DNA Research 20, 437–448, (2013)
Advance Access publication on 19 May 2013
doi:10.1093/dnares/dst021

Genome-Wide Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants

Hai Du[1,2], Yong-Bin Wang[1], Yi Xie[1], Zhe Liang[3], San-Jie Jiang[4], Shuang-Shuang Zhang[1], Yu-Bi Huang[1,*], and Yi-Xiong Tang[2,*]

Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region of Ministry of Agriculture, Maize Research Institute of Sichuan Agricultural University, Chengdu, Sichuan 611130, China[1]; Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China[2]; Department of Plant and Environmental Sciences, Norwegian University of Life Sciences, PO Box 5003, Ås N-1432, Norway[3] and Department of Food and Bioproduct Sciences, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A8[4]

*To whom correspondence should be addressed. Tel. +86-13908160283.
Email: yubihuang@sohu.com (Y.-B.H.); tangyx@mail.caas.net.cn (Y.-X.T.)

Edited by Dr Satoshi Tabata
(Received 18 December 2012; accepted 23 April 2013)

Abstract

MYB proteins constitute one of the largest transcription factor families in plants. Recent evidence revealed that MYB-related genes play crucial roles in plants. However, compared with the R2R3-MYB type, little is known about the complex evolutionary history of MYB-related proteins in plants. Here, we present a genome-wide analysis of MYB-related proteins from 16 species of flowering plants, moss, *Selaginella*, and algae. We identified many MYB-related proteins in angiosperms, but few in algae. Phylogenetic analysis classified MYB-related proteins into five distinct subgroups, a result supported by highly conserved intron patterns, consensus motifs, and protein domain architecture. Phylogenetic and functional analyses revealed that the Circadian Clock Associated 1-like/R-R and Telomeric DNA-binding protein-like subgroups are >1 billion yrs old, whereas the I-box-binding factor-like and CAPRICE-like subgroups appear to be newly derived in angiosperms. We further demonstrated that the MYB-like domain has evolved under strong purifying selection, indicating the conservation of MYB-related proteins. Expression analysis revealed that the MYB-related gene family has a wide expression profile in maize and soybean development and plays important roles in development and stress responses. We hypothesize that MYB-related proteins initially diversified through three major expansions and domain shuffling, but remained relatively conserved throughout the subsequent plant evolution.

Key words: MYB-related transcription factors; classification; evolution; phylogenetic analysis; expression profile analysis

1. Introduction

MYB proteins are characterized by a conserved DNA-binding domain and constitute one of the largest families of transcription factors (TFs) in plants, which are classified into four major groups according to the number of adjacent repeats in the DNA-binding domain.[1] All four groups are found in plants. The most common is the 2R-MYB group. The second group comprises a heterogeneous collection of R3- or 1R-MYB type proteins, hereafter referred to as MYB-related proteins, which usually contain a single MYB repeat. The third and fourth groups are composed of 3R-MYB and 4R-MYB type proteins, respectively. These latter groups consist of only 1−5 members.

The first plant MYB-encoding gene, *C1* (2R-MYB), was isolated from maize (*Zea mays*).[2] Accordingly, research on MYB genes has mainly focused on the 2R-MYB gene family because of its large size.[1] In the last two decades, a vast number of plant 2R-MYB genes have been shown to play important roles in many plant-specific processes. The first plant MYB-related gene (*MybSt1*) was isolated from potato.[3] The numerous MYB-related genes subsequently identified play key roles as transcriptional regulators,[3,4] circadian clock-associated repressors,[5,6] and telomeric repeat-binding proteins[7,8] in diverse biological processes. To date, genome-wide analyses of 2R-MYB proteins have been conducted in numerous plant species based on sequenced genomes[1,9–11] However, comprehensive analyses of MYB-related proteins in major land plants are still lacking. Accordingly, the evolutionary relationships between plant MYB-related proteins remain unknown, necessitating a detailed survey and classification of disparate evolutionary groups.

To understand the evolutionary history of plant MYB-related genes, we identified MYB-related proteins at the genome-wide level and performed structural and evolutionary analyses across distantly related plant evolutionary lineages, including eudicots, monocots, a gymnosperm, a bryophyte, five chlorophyte species, and a rhodophyte. Subsequently, we assessed the origins, patterns of differentiation, and expansion of different phylogenetic subgroups of this gene family. In addition, we analysed the expressions of MYB-related genes in different tissues and developmental stages and under stress treatments.

2. Materials and methods

2.1. *Sequence retrieval*

We performed a BLASTP search among sequenced genomes of land plants in Phytozome (http://www.Phytozome.net) using well-known plant MYB-related proteins as queries. The species represented a broad range of the plant lineages from unicellular green algae to multicellular plants (http://www.jgi.doe.gov/). To verify the reliability of our results, all putative non-redundant sequences were assessed with PROSITE profiling[12] and SMART analysis,[13] respectively.

2.2. *Multiple sequence alignments*

Multiple alignments of MYB domains in candidate genes were performed using the MAFFT version 7 software under default parameters.[14] Nucleotide substitution levels were calculated using the HyPhy version 2.0.[15] The HyPhy batch Quick Selection Detion.bf was used to estimate site-by-site variation in rates.

2.3. *Phylogenetic analysis*

A neighbour-joining (NJ) tree was constructed using the MEGA version 5 software,[16] based on the alignment of MYB domains. To determine the statistical reliability, we conducted bootstrap analysis with the following parameters: *p*-distance and pairwise deletion. Bootstrap analysis was performed with 1000 replicates.

2.4. *Detection of conserved motifs*

Conserved motifs of MYB-related proteins were identified statistically with the MEME program.[17] The following parameter settings were used: maximum number of motifs, 100; minimum width of motif, 6; and maximum width of motif, 300. All putative motifs with expected values of $<1E-30$ were discarded. In addition, we used the PFAM tool to identify whether any remaining motifs matched well-known motifs.[18]

2.5. *Gene expression analysis*

Maize and soybean public expression datasets were obtained from the Plant Expression Database (PLEXdb).[19] Additionally, maize and soybean microarray-based datasets, with accession numbers GSE16567, GSE40052, GSE19501, GSE10023, GSE31188, GSE31763, GSE15100, GSE35427, and GSE18827, were downloaded from the PLEXdb. A hierarchical cluster was created using the Cluster 3.0.[20]

3. Results and discussion

3.1. *Identification of MYB-related proteins in plants*

To identify MYB-related proteins in land plants, we implemented BLASTP searches of the complete genomes of the red alga (*Cyanidioschyzon merolae*); the chlorophytes (*Volvox carteri*, *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, and *Chlorella vulgaris*); the moss (*Physcomitrella patens*); the lycophyte (*Selaginella moellendorffii*); the eudicots (*Arabidopsis thaliana*, *Citrus sinensis*, *Populus trichocarpa*, *Glycine max*, *Vitis vinifera* L., and *Solanum lycopersicum*), and the monocots (maize and *Brachypodium distachyon*). Each matching sequence was then used to search the respective genome databases until no new sequences were found.

Further analyses focused only on proteins with full-open reading frames. We referred to the sequences of MYB-related proteins in the Plant Transcription Factor Database (PlantTFDB)[21] and PlnTFDB,[22] and reconfirmed the sequences by comparative analysis. Given the substantial sequence divergence of MYB-related genes, their identification should be careful manual checking. GARP-like TFs are often confused with MYB-related TFs,[23] because they contain a consensus sequence (SHLQKY) very similar to that of CCA1-like proteins (SHAQK(Y/F)F).[4] However, they contain only 1 of the 3 regularly spaced Trp (W) residues found in the MYB domain.[23,24] Thus, we excluded GARP-like TFs in this study.

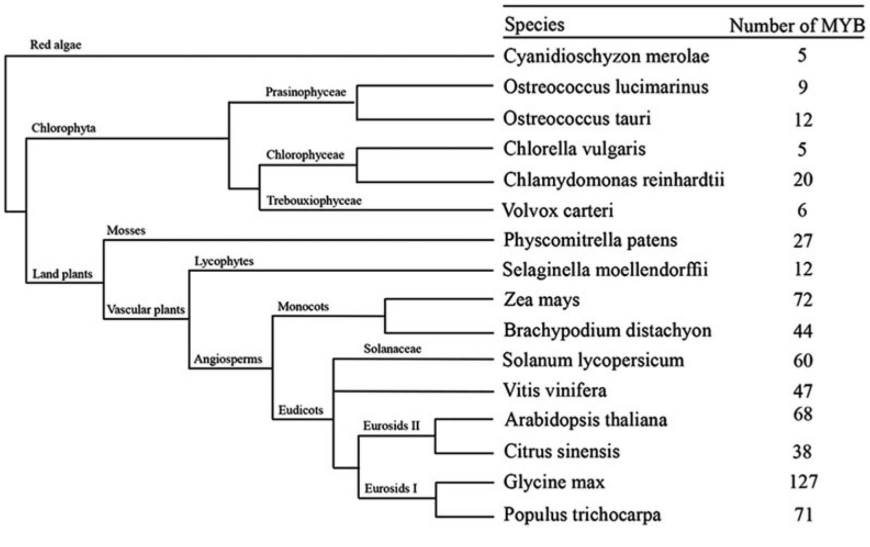
After removing incomplete or redundant sequences, and predicted alternative splice variants, we identified 623 MYB-related genes (Supplementary Table S1). We also identified a large number of putative MYB-related proteins in angiosperms, but only a small number in land plants that diverged earlier (Fig. 1). This suggests that a huge expansion occurred after the evolution of angiosperm plants. In this study, we excluded false positives of MYB-related proteins according to the Yanhui's criterion.[25] Interestingly, we retrieved four genes not previously annotated as MYB-related genes in *Arabidopsis*. In addition, we retrieved a small number of genes in unicellular green algae and red alga (Fig. 1), which suggests that MYB-related proteins arose before plants transitioned from water to land.

3.2. *Phylogenetic analysis of MYB-related proteins*

To investigate the evolution of plant MYB-related proteins, we constructed an NJ tree (Fig. 2 and Supplementary Fig. S1) based on the alignment of the MYB domains. Based on the topology and clade support values, the 623 MYB-related proteins were classified into five major subgroups with robust bootstrap support (generally ≥60%), CCA1-like/R-R, I-box-like, CPC-like, TRF-like, and TBP-like (Fig. 2).

The CCA1-like/R-R and TBP-like subgroups were the largest of the five subgroups. All subgroups were present in monocots and eudicots (Fig. 2), indicating that the appearance of most MYB-related genes in plants predates the divergence of monocot/eudicots. Meanwhile, in contrast to 2R-MYB genes,[10,11] no species-specific subgroups and/or clades were observed, implying that MYB-related genes were more conserved during evolution. In addition, MYB-related genes from the same lineage tended to cluster together in the phylogenetic tree and were not equally represented within a given clade, suggesting that they experienced duplications after the lineages diverged. Additional features used for validation, discussed below, strongly supported the reliability of the clustering results.

The CCA1-like/R-R subgroup comprised of four major clades (Clades I–IV) with different intron patterns (a–d; Figs 2 and 3). The common characteristic of this subgroup is a highly conserved motif, SHAQK(Y/F)F,[4] in the third helix of the MYB domain. Whereas most of the plant MYB-related proteins contain a single MYB domain, R-R proteins, similar to 2R-MYB proteins, have two MYB domains. However, unlike in 2R-MYB proteins, the two MYB repeats in R-R proteins are separated, in the N-terminal and middle regions, respectively. The second repeats are more closely related to the MYB domains of CCA1-like proteins, and they clustered as the second clade of the CCA1-like/R-R subgroup. The TBP-like subgroup contained five distinct clades (Fig. 2, Clades I–V) with conserved characteristics. Although the bootstrap value of the TBP-like subgroup node was low, the reliability of the clustering was supported by the presence of the consensus motif LKDKW(R/K)(N/T),[26] the intron patterns, and the architectures of the non-MYB motifs (Fig. 2). The TRF-like subgroup comprised of a limited genes from all land plants investigated, except for *Selaginella*. This suggests that the TRF-like subgroup was also conserved during evolution. The CPC-like subgroup consisted of two distinct clades (Fig. 2, I and II); one contained angiosperm genes, and the other contained algae genes. Interestingly, members of the



Species	Number of MYB
Cyanidioschyzon merolae	5
Ostreococcus lucimarinus	9
Ostreococcus tauri	12
Chlorella vulgaris	5
Chlamydomonas reinhardtii	20
Volvox carteri	6
Physcomitrella patens	27
Selaginella moellendorffii	12
Zea mays	72
Brachypodium distachyon	44
Solanum lycopersicum	60
Vitis vinifera	47
Arabidopsis thaliana	68
Citrus sinensis	38
Glycine max	127
Populus trichocarpa	71

Figure 1. Phylogenetic relationships between all species investigated in this study. The total number of MYB-related proteins found in each genome is indicated on the right.

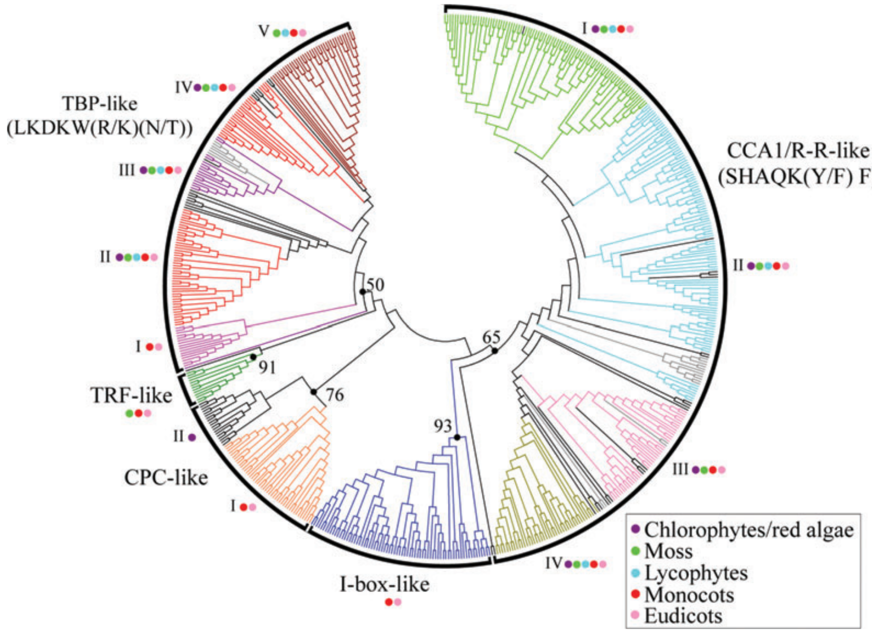


Figure 2. NJ analysis of 623 plant MYB-related proteins. The proteins clustered into five major subgroups, CCA1-like/R-R, I-box-like, CPC-like, TRF-like, and TBP-like. The numbers beside the branches represent bootstrap support values from 1000 replications. The coloured lines indicate the intron pattern as shown in Fig. 3. The coloured dots symbolize the species to which the proteins in each clade belong. The major clades of each subgroup are numbered consecutively.

plant-only clade were characterized by very short sequences without transcription–activation domains. The lack of moss and lycophyte proteins in this clade suggests that either these genes were lost in early diverged land plants during expansion or were evolved after gymnosperms diverged. The high bootstrap values for the node supported that the two clades likely clustered in a subgroup.

In our phylogenetic analysis, most of the MYB-related genes fell into a subgroup, except for *AtMYBR48*, which was classified as an orphan gene.

3.3. *Conserved characteristics in the MYB domain*

Alignment analysis revealed the MYB domains of the five subgroups are remarkable divergence, but are highly similar within each subgroup or clade (Fig. 3). Similar to 2R-, 3R-, and 4R-MYB proteins,[10] MYB-related proteins also contained the three evenly distributed Trp (W) residues characteristic of MYB repeats. However, 13% and 65% of the plant MYB-related genes had a substitution at either the first or the third Trp (W) residue in the MYB domain, respectively. Most members of the TRF-like and TBP-like subgroups contained the three W residues. In contrast, the third W residue was often substituted by Ala (A) and Tyr (Y) in CCA1-like/R-R and I-box-like subgroups, respectively (Fig. 3). While the first W residue was substituted by Phe (F) in all members of the fourth clade of the TBP-like subgroup and in most members of the CPC-like subgroup.

Interestingly, despite the divergence of the individual MYB domains, the consensus sequences SHAQK(Y/F)F and LKDKW(R/K)(N/T) were highly conserved in the MYB domains of the CCA1-like/R-R and TBP-like subgroups, respectively (Figs 2 and 3), thus providing unique criteria for identifying these types of MYB proteins. The existence of highly conserved, subgroup-specific sites in the MYB domains also indicates a common origin, despite variability among the different subgroups.

The third helix in the MYB domain plays a major role in recognizing *cis*-elements in target genes, whereas the conserved W residues are important for forming the hydrophobic core and maintaining the three-dimensional structure of the MYB repeat. This suggests that the molecular structures and biological functions of each subgroup and/or clade were highly conserved during evolution.

3.4. *Conservation of intron/exon structure within MYB domains*

To determine the intron patterns of MYB-related genes, we analysed the intron distribution in regions encoding MYB domains. Most of the plant MYB-related genes (∼87%) were disrupted by intron(s), with up to two introns. In contrast, ∼13% of MYB-related genes did not contain introns (Fig. 3, I-box-like subgroup).

Our results revealed that the intron patterns (Fig. 3a–k), formed by relative position and phase, were highly

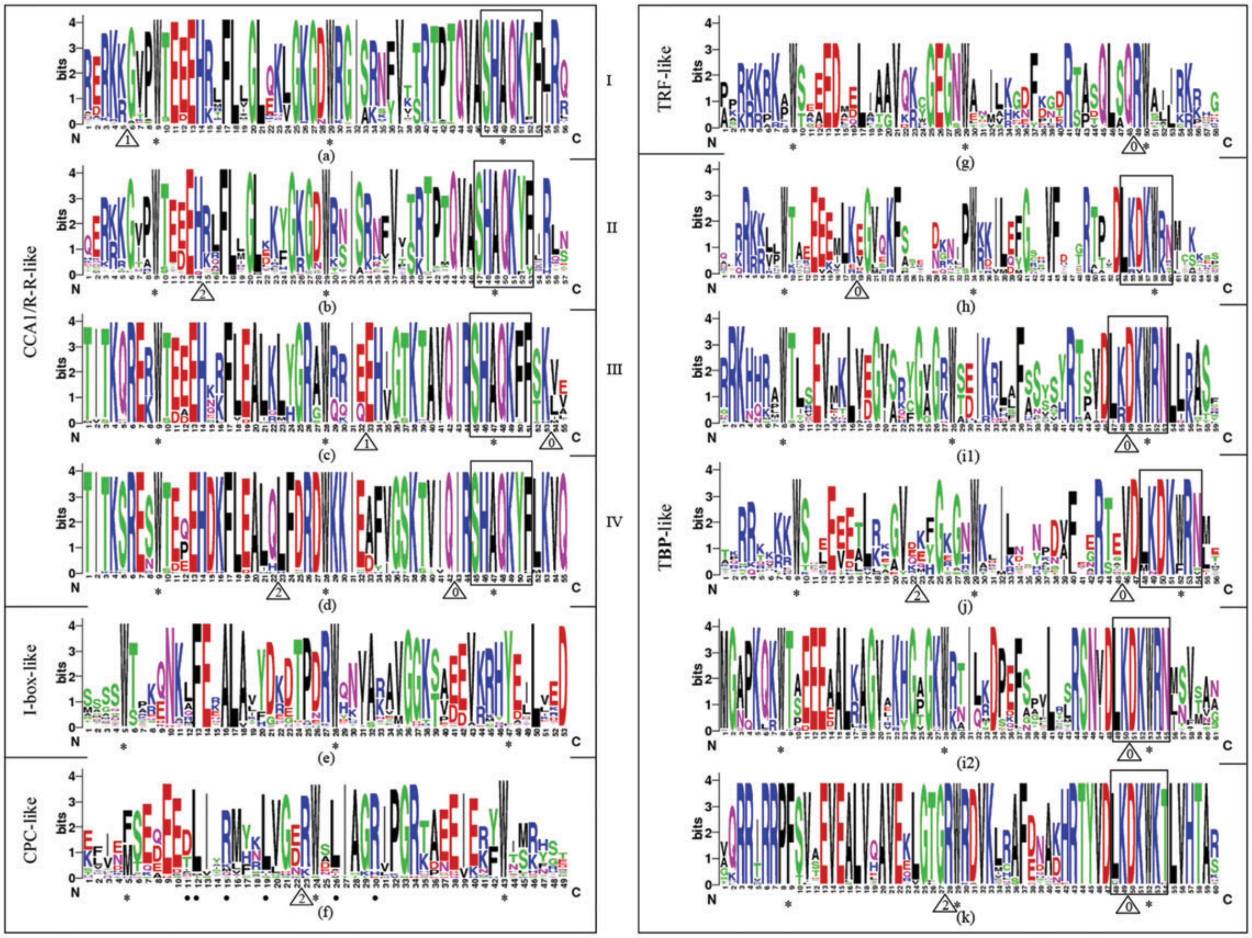


Figure 3. Sequence logos of the MYB domains of plant MYB-related proteins. The bit score indicates the information content for each position in the sequence. Asterisks indicate conserved Trp (W) residues in the MYB domain. Dots indicate the conserved motif (DLx2Rx3Lx6Lx3R). Black boxes indicate conserved motifs in the MYB domains. The intron patterns of land plant MYB-related genes are denoted a–k. White triangles indicate the locations of introns, and the number within each triangle indicates the splicing phases of introns. The corresponding clades in the NJ tree (Fig. 2) are listed on the right, for reference.

conserved in all subgroups and/or clades of plant MYB-related genes. The highly conserved intron patterns within subgroups or clades provided an independent criterion for testing the reliability of our phylogenetic analysis (Fig. 2). The intron patterns in algae were not conserved and generally quite different from those in land plants. However, an algae CCA1-like gene, *VcMYBR05*, showed the same intron pattern (Fig. 3c) as that in the third clade of the CCA1-like/R-R subgroup, strongly supporting their common origin.

In addition, the intron phases were highly conserved in plant MYB-related genes within each subgroup. For instance, pattern (a) always was in Phase 1, while pattern (d) was consistently in Phases 2 and 0 (Fig. 3), resulting in a significant excess of non-symmetrical exons. This suggests that splicing phases were also highly conserved during the evolution. Overall, our results indicate a strong correlation between the phylogeny and exon/intron structure of the MYB-related gene family.

3.5. *Molecular evolution of plant MYB-related genes*

To analyse the selective pressures acting during the expansion of plant MYB-related genes, we investigated the influences of selective constraints on the MYB domains. By globally fitting an evolutionary model, we first calculated the dN/dS ratios for each subgroup. The dN/dS values were substantially <1 in all subgroups, providing a crude indication that the strong purifying selection has been maintained across land plants (Supplementary Table S2). At the individual codon level, most of the residues were under significant negative selection ($P < 0.05$).

Because the CCA1-like/R-R and TBP-like subgroups subdivided into several clades (Fig. 2), the preceding method merely estimated the dN/dS ratio across each subgroup, without considering variations among clades in the large subgroups. Therefore, we estimated the dN/dS ratios for the clades of the CCA1-like/R-R and TBP-like subgroups (Supplementary Table S2).

In general, the dN/dS values of individual clades were lower than that of the subgroups. However, the dN/dS values of some individual clades were higher than that of the subgroups, and the number of residues under significant negative selection was reduced ($P < 0.05$). However, no clades showed dN/dS values >1, suggesting that different clades were subjected to different strengths of purifying selection. For example, in TBP-like subgroup clades, the dN/dS values ranged from 0.06 to 0.32, while in CCA1-like/R-R subgroup clades, the dN/dS values were <0.11. Thus, our dN/dS analysis suggests that selective constraints have remained stable throughout the evolution of MYB-related genes in land plants.

3.6. *Distribution of MYB domains and non-MYB motifs in plants*

The MYB domains were found throughout the entire coding region of MYB-related proteins, even within different clades of a subgroup (Fig. 4). For example, within the TBP-like subgroup, the MYB domain is at the N-terminal region in the second clade and at the C-terminal region in the third clade. Similarly, the MYB domains of the CCA1-like/R-R subgroup are located either at the N-terminal or at the middle region. Thus, the location of MYB domains is less conserved in MYB-related proteins than in 2R-MYB proteins.[11] These results illustrate the variability in the relative locations of the MYB domain and the high divergence of MYB-related proteins.

Because sequences outside of the MYB domains are quite divergent, non-conserved subgroup-specific motifs were detected. However, we identified 34 clade-specific motifs in the CCA1-like/R-R, TRF-like, and TBP-like subgroups (Supplementary Table S3). No motifs were detected in the CPC-like or I-box-like subgroup, because they lack the C-terminal (Fig. 4). Motifs 10, 11, 13, 14, and 16 in the CCA1-like/R-R subgroup and motifs 19, 21, 29, 30, 31, and 33 in the TBP-like subgroup were found only in angiosperms, suggesting that they are angiosperm-specific motifs that originated after the evolution of angiosperms. Motifs 1, 9, and 34 were adjacent to MYB domains, indicating that they co-evolved with the corresponding MYB domain (Fig. 4). In addition, motifs 1, 9, and 23 were present in most of the chlorophyta and/or red algae MYB-related proteins, suggesting that they are ancient. Overall, the protein architectures of closely related members in a specific clade were remarkably conserved, indicating a common origin and/or close relationship.

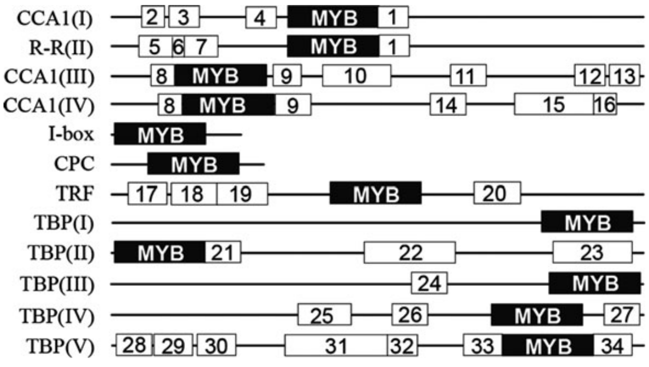


Figure 4. Architecture of conserved protein motifs in plant MYB-related subgroups and/or clades. An idealized representation of a typical member of each clade is shown, with the MYB domain and conserved motifs drawn as numbered boxes. The diagrams are not drawn to scale.

We also queried the PFAM database of protein domains using the candidate non-MYB motifs. With the exception of motif 22, none of the 34 conserved motifs corresponded to known domains. Motif 22, shared by members of the second clade of the TBP-like subgroup, showed significant homology with linker histones H1 and H5, which bind the nucleosome as a major component of chromatin and play a role in chromatin dynamics.[27] In addition, in the same clade, we identified a region near the C-terminus that may form a coiled-coil domain (motif 23). Such domains, found in many TFs, are predicted to stabilize protein dimer formation.[28] The presence of motifs 22 and 23 verified our phylogenetic classification and suggested a specific role for this type of MYB-related protein. Motifs 5–7 formed the first MYB repeat of R-R proteins, demonstrating that the first repeat is less homologous to the typical MYB domain than the second repeat.

The conservation of these additional motifs demonstrates that the diversity of domain architecture has been maintained beyond the core components of the MYB domain, while the presence of clade-specific motifs indicates their recent common origin. Therefore, they may be essential for the function of MYB-related proteins.

3.7. *Expression analysis of MYB-related genes at different developmental stages*

To understand the temporal and spatial expression patterns of MYB-related genes, we compared their expression patterns during maize and soybean development.

Microarray data of 60 different tissues/developmental conditions of maize[29] were used (Fig. 5). Few genes were constitutively expressed in all organs and developmental stages. CCA-like/R-R genes were expressed in most organs examined, with the exception of seeds. However, six genes in a CCA1-like/R-R subgroup clade (*ZmMYBR02, ZmMYBR11*, *ZmMYBR34*, *ZmMYBR42*, *ZmMYBR65*, and *ZmMYBR67*) showed higher expression in seeds than in other organs, which indicated that they may play important roles in seed development (Fig. 5). Similar to CCA1-like/R-R proteins, the I-box-like genes

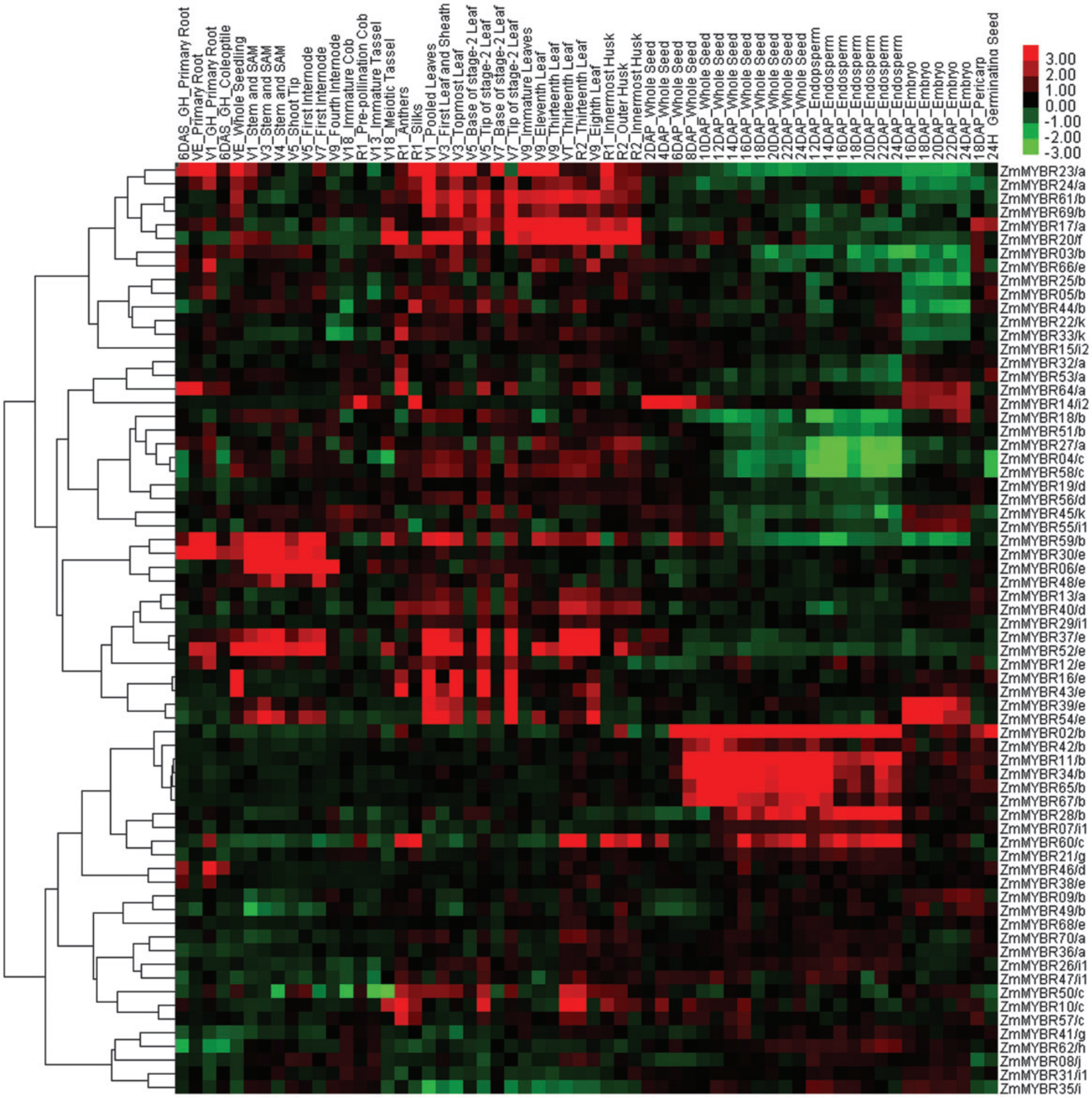


Figure 5. Expression profiles of MYB-related genes in maize across different developmental stages and organs. The genes and their corresponding intron patterns are on the right. The tissues used for expression analysis are indicated at the top of each column. The colour bar represents log2 expression values.

were also expressed abundantly in many maize organs. This may indicate that these two subgroups predominantly contribute to maize development. The expression of I-box-like genes and most of the CCA1-like/R-R genes significantly decreased during seed development, further implying roles as negative regulators in seed development. A CPC-like gene, *ZmMYBR20*, was highly expressed in leaf tissues, which suggested that it may function in leaf development, or it may be restricted by 2R-MYB genes in other developmental stages (see below).

The TRF-like subgroup included only two maize genes, which showed relatively high expression in seeds. Although no TRF-like genes have yet been functionally characterized in plants, their preferential expression in maize seed tissues implies their possible roles in seed development. The TBP-like subgroup, consisting of five clades, contained 14 maize genes. Although members of this subgroup displayed relatively low expression in all examined organs, TBP-like genes have wider expression in maize (Fig. 5). Furthermore, closely related genes generally showed highly similar

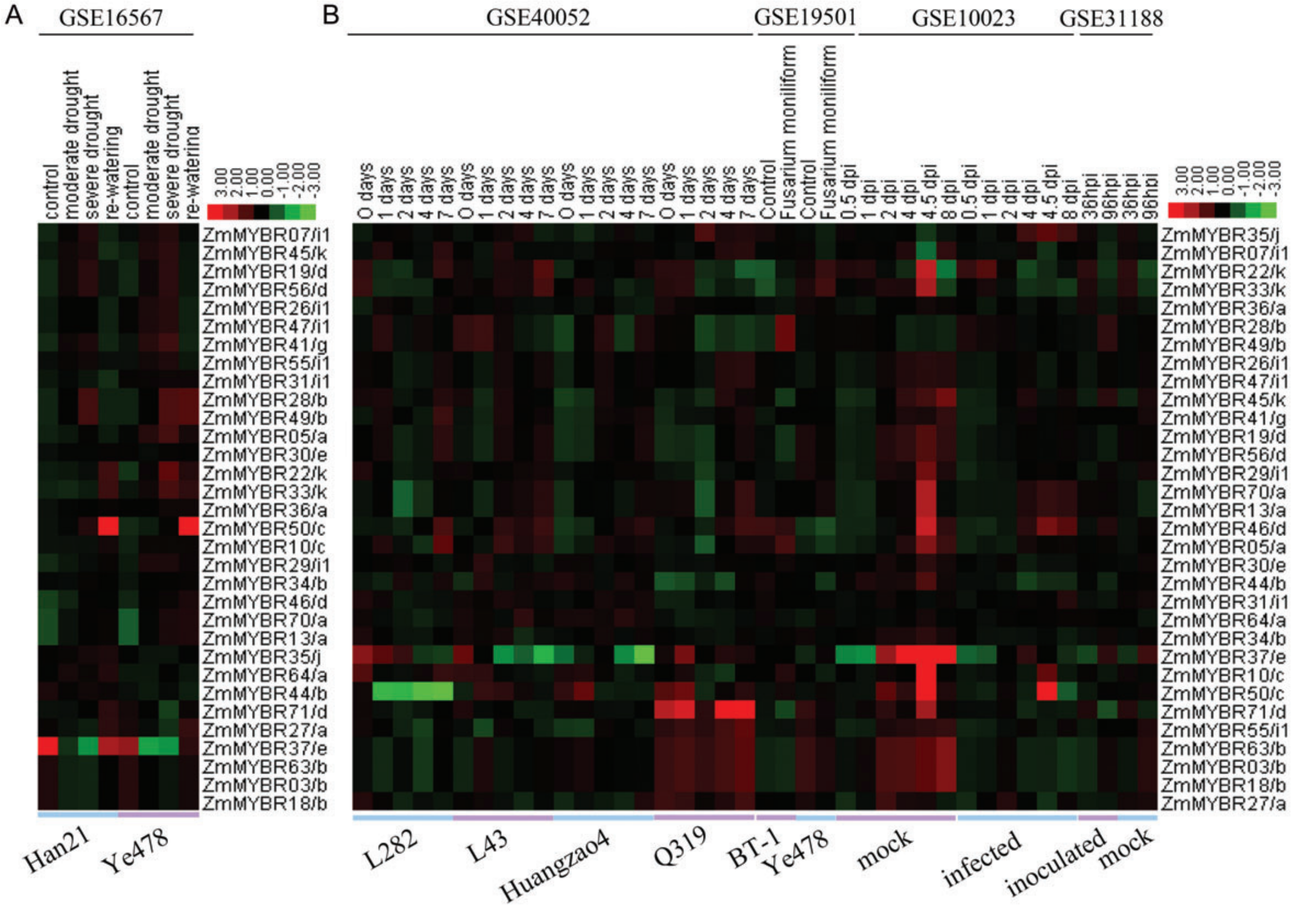


Figure 6. Expression profiles of maize MYB-related genes in response to drought stress or fungal infection. (A) The expression profiles of maize MYB-related genes under drought stress. (B) The expression of maize MYB-related genes after fungal infection.

expression patterns, indicating that they may share similar or overlapping functions.

We next analysed the expression profiles of soybean MYB-related genes.[30] The majority of the 127 soybean MYB-related genes showed wide expressions in the examined tissues. However, 22 soybean genes were not expressed in this dataset, suggesting that they might be pseudogenes. In most of the cases, the expression patterns of MYB-related genes in maize and soybean were very similar (Supplementary Fig. S2). The expression patterns of soybean genes divided into two main groups. Most of the CCA1-like/R-R genes showed prominent responses in the early stage of soybean development, while some TBP-like genes were expressed at higher levels in leaves and seeds. There were also minor differences in the expression patterns of I-box-like and CPC-like genes between maize and soybean. Some soybean I-box-like and CPC-like genes showed higher expression in more tissues, suggesting that these genes may play wider roles in soybean. The high similarity of MYB-related gene expression in maize and soybean indicates functional conservation of this gene family in plants.

3.8. Expression analysis of MYB-related genes under biotic and abiotic stresses

We further examined the roles of maize MYB-related genes in drought stress, based on the microarray data[19] (Fig. 6A). Twenty-six probe sets on the maize 18k GeneChip corresponded to 32 MYB-related genes (five probes represented more than one gene). The high sequence similarity necessitated further experimental confirmation. Most of the maize MYB-related genes were expressed at low levels, but were preferentially expressed under specific stress conditions. The expressions were very similar in tolerant (Han21) and sensitive (Ye478) lines. Among the genes, four CCA1-like/R-R genes (*ZmMYBR19*, *ZmMYBR28*, *ZmMYBR49*, and *ZmMYBR56*), one TRF-like gene (*ZmMYBR41*), and six TBP-like genes (*ZmMYBR07*, *ZmMYBR26*, *ZmMYBR31*, *ZmMYBR45*, *ZmMYBR47*, and *ZmMYBR55*) increased in response to drought stress. In contrast, the expression of five CCA1-like/R-R genes (*ZmMYBR03*, *ZmMYBR18*, *ZmMYBR27*, *ZmMYBR44*, and *ZmMYBR63*) and one I-box-like gene (*ZmMYBR37*) was significantly down-regulated by drought stress and recovered after re-watering. Thus, MYB-related genes likely contribute to the drought response.

To explore the roles of maize MYB-related genes in the response to pathogens, we investigated their expressions after the treatment with *Sphacelotheca reiliana*, *Fusarium moniliforme*, *Ustilago maydis*, or *Colletotrichum graminicola*.[19] As shown in Fig. 6B, the majority of maize genes analysed were differentially expressed over time after inoculation with these four pathogens. In general, the genes showed similar expression patterns in response to each of the pathogens. For example, *ZmMYBR05*, *ZmMYBR45*, and *ZmMYBR56* were up-regulated after infection with the four pathogens. However, some MYB-related genes also showed different expression patterns in different lines in response to the same pathogen. Furthermore, the expression of maize MYB-related genes varied more with time after *U. maydis* infection. Taken together, our results showed that MYB-related genes might participate in the maize pathogen response.

We used the Illumina transcriptome sequencing data[31] to assess the expressions of soybean MYB-related genes under pathogen stress (Supplementary Fig. S3A). Most of the soybean MYB-related genes were induced after infection with *Bradyrhizobium japonicum*. Moreover, the expression patterns differed significantly between root hair and stripped root samples: many genes exhibited much higher expression in root hairs than in stripped roots. In most of the cases, the soybean genes were differentially up-regulated upon *B. japonicum* infection. However, some genes were reduced. To extend the expression analysis of soybean MYB-related gene, we used the Affymetrix array data housed within the PLEXdb.[20] Thirty-two probes corresponded to individual soybean genes, of which 22 matched 2 genes and 1 matched 3 genes. Most of the soybean genes were strongly induced in hypocotyls infected with *Phytophthora sojae* (Supplementary Fig. S3B), which suggests these genes also contribute to the pathogen response. In addition, a few soybean genes, such as *GmMYBR028* and *GmMYBR126*, were up-regulated after infection with aphids, indicating a possible function.

3.9. *Evolution and divergence of MYB-related proteins*

Our phylogenetic analysis allowed us to assess the origin and evolutionary relationships among different subgroups. In the CCA1-like/R-R subgroup, the inclusion of all five chlorophyte algae and the red alga implies that this subgroup predates the divergence of red algae from the ancestor of land plants 1.5 billion yrs ago.[32] This subgroup contained four major clades (Fig. 2). The first two clades were characterized by the location of the MYB domain and the presence of motif 1 adjacent to the C-terminus of the MYB domain; these two clades were further distinguishable by a number of clade-specific motifs (Fig. 4). Similar results were also observed for the other two clades. Interestingly, all four clades included chlorophyte algae proteins, suggesting that they differentiated from a common ancestor before the origin of land plants. Clades III and IV appeared to be relatively older since they clustered with several red algae proteins (Fig. 2).

In contrast, the I-box-like subgroup seems to have evolved recently in angiosperms. No obvious orthologues were detected in algae, moss, or *Selaginella*. Although the MYB domains of different subgroups were generally quite divergent, those of I-box-like proteins were highly homologous (∼40% identity) to the first MYB repeats of R-R proteins. One significant difference was an amino acid deletion in I-box-like proteins (Supplementary Fig. S4). Consistent with a previous study,[25] when both MYB repeats of R-R proteins were used in the phylogenetic analysis, the second repeats clustered within the CCA1-like/R-R subgroup, while the first repeats clustered within the I-box-like subgroup with high bootstrap values (Supplementary Fig. S5). These results imply that I-box-like proteins evolved from R-R proteins through the gene disruption among the angiosperms ∼415 million yrs ago.[33] During this process, I-box-like proteins likely evolved from the first repeats in R-R proteins, while the second repeats formed the first clade of the CCA1-like/R-R subgroup. This view is supported by the conserved intron patterns of I-box-like genes and the first MYB repeats of R-R genes (both of which are intronless).

The TBP-like subgroup is composed of MYB-related proteins from chlorophyte algae and land plants, suggesting that it is >1 billion yrs old.[34] Among its five major clades, the second, third, and fourth clades are likely the oldest, because they contain algae proteins. While the second and fourth clades of this subgroup share the same intron pattern (i), the positions of their MYB domains differ (Fig. 4). Since both of these clades include algae proteins, domain shifting could have occurred earlier in plants. Similar results were also found in the CCA1-like/R-R subgroup (Fig. 4).

The CPC-like subgroup includes two major clades. Clade II composed of *V. carteri* and *C. reinhardtii* proteins and is sister to Clade I of angiosperm proteins (Fig. 2). Recently, we found that some soybean 2R-MYB genes are alternatively spliced, resulting in a change from 2R-MYB to R3-MYB.[11] Alignment analysis showed that the MYB domains of angiosperm CPC-like proteins had significant homology to the R3 repeats of 2R-MYB proteins. Both sequences contain a conserved motif, [DE]Lx2[RK]x3Lx6Lx3R (Fig. 3), which specifies the interaction with bHLH proteins.[35,36] Phylogenetic tree analysis showed that such R3-MYB proteins clustered within the CPC-like subgroup (data not shown). However, we did not find this motif in algae CPC-like proteins or 2R-MYB proteins. Further phylogenetic analysis revealed that the algae CPC-like proteins and R3 repeats of 2R-MYB proteins also clustered within a clade (data not shown). This result suggests that the algae CPC-like proteins originated from 2R-MYB proteins after the divergence from a common ancestor of land plants. The absence of the [DE]Lx2[RK]x3Lx6Lx3R motif in algae CPC-like and 2R-MYB proteins, as well as in lower land plants, implies that the interaction between MYB and bHLH proteins may be angiosperm specific. The TRF-like subgroup included proteins from angiosperms and moss, but not from lycophytes, suggesting a loss of these genes from lycophytes. Though the subgroup is very small, it is >443 million yrs old.[37]

Our results also showed a gradual increase in the number of MYB-related genes from moss to flowering plants (Fig. 1). This finding suggests the evolutionary diversification of MYB-related proteins through extensive expansion during plant evolution. This expansion appears to have occurred in three important stages. The first stage likely predated the origin of red algae and led to the establishment of CCA1-like/R-R and TBP-like proteins, containing the motifs (SHAQK(Y/F) F and LKDKW(R/K)(N/T), respectively. The second stage may have occurred at the early origin of land plants to establish the diversity of MYB domains, intron patterns, and non-MYB motifs. The third stage may have occurred after the split between gymnosperms and angiosperms, as reflected in the greater size in angiosperms within each subgroup (Fig. 2). Despite several rounds of gene duplications and loss in different plant lineages, these subgroups have remained highly conserved throughout the plant evolution. Chromosomal distribution analysis revealed that MYB-related genes were distributed throughout all corresponding chromosomes in each species (data not shown). Compared with 2R-MYB proteins,[10,11] we detected fewer tandem duplication events in the MYB-related gene family, which suggests that its major expansion is genome-wide duplication.

Furthermore, CCA1-like and TBP-like genes are also present in other eukaryotes, including fungi and metazoans (data not shown). Taken together, our results indicate that CCA1-like and TBP-like genes are much older than previously thought.

3.10. Functional diversity of MYB-related genes

Putative orthologues in each subgroup and/or clade indicate conserved physiological functions. Therefore, we performed a comparative phylogenetic analysis of *Arabidopsis* and other well-known plant MYB-related proteins (Supplementary Fig. S6). Supplementary Table S4 summarizes the functions of plant MYB-related genes.

CCA-1like/R-R genes are best known for their involvement in circadian rhythm regulation, and are highly conserved in plants (Supplementary Table S4), which implies that the functional divergence of clock machinery occurred before the divergence of land plants. Consistently, in the analysis of PLEXdb microarray data,[19] we identified *AtCCA1* gene homologues in maize and soybean that are involved in circadian rhythmicity (Supplementary Fig. S7A). In our study, CCA1-like genes involved in circadian rhythmicity divided into two clades with different intron patterns (c and d). This may explain why they have similar functions but through different mechanisms. Moreover, we observed relatively high expression of MYB-related genes in the flower tissues of maize and soybean, consistent with a role in regulating floral development.[38] Taken together, these results indicate the functional diversity and conservation of CCA1-like/R-R proteins and their crucial roles in plant development.

The cooperative interaction between MYB and bHLH TFs is a classical example of combinatorial regulation. Two types of MYB TFs, 2R-MYB (WER) and CPC-like (or R3-MYB), have similar functions in cell-fate determination.[39,40] Both of which contain the conserved motif DLx2Rx3Lx6Lx3R in their MYB domains, which is involved in MYB−bHLH interactions[35] (Fig. 3). CPC-like proteins can interact with bHLH proteins, thereby competing with the 2R-MYB protein in the regulation of plant development.[41] In the present study, one soybean CPC-like gene, *GmMYBR79*, was expressed in the soybean root hair (Supplementary Fig. S3A), implying a similar role in soybean hair development. Recently, CPC-like genes were shown to down-regulate anthocyanin synthesis by a similar mechanism.[42] In our expression analysis, one maize (*ZmMYBR20*) and two soybean (*GmMYBR78* and *GmMYBR80*) CPC-like genes showed high expression in flower tissues (Fig. 5 and Supplementary Fig. S2), suggesting that they may regulate anthocyanin synthesis via similar mechanisms. These results demonstrate the close relationship between CPC-like proteins and 2R-MYB proteins and the functional conservation of this motif during the evolution.

Members of the I-box-like subgroup are also key developmental regulators in various plant tissues (Supplementary Table S4). Consistently, our results showed that I-box-like genes have broad expression profiles in maize and soybean (Fig. 5 and Supplementary Fig. S2). To date, the most well-known role of I-box-like genes is the regulation of floral asymmetry.[43,44] Interestingly, the I-box-like genes not only showed high homology and similar expression patterns with R-R-like genes, but also appeared to have similar functions in flower development. This further demonstrates that I-box-like genes evolved from R-R-like genes and maintained functions similar to those of R-R-like proteins. A relatively small number of TBP-like genes have been functionally characterized (Supplementary Table S4). Most of the known TBP-like genes encode telomere-binding proteins.[8] The strong divergence of this subgroup implies that its members might have additional diverse functions.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This project was supported by the State Key Basic Research and Development Plan of China (2011Cb100106), the National Natural Science Foundation of China (31071354), and the National Transgenic Program (2011ZX08003001).

References

1. Stracke, R., Werber, M. and Weisshaar, B. 2001, The R2R3−MYB gene family in *Arabidopsis thaliana*, *Curr. Opin. Plant Biol.*, **4**, 447−56.
2. Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A. and Saedler, H. 1987, The regulatory c1 locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators, *EMBO J.*, **6**, 3553−8.
3. Baranowskij, N., Frohberg, C., Prat, S. and Willmitzer, L. 1994, A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator, *EMBO J.*, **13**, 5383−92.
4. Rose, A., Meier, I. and Wienand, U. 1999, The tomato I-box binding factor LeMYBI is a member of a novel class of myb-like proteins, *Plant J.*, **20**, 641−52.
5. Wang, Z.Y. and Tobin, E.M. 1998, Constitutive expression of the CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression, *Cell*, **93**, 1207−17.
6. Schaffer, R., Ramsay, N., Samach, A., et al. 1998, The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering, *Cell*, **93**, 1219−29.
7. Yu, E.Y., Kim, S.E., Kim, J.H., Ko, J.H., Cho, M.H. and Chung, I.K. 2000, Sequence-specific DNA recognition by the myb-like domain of plant telomeric protein RTBP1, *J. Biol. Chem.*, **275**, 24208−14.
8. Marian, C.O., Bordoli, S.J., Goltz, M., et al. 2003, The maize single myb histone 1 gene, Smh1, belongs to a novel gene family and encodes a protein that binds telomere DNA repeats in vitro, *Plant Physiol.*, **133**, 1336−50.
9. Matus, J.T., Aquea, F., and Arce-Johnson, P. 2008, Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes, *BMC Plant Biol.*, **8**, 83−98.
10. Du, H., Fang, B.R., Yang, S.S., Huang, Y.B., and Tang, Y.X. 2012, The R2R3-MYB transcription factor gene family in maize, *PLoS One*, **7**, e37463.
11. Du, H., Yang, S.S., Liang, Z., et al. 2012, Genome-wide analysis of the MYB transcription factor superfamily in soybean, *BMC Plant Biol.*, **12**, 106.
12. Apweiler, R., Attwood, T.K., Bairoch, A., et al. 2001, The InterPro database, an integrated documentation resource for protein families, domains and functional sites, *Nucleic Acids Res.*, **29**, 37−40.
13. Letunic, I., Copley, R.R., Schmidt, S., et al. 2004, SMART 4.0: towards genomic data integration, *Nucleic Acids Res.*, **32**, D142−4.
14. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772−80.
15. Kosakovsky, Pond, S.L., Frost, S.D.W., and Muse, S.V. 2005, HyPhy: hypothesis testing using phylogenies, *Bioinformatics*, **21**, 676−9.
16. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.*, **28**, 2731−9.
17. Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. 2006, MEME: discovering and analyzing DNA and protein sequence motifs, *Nucleic Acids Res.*, **34**, W369−73.
18. Sonnhammer, E.L., Eddy, S.R. and Durbin, R. 1997, Pfam: a comprehensive database of protein domain families based on seed alignments, *Proteins*, **28**, 405−20.
19. Dash, S., Van Hemert, J., Hong, L., Wise, R.P. and Dickerson, J.A. 2012, PLEXdb: gene expression resources for plants and plant pathogens, *Nucleic Acids Res.*, **40**, D1194−201.
20. de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. 2004, Open source clustering software, *Bioinformatics*, **20**, 1453−4.
21. Zhang, H., Jin, J.P., Tang, L., et al. 2011, PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database, *Nucleic Acids Res.*, **39**, D1114−7.
22. Pérez-Rodríguez, P., Riaño-Pachón, D.M., Corrêa, L.G., Rensing, S.A., Kersten, B., and Mueller-Roeber, B. 2010, PlnTFDB: updated content and new features of the plant transcription factor database, *Nucleic Acids Res.*, **38**, D822−7.
23. Zhai, H., Bai, X., Zhu, Y., et al. 2010, A single-repeat R3-MYB transcription factor MYBC1 negatively regulates freezing tolerance in Arabidopsis, *Biochem. Biophys. Res. Commun.*, **39**, 1018−23.
24. Riechmann, J.L. and Ratcliffe, O.J. 2000, A genomic perspective on plant transcription factors, *Curr. Opin. Plant Biol.*, **3**, 423−34.
25. Yanhui, C., Xiaoyuan, Y., Kun, H., et al. 2006, The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family, *Plant Mol. Biol.*, **60**, 107−24.
26. Feldbrügge, M., Sprenger, M., Hahlbrock, K., and Weisshaar, B. 1997, PcMYB1, a novel plant protein

containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit, *Plant J.*, **11**, 1079–93.

27. Farkas, D.H. 1996, *DNA Simplified: the Hitchhiker's Guide to DNA*, AACC Press: Washington, DC.
28. Lupas, A., Van Dyke, M., and Stock, J. 1991, Predicting coiled coils from protein sequences, *Science*, **252**, 1162–4.
29. Sekhon, R.S., Lin, H., Childs, K.L., et al. 2011, Genome-wide atlas of transcription during maize development, *Plant J.*, **66**, 553–63.
30. Severin, A.J., Woody, J.L., Bolon, Y.T., et al. 2010, RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome, *BMC Plant Biol.*, **10**, 160.
31. Libault, M., Farmer, A., Joshi, T., et al. 2010, An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants, *Plant J.*, **63**, 86–99.
32. Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., and Bhattacharya, D. 2004, A molecular timeline for the origin of photosynthetic eukaryotes, *Mol. Biol. Evol.*, **21**, 809–18.
33. Kenrick, P. and Crane, P. 1997, The origin and early evolution of plants on land, *Nature*, **389**, 33–9.
34. Heckman, D.S., Geiser, D.M., Eidell, B.R., Stauffer, R.L., Kardos, N.L., and Hedges, S.B. 2001, Molecular evidence for the early colonization of land by fungi and plants, *Science*, **293**, 1129–33.
35. Zimmermann, I.M., Heim, M.A., Weisshaar, B., and Uhrig, J.F. 2004, Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins, *Plant J.*, **40**, 22–34.
36. Lin-Wang, K., Bolitho, K., Grafton, K., et al. 2010, An R2R3 MYB transcription factor associated with regulation of the anthocyanin biosynthetic pathway in Rosaceae, *BMC Plant Biol.*, **10**, 50.
37. Steemans, P., Herisse, A.L., Melvin, J., et al. 2009, Origin and radiation of the earliest vascular land plants, *Science*, **324**, 353.
38. Fujiwara, S., Wang, L., Han, L., et al. 2008, Post-translational regulation of the Arabidopsis circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins, *J. Biol. Chem.*, **234**, 23073–83.
39. Kirik, V., Simon, M., Hulskamp, M., and Schiefelbein, J. 2004, The ENHANCER OF TRY AND CPC1 (ETC1) gene acts redundantly with TRIPTYCHON and CAPRICE in trichome and root hair cell patterning in Arabidopsis, *Dev. Biol.*, **268**, 506–13.
40. Tominaga, R., Iwatam, M., Sanom, R., Inouem, K., Okadam, K., and Wadam, T. 2008, Arabidopsis CAPRICE-LIKE MYB 3 (CPL3) controls endoreduplication and flowering development in addition to trichome and root hair formation, *Development*, **135**, 1335–45.
41. Pesch, M. and Hülskamp, M. 2011, Role of TRIPTYCHON in trichome patterning in Arabidopsis, *BMC Plant. Biol.*, **11**, 130.
42. Zhu, H.F., Fitzsimmons, K., Khandelwal, A., and Kranz, R.G. 2009, CPC, a single-repeat R3 MYB, is a negative regulator of anthocyanin biosynthesis in Arabidopsis, *Mol. Plant*, **2**, 790–802.
43. Almeida, J., Rocheta, M., and Galego, L. 1997, Genetic control of flower shape in *Antirrhinum majus*, *Development*, **124**, 1387–92.
44. Corley, S.B., Carpenter, R., Copsey, L., and Coen, E. 2005, Floral asymmetry involves an interplay between TCP and MYB transcription factors in *Antirrhinum*, *Proc. Natl Acad. Sci. USA*, **102**, 5068–73.