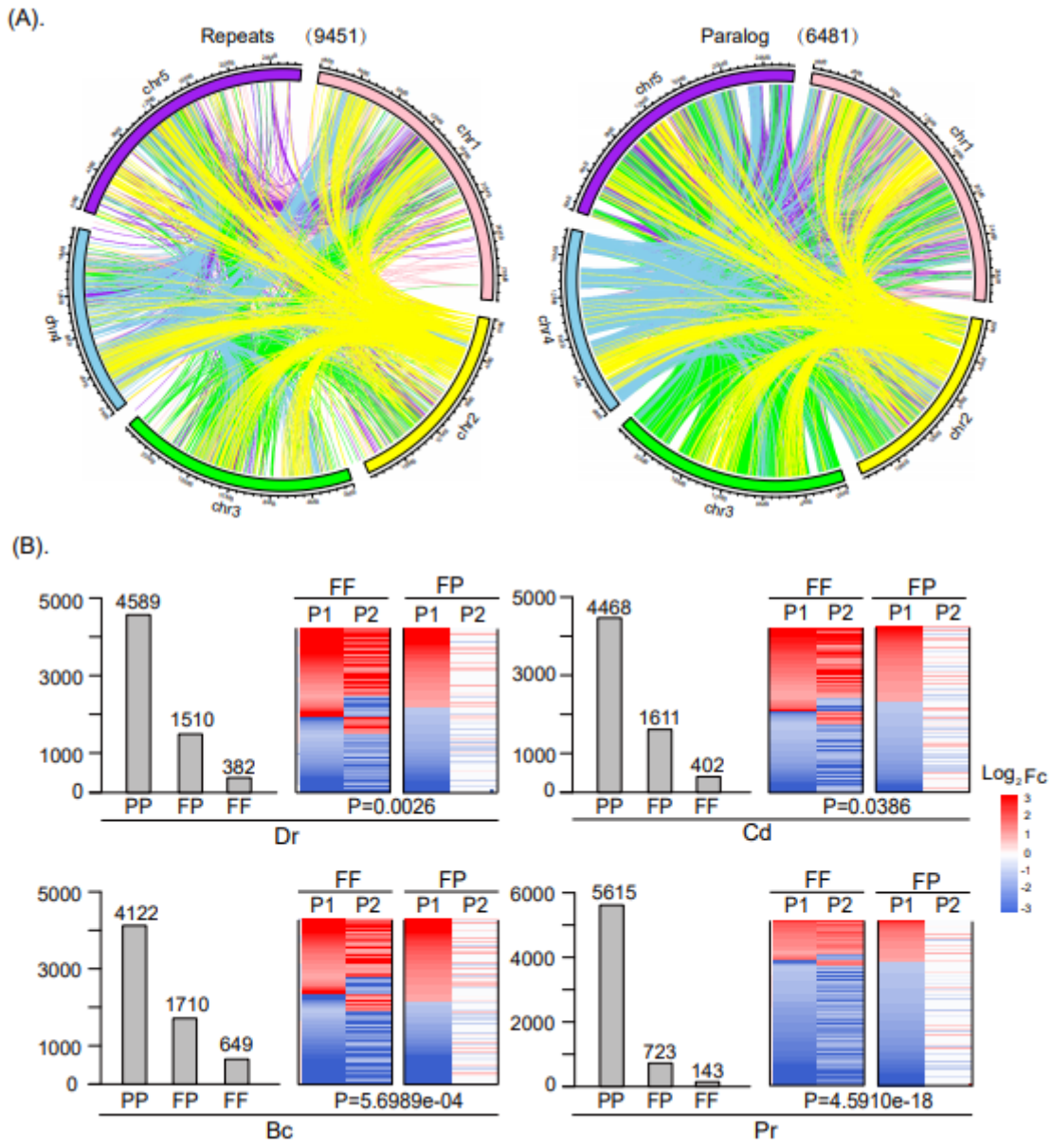

Figure 1

Distributions and expression classifications of the paralogs. (A) The distributions of 9,451 repetitive sequences and 6,481 paralogs throughout the chromosomes. (B) The identification and expression of the three types of paralogs under four different types of stress, including two biotic stresses (infection by the necrotrophic fungus Botrytis cinerea, Bc and herbivory by the chewing larvae of Pieris rapae, Pr) and two abiotic stresses (drought and cold).

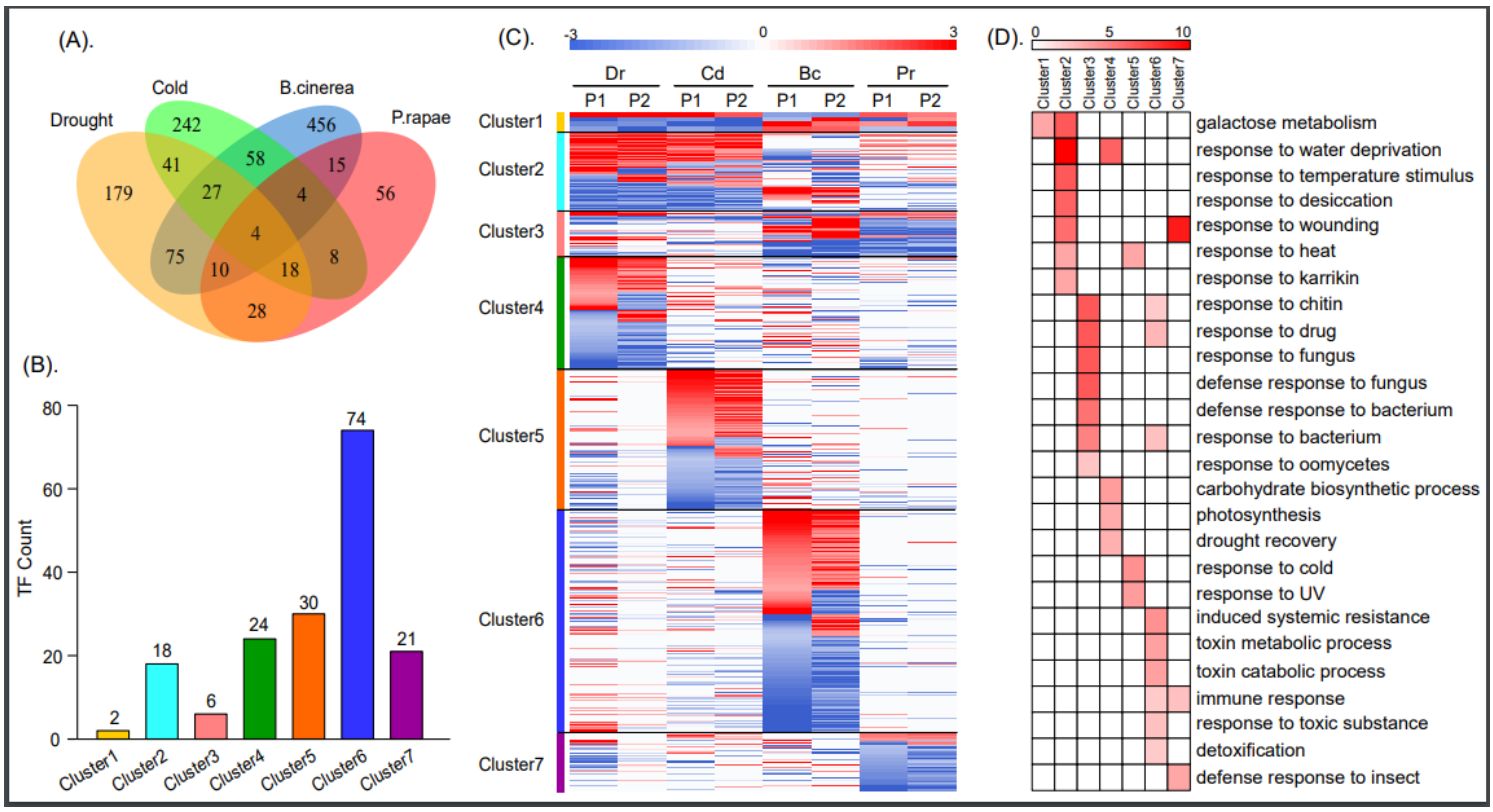

Figure 2

The differential expression patterns and functional enrichment of the FF paralogs under four different types of stress. (A) Venn diagram of the FF paralogs under four different types of stress. (B) The number of transcription factors in each cluster of FF paralogs. (C) Heatmap of seven expression modules of the FF paralogs under four different types of stress. Color bars represent the log2|FC| values, with red representing up-regulation and blue representing down-regulation. (D) The functional enrichment of seven clustered paralogs in the heatmap.

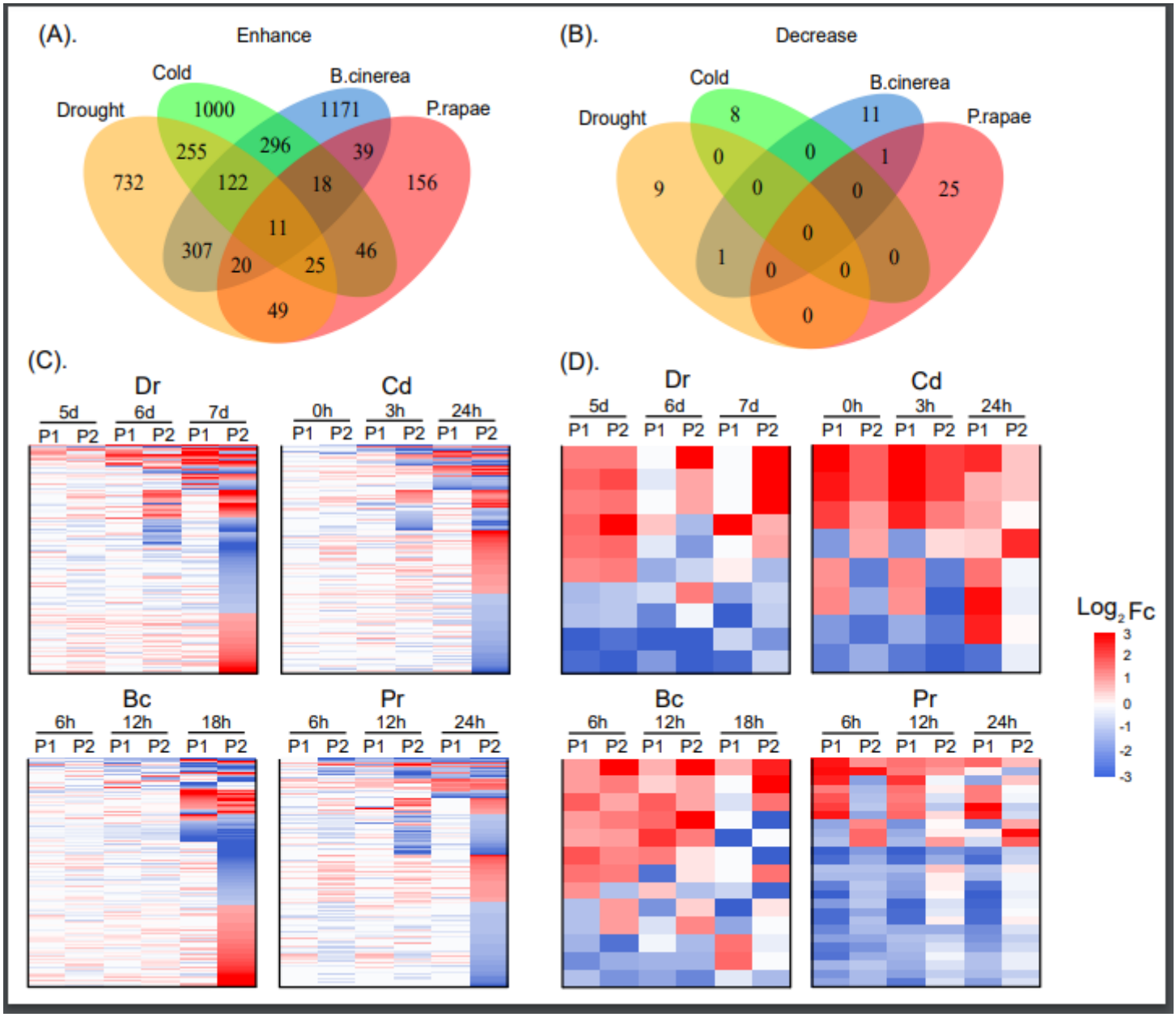

Figure 3

The expression pattern of enhancing and decreasing paralogs under different types of stress. (A) Venn diagram of paralogs with enhancing expression patterns under four different types of stress. (B) Venn diagram of paralogs with decreasing expression patterns under four different types of stress. (C) The heatmaps of paralogs with enhancing expression patterns under each stress condition. (D) The heatmaps of paralogs with decreasing expression patterns under each stress condition.

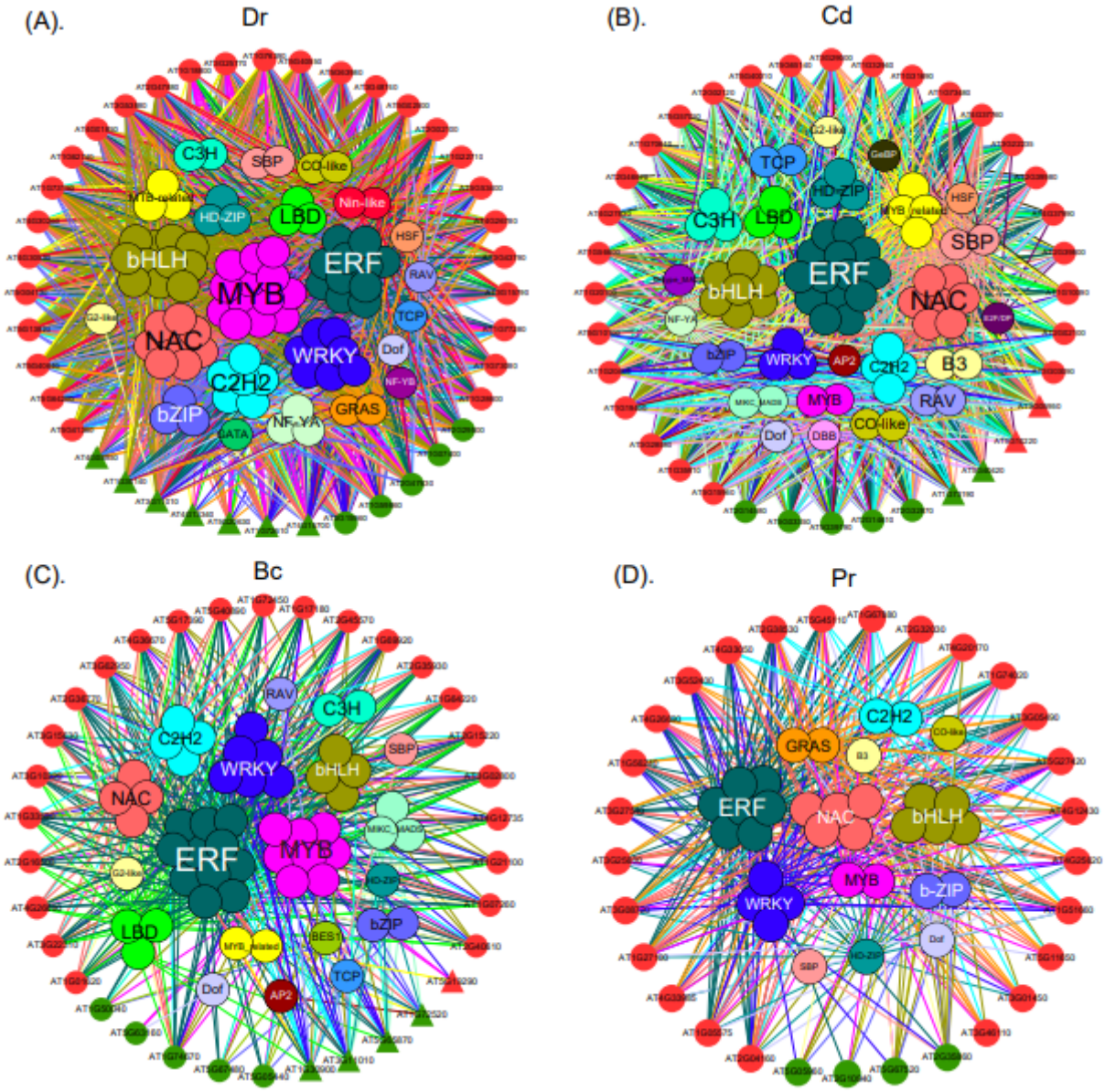

Figure 4

The co-expression networks of DEPs with enhancing and decreasing patterns under four different types of stress. The outer circle represents differentially expressed paralogs. Red (left) and green (right) represent paralogs with enhancing and decreasing expression patterns, respectively. Triangles represent the up-regulated genes, while circles represent the down-regulated genes. The inner circle represents the co-expressed TFs.

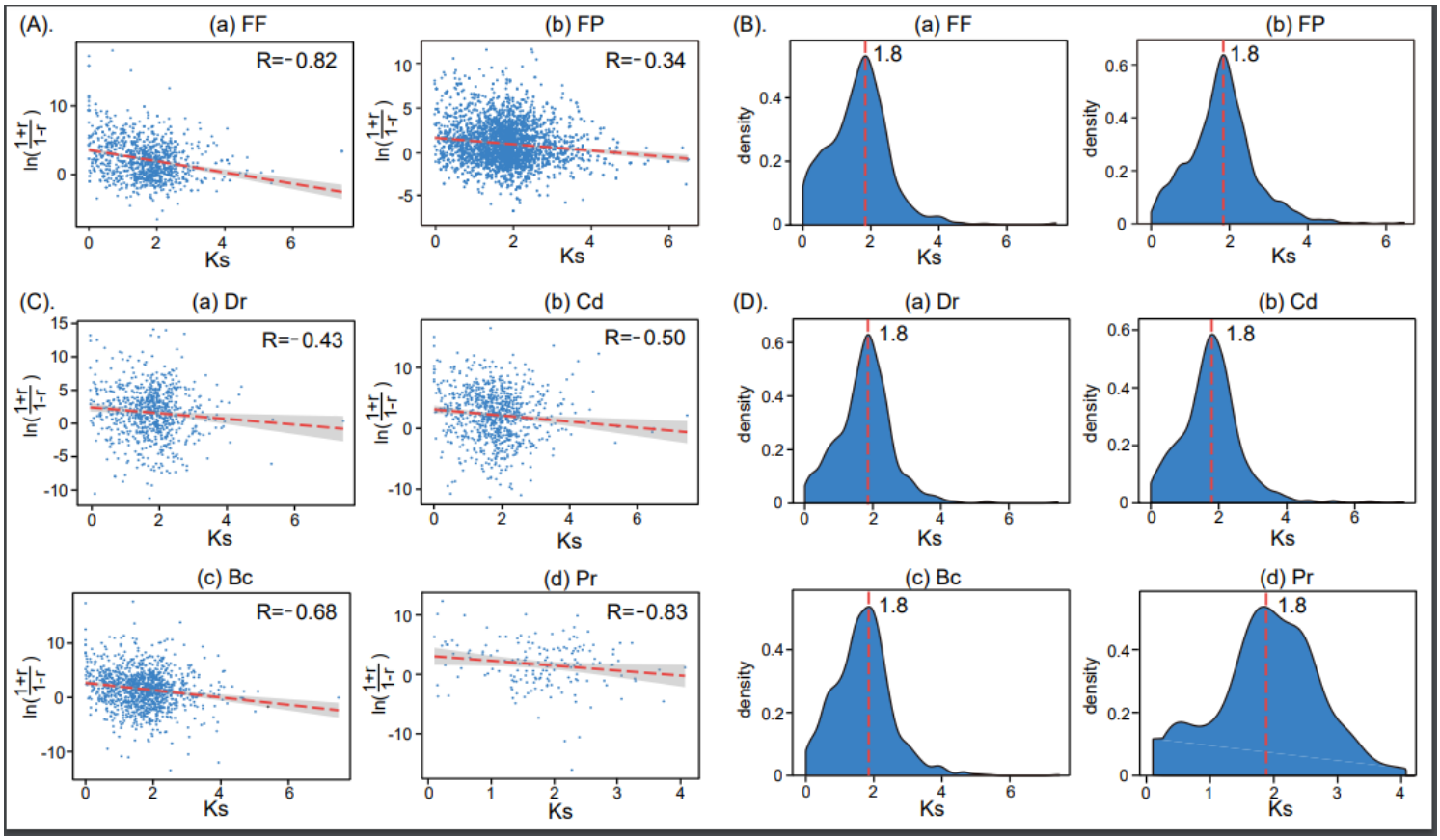

Figure 5

The regression results of the expression divergences and sequence divergences. (A) The regression results of FF (a) and FP (b) paralogs under all four types of stress. (B) The density plot of Ks values of FF (a) and FP (b) paralogs under all four types of stress. (C) The regression results of paralogs with enhancing and decreasing expression under each stress condition. (a) Dr, (b) Cd, (c) Bc and (d) Pr. (D) The density plot of Ks values of paralogs with enhancing and decreasing expression under each of the four types of stress. (a) Dr, (b) Cd, (c) Bc and (d) Pr.

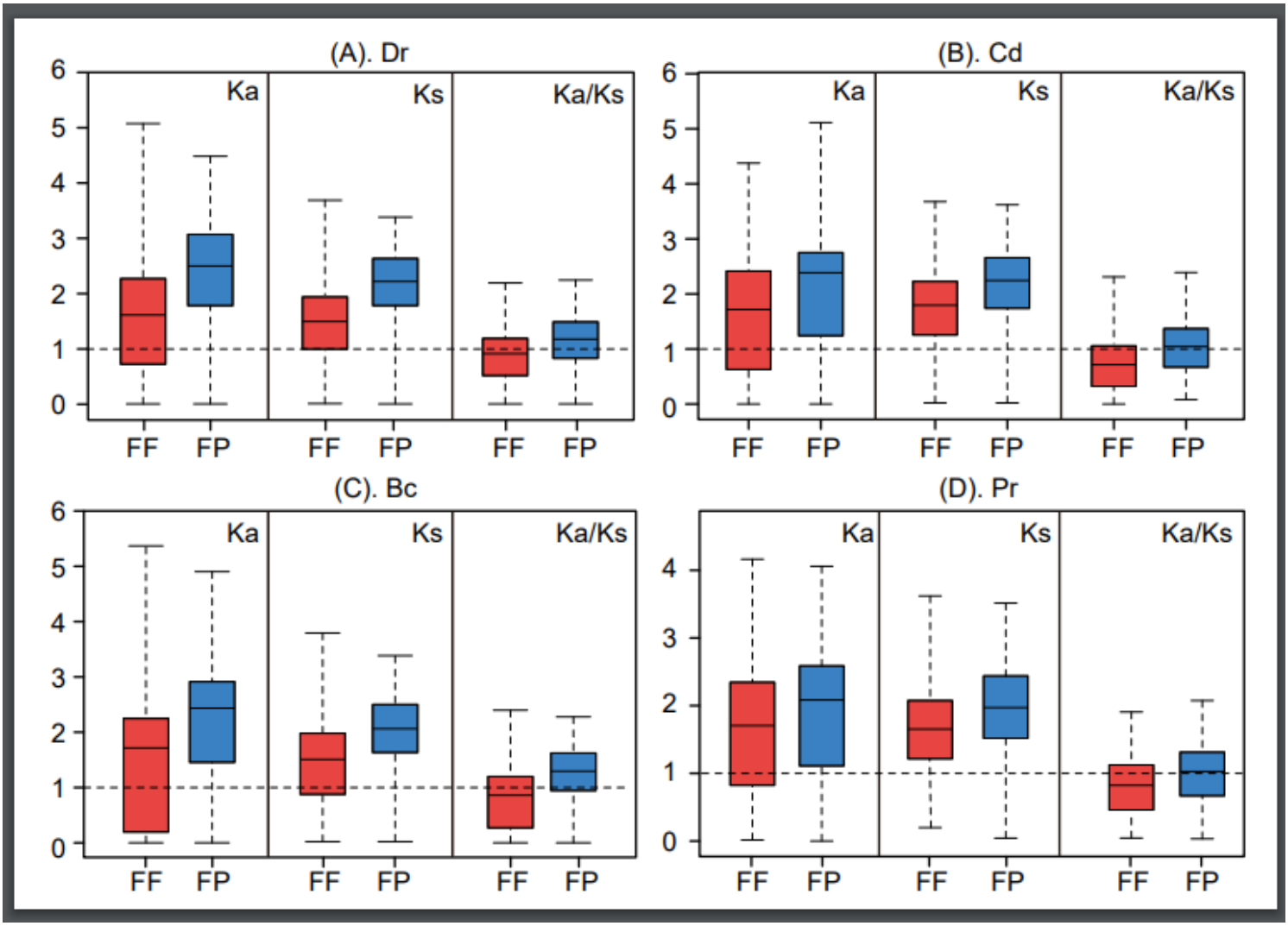

Figure 6

Boxplot of Ka, Ks and Ka/Ks of FF and FP DEPs under four different stress conditions.

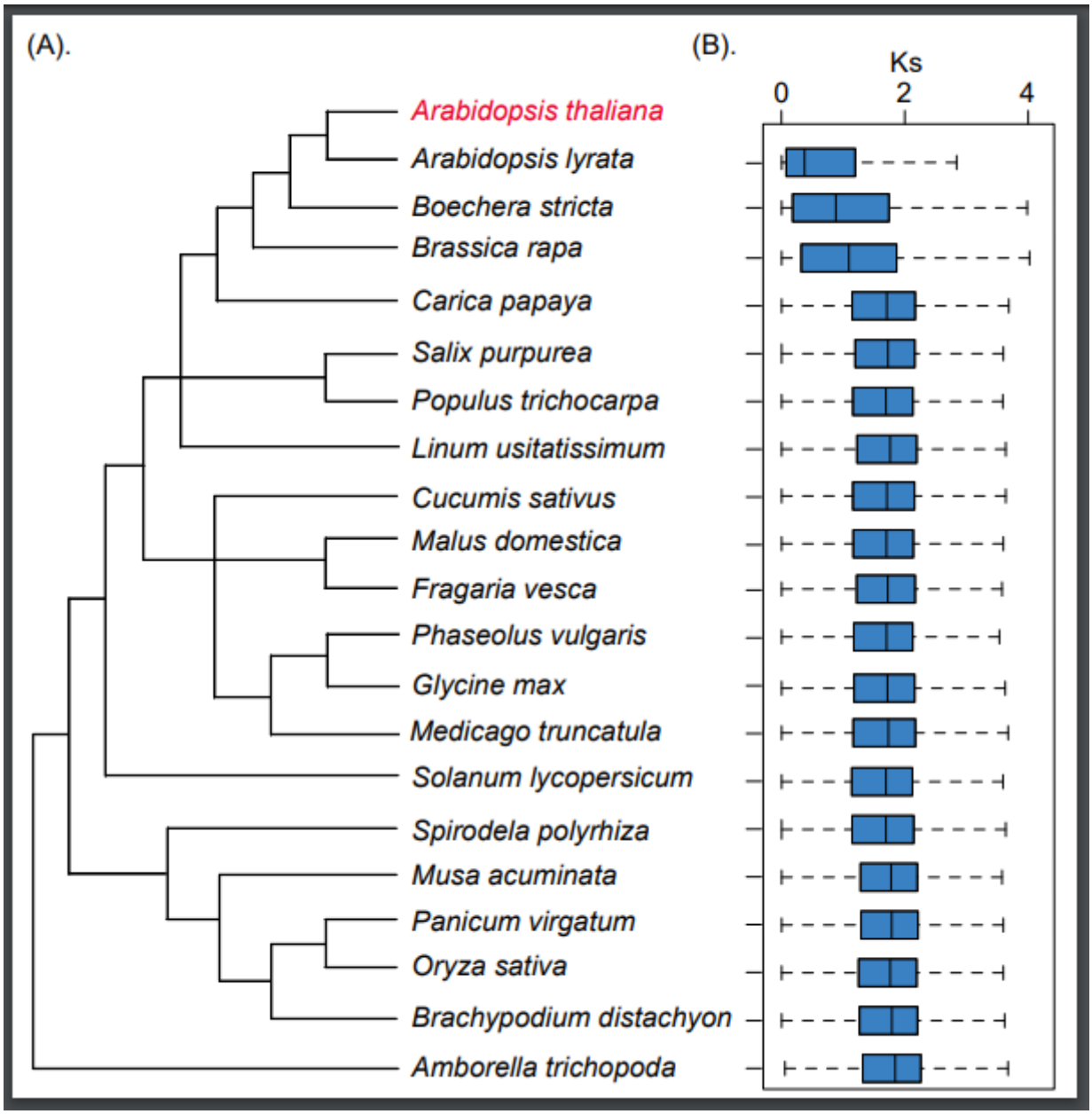

Figure 7

(A) The phylogenetic tree of the 21 species presented in this study. (B) The boxplot of the Ks values of the paralogs between Arabidopsis thaliana and the other 20 species.

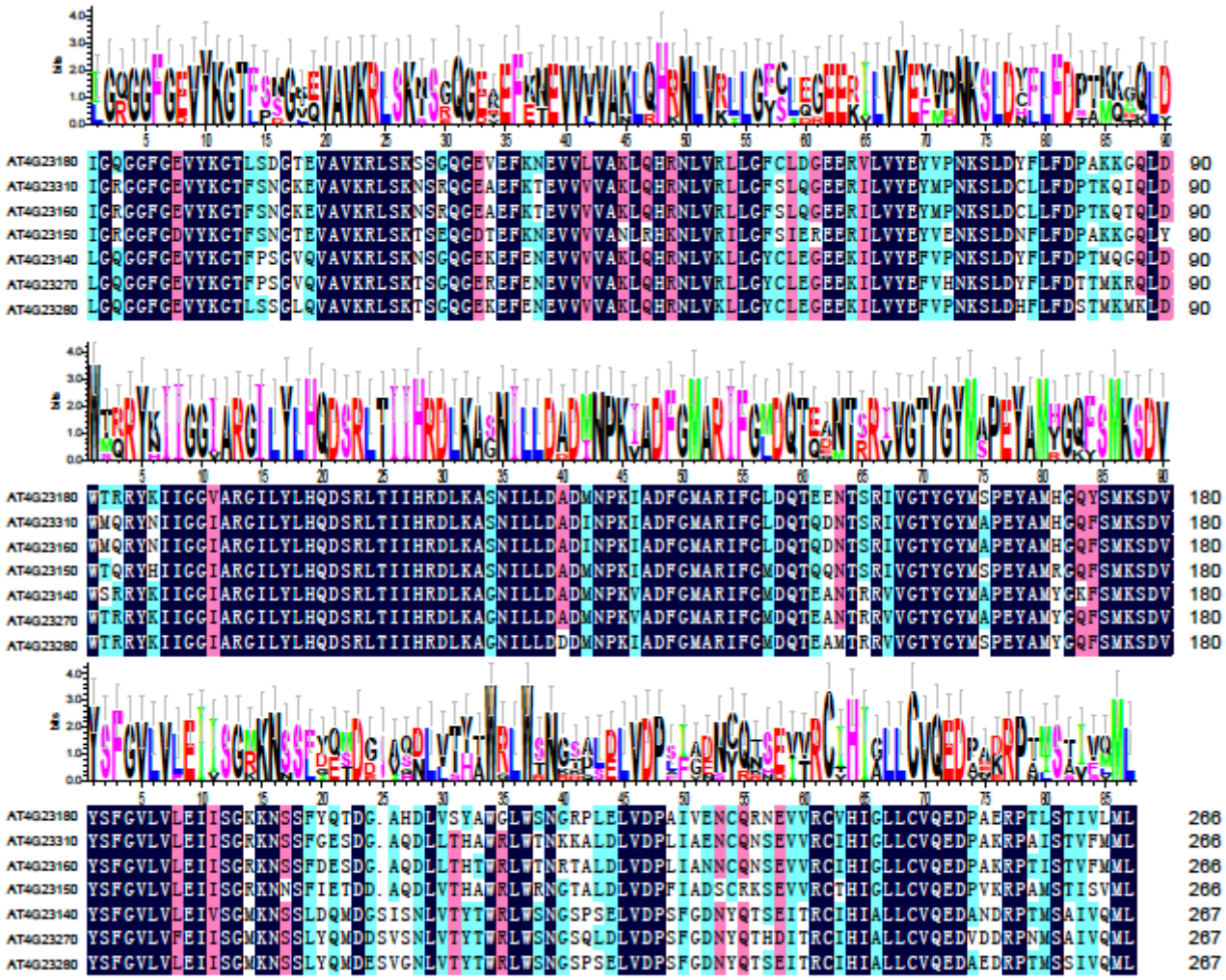

Figure 8

The conserved protein domain sequences of the paralogs in all 21 species.

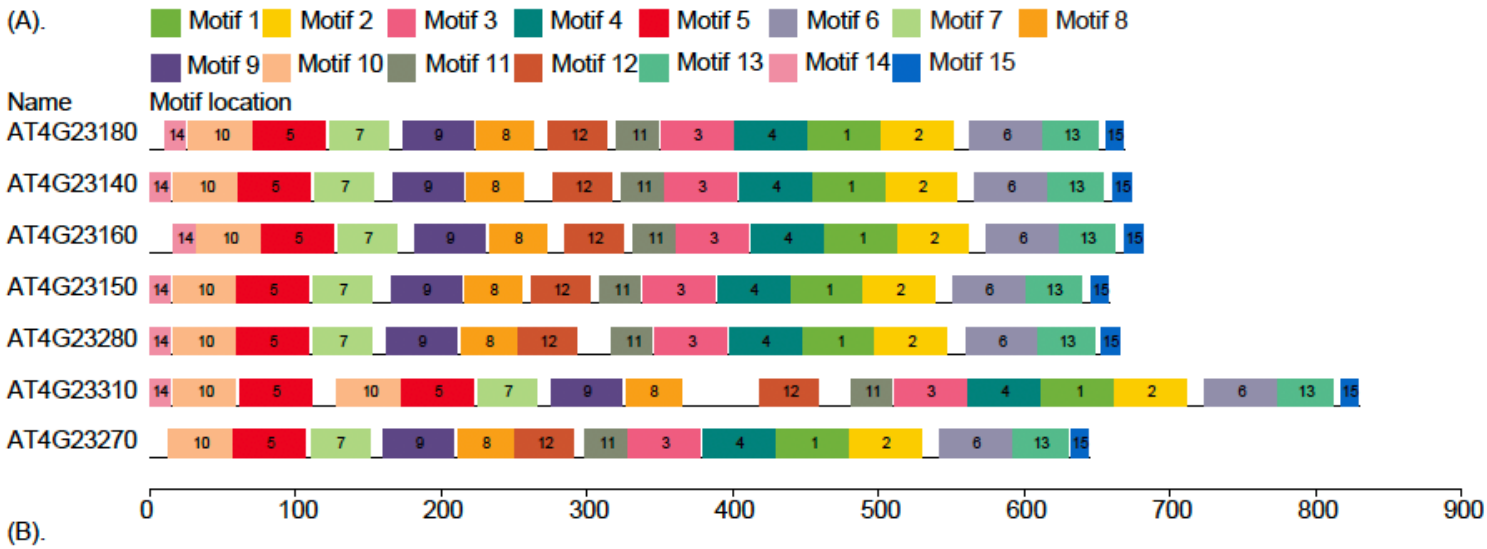

Figure 9

The conserved motifs of the paralogs in all 21 species.

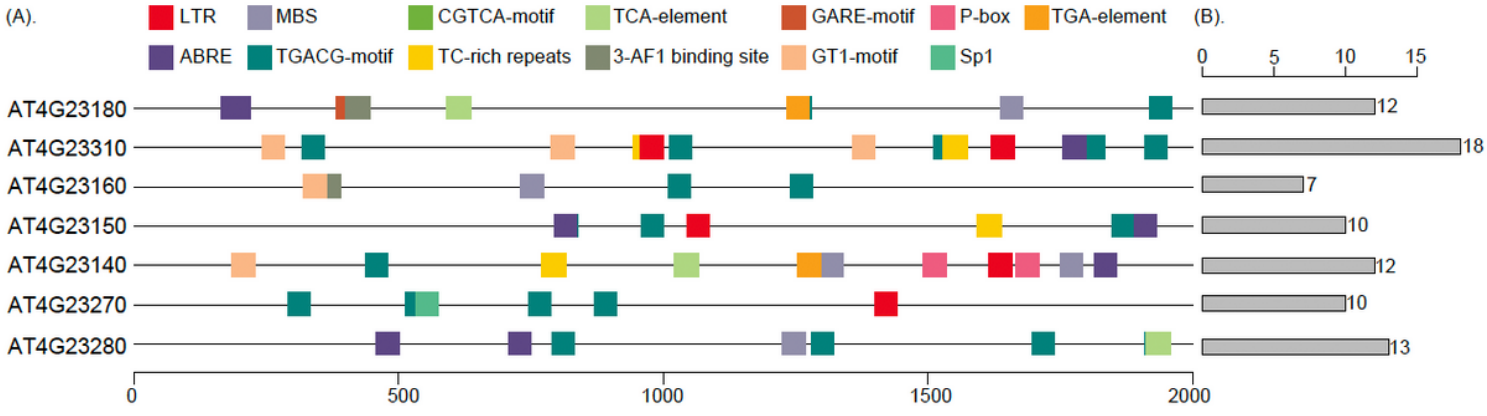

Figure 10

(A) The top ten cis-elements of STKc_IRAK in the 2000-bp promoter sequence. (B) The number of cis-elements in each gene.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fig.S4.pdf
- TableS8.xlsx
- TableS7.xlsx
- TableS2.xlsx
- TableS11.xlsx
- TableS3.xlsx
- Equations.docx
- TableS6.xlsx
- TableS9.xlsx
- TableS10.xlsx
- Fig.S2.pdf
- TableS1.xlsx
- TableS4.xlsx
- TableS5.xlsx
- Fig.S1.pdf
- Fig.S3.pdf