Research

Open Access

Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of *Arabidopsis* and *Oryza sativa* L. ssp. *indica*

Cizhong Jiang*, Xun Gu*† and Thomas Peterson*

Addresses: *Department of Genetics, Development and Cell Biology, and Department of Agronomy, Iowa State University, Ames, IA 50011, USA. †LHB Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA.

Correspondence: Thomas Peterson. E-mail: thomasp@iastate.edu

Published: 29 June 2004

Genome ***Biology*** 2004, **5:**R46

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/7/R46

Received: 22 December 2003
Revised: 23 March 2004
Accepted: 29 May 2004

Abstract

Background: Myb proteins contain a conserved DNA-binding domain composed of one to four repeat motifs (referred to as R0R1R2R3); each repeat is approximately 50 amino acids in length, with regularly spaced tryptophan residues. Although the Myb proteins comprise one of the largest families of transcription factors in plants, little is known about the functions of most Myb genes. Here we use computational techniques to classify Myb genes on the basis of sequence similarity and gene structure, and to identify possible functional relationships among subgroups of Myb genes from *Arabidopsis* and rice (*Oryza sativa* L. ssp. *indica*).

Results: This study analyzed 130 Myb genes from *Arabidopsis* and 85 from rice. The collected Myb proteins were clustered into subgroups based on sequence similarity and phylogeny. Interestingly, the exon-intron structure differed between subgroups, but was conserved in the same subgroup. Moreover, the Myb domains contained a significant excess of phase 1 and 2 introns, as well as an excess of nonsymmetric exons. Conserved motifs were detected in carboxy-terminal coding regions of Myb genes within subgroups. In contrast, no common regulatory motifs were identified in the noncoding regions. Additionally, some Myb genes with similar functions were clustered in the same subgroups.

Conclusions: The distribution of introns in the phylogenetic tree suggests that Myb domains originally were compact in size; introns were inserted and the splicing sites conserved during evolution. Conserved motifs identified in the carboxy-terminal regions are specific for Myb genes, and the identified Myb gene subgroups may reflect functional conservation.

Background

Regulation of gene expression at the level of transcription controls many important biological processes in a cell or organism. The process of transcription recruits a number of different transcription factors, which can be activators, repressors, or both [1]. Genome-wide comparisons have revealed the diversity in the regulation of transcription during evolution. With the completion of *Arabidopsis* genome sequencing, 5% of its genome was found to encode more than 1,500 transcription factors [2]. On the basis of sequence similarities, transcription factors have been classified into

families. In plants, Myb factors comprise one of the largest of these families.

Myb proteins are defined by a highly conserved DNA-binding domain (termed the Myb domain) composed of one to four helix-turn-helix motifs, which exist as tandem repeats (referred to as R0R1R2R3) in a single Myb protein. Each repeat is about 50 amino acids long, with regularly spaced tryptophan residues, and forms three α-helices The third α-helix has a recognition role during DNA binding [3]. The three-dimensional structure of the Myb domain in the Protein Data Bank (PDB) shows that the DNA recognition α-helix interacts with the DNA major groove. Moreover, previous research indicated that five amino-acid residues in the helix-turn-helix motif bind directly to the major groove [4]. It should be noted that sequences outside the Myb domain are highly divergent.

The first Myb gene found was the v-*Myb* oncogene from the avian myeloblastosis virus [5]. Subsequently, members of the Myb gene family were identified in diverse plants and animals [6,7]. Previous research showed that animal genomes encode relatively few Myb genes [7]. In contrast, flowering plants contain large numbers of Myb genes with very diverse structures and functions [6]. To date, the precise functions of most plant Myb genes are unknown, although some well studied examples suggest important roles for Myb genes in regulation of secondary metabolism, cellular morphogenesis, pathogen resistance, and responses to growth regulators and stress [6,8].

With the completion of *Arabidopsis* and rice (*Oryza sativa* L. ssp. *indica*) genome sequencing [9], the entire complement of Myb genes can be identified and described. However, a great deal of experimental work is required to determine the specific biological function of each gene. In *Arabidopsis*, R2R3 Myb gene-expression levels were determined in more than 20 different growth conditions; the results indicated that Myb genes were specifically expressed in different tissues and physiological conditions [10]. To obtain further functional information on *Arabidopsis* Myb genes, a process of reverse genetics was applied to isolate insertion mutants. In all, 47 insertion mutants were detected in 36 distinct Myb genes by screening a total of 73 genes. However, none of the insertions gave rise to morphological phenotypes visible in soil-grown plants [11]. The redundancy of Myb genes may diminish the efficiency of the molecular approach by complementation of function. No similar research has been done in rice Myb genes. Here, we have used phylogenetic and computational methods to classify Myb genes in subgroups. The resulting subgroup classification and putative functional conserved motif identification may be useful for research on agronomic traits in rice, which is the most important crop for human consumption, and an important model for other cereal grains.

Results and discussion

Expansion of Myb genes in *Arabidopsis* and rice

The Myb gene family has broadly expanded in plants during evolution. The amplification of the Myb gene family occurred before the divergence of monocots and dicots [12]. In our study, 130 Myb genes were found in the *Arabidopsis* genome and 85 in *Oryza sativa* L. ssp. *indica*. The large size of this gene family was also confirmed in *Zea mays* and sorghum [12]. Although most plant Myb genes contain only two repeats, there have been three-repeat Mybs reported in *Arabidopsis* [1], maize [13] and other plants [12]. To date, only three-repeat Myb genes have been detected in animals, and it has been proposed that two-repeat Myb genes died out in the animal lineage [12]. The broad presence of three-repeat Myb genes in diverse species indicates the antiquity of these genes. Using homology search in the GenBank non-redundant database, two three-repeat Myb proteins (accession numbers NP_913483 and BAC79618.) were identified in *Oryza sativa* L. ssp. *japonica*. However, no three-repeat Mybs were detected in rice (*indica*) in our study. This could be due to the incompleteness of the *Oryza indica* dataset.

Figure 1 *(see following page)*
Phylogeny, subgroup designations, and carboxy-terminal motifs in Myb proteins from *Arabidopsis* and rice. The phylogenetic tree on the left represents 130 Myb genes from *Arabidopsis*, 85 from rice, and 43 from other plants, which are clustered into 42 subgroups (triangles) and seven singletons (lines). The 19 gray subgroups contain conserved carboxy-terminal motifs. The arrow indicates a large cluster of genes involved in the phenylpropanoid biosynthetic pathway or ABA response. The scale bar under the tree respresents 0.2 substitutions. Some 'landmark' Myb proteins are listed in parentheses for functional reference. The uncompressed tree with full taxa names is available as Additional data file 7. Comparison of the subgroup designations used in this study with that in [1] is described in Additional data file 1. The four blocks (A-D) in the center of the diagram indicate the distribution of the four major splicing patterns in the Myb R2R3 domains; see text for details. The motifs on the right were detected using MEME and drawn to scale. The Myb R2 and R3 repeats are indicated. The black boxes indicate the extension motifs following the R3 repeat. The gray boxes represent the motifs identified in the previous report [1], and the white boxes are the motifs newly discovered here. The thin lines indicate coding regions lacking a detectable motif, with a polypeptide length indicated by the number above the diagonal slash marks. The scale bar is equivalent to 50 amino-acid residues.

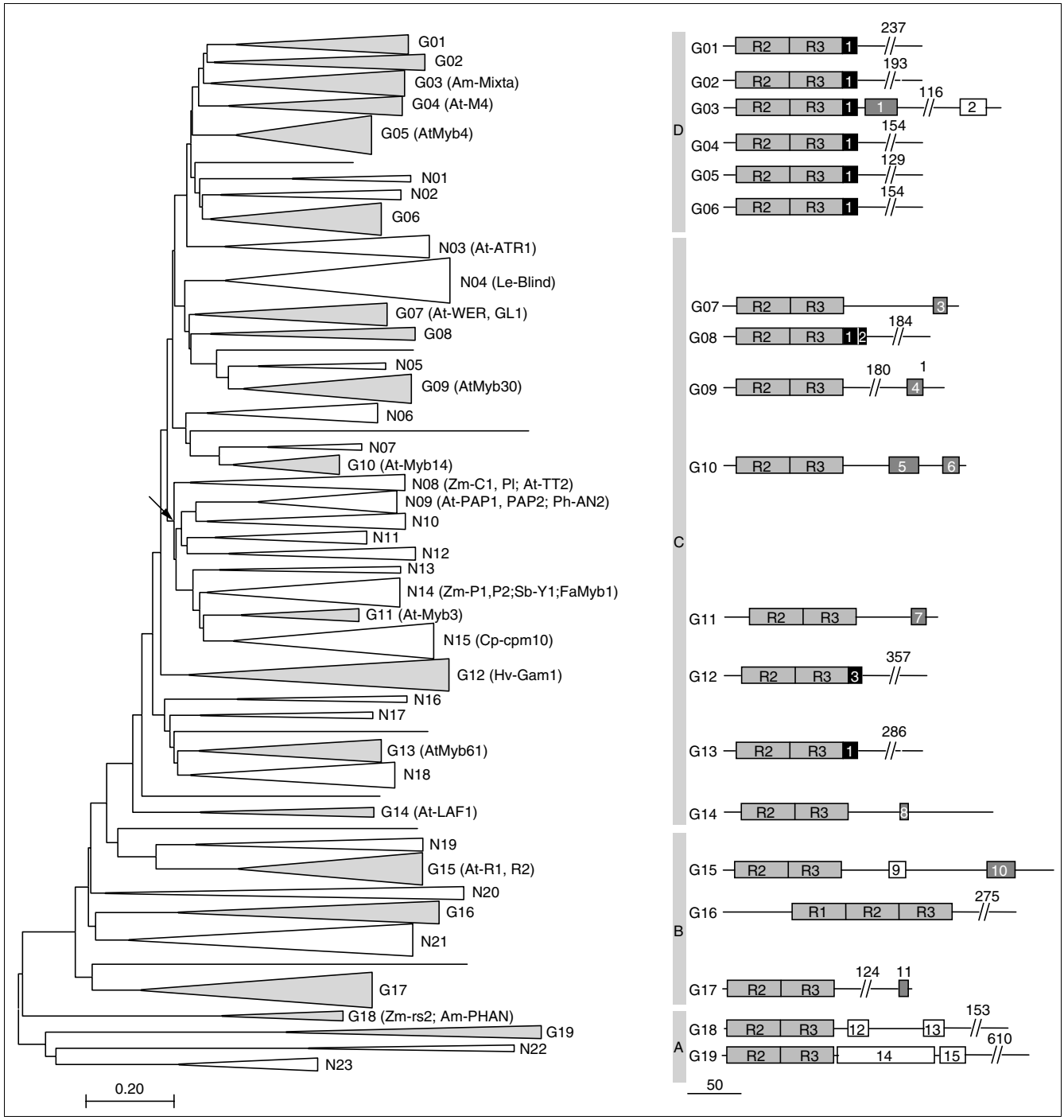

Figure 1 *(see legend on previous page)*

Topology of Myb gene phylogeny

On the basis of sequence similarity and the topology of the phylogeny, we clustered the Myb genes into 42 subgroups, ranging in size from two to 14 Myb genes (Figure 1). The phylogenetic topology and subgroup structures are consistent with previous reports [1,10]. The detailed comparison is described in Additional data file 1 (also available with all other supplementary material at [14]). However, because of the large number of taxa, the bootstrap values are low (data not shown). Therefore, we sought other evidence to support the reliability of the subgroup designations.

Interestingly, AtMyb33, 65, 101, 104 and At3g60460 were complementary, with few mismatches, to *Arabidopsis* Myb

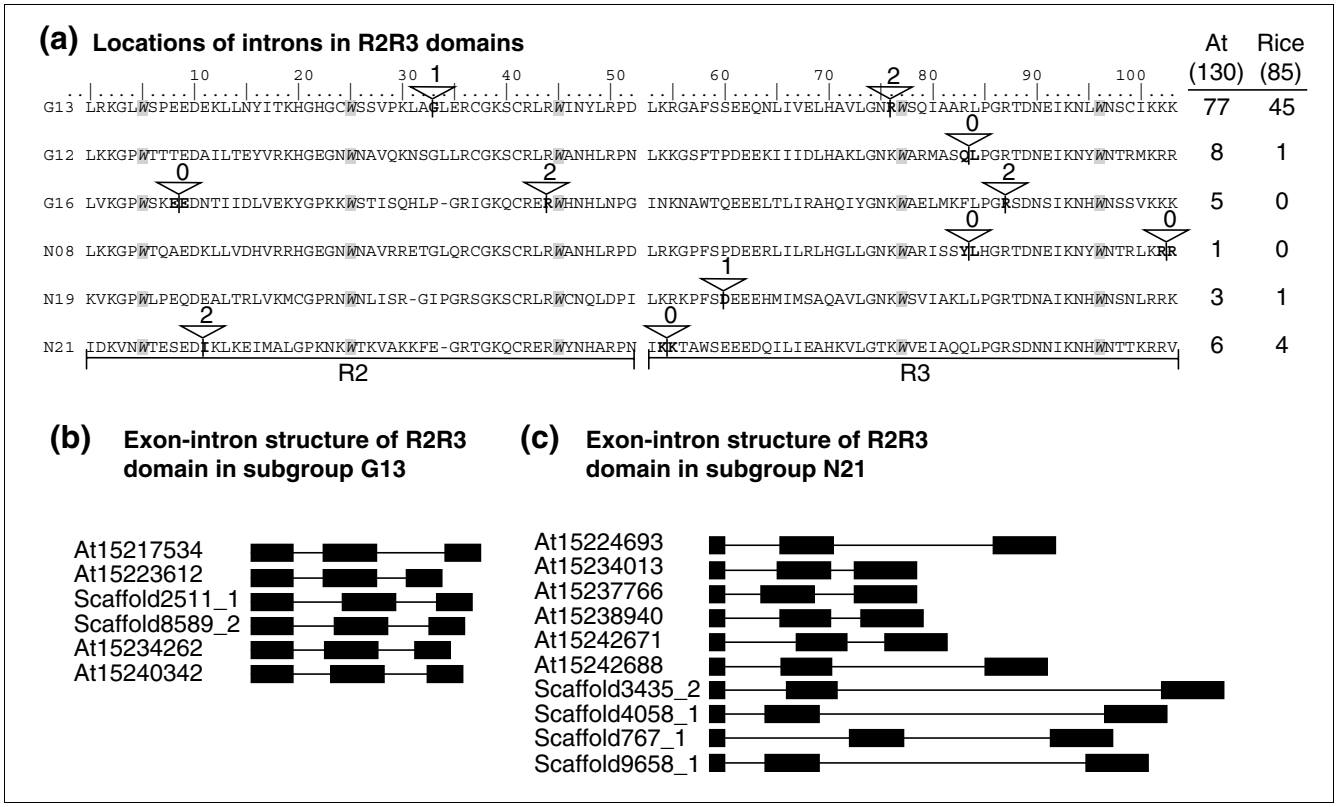

Figure 2

Intron-exon structure of Myb genes. **(a)** Locations of introns 1 and 2 splice sites in R2R3 domains. Six representative Myb R2R3 domain sequences are shown. The extent of the R2 and R3 repeats is indicated at bottom of the alignment. The triangles indicate the positions of the splice sites, and the numbers above the triangles indicate the phases of introns. Subgroup G13 represents the major splicing pattern; that is, 77 (of 130) Myb genes in *Arabidopsis*, and 45 (of 85) Myb genes in rice have this splicing pattern. The shaded W residues indicate the regularly spaced tryptophan residues. The representative sequences of the six subgroups are: G13, At1g57560; G12, At2g32460; G16, At4g32730; N08, Scaffold479_5; N19, At2g39880; N21, At2g25230. The table at the right of the alignment lists the number of Myb genes with each splicing pattern and the total number of Myb genes in *Arabidopsis* and rice. Note that 22 *Arabidopsis* and 18 rice genes have the typical G13 splicing pattern, except that they lack either intron 1 or intron 2. Additionally, six *Arabidopsis* and 12 rice genes have no introns within the R2R3 domain. Finally, two *Arabidopsis* and four rice Myb genes have other atypical splicing patterns (data not shown). **(b,c)** The conserved exon-intron structure of all member genes in subgroups G13 and N21. Boxes and lines indicate exons and introns, respectively. Additional examples are provided in Additional data file 8.

microRNA (noncoding RNA) miR159 [15]. The sequence is 21 nucleotides long and located in the 3' untranslated region (3' UTR) of all five genes. MicroRNAs are proposed to act as regulators of gene expression through interactions with complementary mRNA sequences. Importantly, these five *Arabidopsis* Mybs are located in subgroup G12 (Figure 1). This clustering provides additional evidence for the reliability of the subgroup designations in our analysis.

Conserved gene structure within each subgroup supports the subgroup designations

The phylogenetic topology and subgroup structures are based on sequence comparisons of the complete predicted Myb genes. To test the reliability of the subgroup designations using independent criteria, we investigated the exon-intron structure of Myb genes subgroup by subgroup. A majority of *Arabidopsis* (59%) and rice (53%) Myb genes have a conserved splicing pattern of three exons and two introns in R2R3 domains (represented by subgroup G13; Figure 2a). Either or both of the two introns are absent in 19% of *Arabidopsis* Myb genes and 12% of rice Myb genes. Variable splicing patterns different from G13 were detected in 22% of *Arabidopsis* and 35% of rice Myb genes, respectively (data not shown). Strikingly, the exon-intron structure is conserved within each subgroup, but varies between subgroups (Figures 2b,c). This supports the subgroup designations from the independent criterion of splicing pattern.

Interestingly, the Myb gene splicing patterns constitute four major blocks in the Myb gene phylogeny (Figure 1). Block A lacks both introns 1 and 2. There are three splicing patterns in block B: subgroup G15 lacks both introns; subgroup G17 lacks only intron 2; and the remaining genes have altered splicing sites when compared to subgroup G13. Myb genes in block C have the major splicing pattern (81.2%) typified by G13, with some individual genes lacking intron 1 (9.4%), intron 2 (4.7%)

or both introns (1.9%), or having minor splicing patterns (2.8%). In contrast, 58.2% of Myb genes in block D retain the typical splicing sites, and the rest lack only intron 1 (G02, G05 and half of the genes in G06).

In addition to splice-site locations, we also examined the position of splicing with respect to the open reading frame (ORF) - the intron phase. The splicing of each intron is designated as occurring in one of three phases: in phase 0, splicing occurs after the third nucleotide of the first codon; in phase 1, splicing occurs after the first nucleotide of the single codon; and in phase 2, splicing occurs after the second nucleotide. Figure 2a shows not only the conserved locations in the Myb-domain protein sequences but also the conserved phases of introns within the same subgroup. Moreover, there is a significant excess of phase 1 and 2 introns as well as an excess of nonsymmetric exons in Myb genes. Symmetric exons are exons that are flanked by introns of the same phase.

According to the intron-early theory [16], an excess of phase 0 introns and symmetric exons may facilitate exon shuffling by avoiding interruptions of the ORF, and thus could accelerate the rate of recombinational fusion and exchange of protein domains. Our results suggest that ancient Myb genes had a compact size without introns. During evolution, under some unknown mechanisms, introns were inserted into Myb domains and resulted in the observed splicing patterns. One splicing pattern remained unchanged in the subsequent gene amplification, resulting in the major splicing pattern typified by G13. Consistent with this, transposition of introns occurs very infrequently during evolution [17]. This intron-gain model is consistent with previous results showing that numerous introns have been inserted into plants and retained in the genome [18]. A similar approach to gene classification using intron/exon structure has been applied in the kinesin family [19] and the bHLH family [20], and the results support a similar evolutionary pattern.

Although the splicing sites are conserved, the sizes of both introns vary greatly for different Myb genes. Approximately 85% of introns 1 and 2 of Myb genes is shorter than 300 bp in *Arabidopsis* and rice. Detailed information about the distribution of intron sizes of Myb genes is available in Additional data file 2. It is worth noting that the size of intron 2 of maize *p1* and *p2* orthologs is very large, around 5 kb. This intron-size information may be helpful for aligning expressed sequence tags (ESTs) with genomic sequences.

Strikingly, a 743-base fragment was found in intron 2 of maize *P1-rr* and *P1-wr* alleles, but not in *P1-rw* and *p2* alleles. A 10-base direct repeat (5'-TGATTTTGAC-3') flanks this fragment. Interestingly, no *Ac* elements were found inserted in its adjacent 3.2-kb intronic region, but frequent *Ac* insertion occurred in other regions. This could be due to a particular chromatin structure refractory to *Ac* insertion in this region [21]. BLAST search detected this fragment (94% identity over 723 base-pairs (bp)) at one other locus in the maize genome, but with a new flanking direct repeat (5'-GGATATCCA-3'). The GenBank accession number is AF466202 (located 84795..85689, 12 March 2002 version). These results are consistent with a previous proposal that some transposable elements could insert into the genome as intronic sequences, a mechanism that has been proposed for the insertion of nuclear introns [22].

Topology of Myb gene phylogeny may reflect functional conservation

Most plant Myb genes are thought to encode transcription factors that activate or repress target gene expression either independently or together with cofactors. The topology of Myb phylogeny (Figure 1) indicates that some Myb genes in the same subgroup have the same function and that some Myb genes with similar functions are located in the same subgroup. For example, two Myb orthologs, snapdragon *PHAN* gene and maize *rs2 gene*, are located in subgroup G18, and both are involved in organ development: *PHAN* has been shown to regulate the development of the proximo-distal axis and dorso-ventral asymmetry of lateral organs such as leaves, bracts and petal lobes [23], while the *rs2* gene controls the development of maize lateral organ primordia by repressing expression of *knox* (*knotted1*-like homeobox) genes that are required for the normal initiation and development of lateral organs [24]. In another example, the *Arabidopsis* genes *GL1* and *WER* located in subgroup G07 are both involved in epidermal cell development: *GL1* activates the *GLABRA2* homeobox gene for trichome (hair cell) development in some parts of the leaf and in the stem [25,26], while *WER* controls the formation of the root epidermis by regulating expression of the *GLABRA2* gene [27].

Similar results are observed for Myb genes involved in the phenylpropanoid biosynthetic pathway (Figure 1, subgroups N08, N09, N14): *C1* [28], *Pl*, *TT2* [29], *AN2* [30], *p1* [31], *p2* [32], *FaMyb1* [33], *PAP1* and *PAP2* [34]. These genes all encode a transcription factor that activates enzymes for phenylpropanoid synthesis, except that the *FaMyb1* transcription factor suppresses anthocyanin and flavonol accumulation [33]. In addition, the functional conservation among some Myb genes during evolution could be observed in the cell-cycle protein *CDC5* (Figure 1, G19). The CDC5 protein performs an essential function in cell-cycle control at G2/M, and also participates in pre-mRNA splicing [35].

Carboxy-terminal motifs

The extent of the Myb R1, R2 and R3 repeats is based on similarity to the previously-published consensus Myb repeat sequences [36]. We used computational methods to identify additional conserved sequences downstream of the Myb repeats. A total of 18 motifs were identified in the carboxy-terminal regions, with each motif ranging in size from 9 to 32 amino acids (Figure 1). An exceptionally large domain (91 residues) was found in subgroup G19, gene *CDC5*, which is a con-

Table 1

Consensus sequences of carboxy-terminal motifs

Motif	Alias	E-value	Consensus sequences
E1	24	8.3e-081	LxxMGIDPVTH[KR]P
E2		2.4e-058	FSHLMAEI
E3	18	5.4e-062	QRAGLPLYPpE[IV]
M1	9	4.1e-073	Gq[SA]KnAAxLSH[MT]AQWESARLEAEARLARESKL
M2		7.8e-039	exe[DE]NKNYWNSI[LF]NIV[ND]SSpSdSs
M3	15	1.7e-041	WV[HL][ED]D[DE]FELS[ST]L[TV][MN]M
M4	1.2	3.9e-032	QGsLSL[IF]EKWLFd[DE]Q[SG]
M5	2	2.3e-025	DISNsNKDsatSsEDvlAilDeSFWSeVv
M6	2	4.3e-033	drNdKgYNhDMEFWFD
M7	19	2.2e-014	DQ[ST]gENYWg[MV]DD[IL]W[PS]
M8	16	1.5e-013	PxLfFSEWI
M9		4.1e-031	PGSP[ST]GSD[VR]SD[SL]S[HT][GI]
M10	22.2	1.0e-120	GEFM[AT][VA][VM]QEMI[KR][AT]EVRSYMAe[MV][QG]xx[NA]G[GC]G
M11	21	1.2e-047	[PV]pF[FI]DFLGVG
M12		1.5e-150	Pixx[GS][KR]Y[DE][HW][IL]LExFAEKLVKERP
M13		5.4e-112	SPSVTLSL[SA][PS][SA][TA]VA[PA]aP[PA]aP
M14		3.4e-079	YDa[AN]DdPRkLRPGEIDPNPEaKPARPDPVDMDEDEKEMLSEARARLANTrGKKAKRKAREK QLEeARRLAsLQKRRELKAAGIdgrhrKRK
M15		5.3e-020	IDYNAEIPFEK[KR][AP]paGFYDTaDEDRp[AN]D

Alias indicates the corresponding motifs identified by Stracke *et al.* [1]. E-value was calculated by MEME. Consensus sequences follow the criteria of Joshi *et al.* [44]: a single capital letter is given if the relative frequency of a single residue at a certain position is greater than 50% and greater than twice that of the second most frequent residue. When no single residue satisfied these criteria, a pair of residues was assigned as capital letters in brackets if the sum of their relative frequencies exceeded 75%. If neither of these two criteria was fulfilled, a lower-case letter was given if the relative frequency of a residue is greater than 40%. Otherwise, x is given.

served Myb paralog that originated prior to Myb-family amplification [12]. In addition, Myb genes maize *C1*, *Pl* and AtMyb123 (*TT2*) in subgroup N08 have a nine-amino-acid motif previously reported [1,37]. This motif has a high e-value (4.5e-008), so it was excluded from our analysis. Three other motifs identified by Stracke *et al.* [1] were excluded from our analysis because of their high e-values (see Additional data file 3).

Interestingly, three short fragments directly following the Myb R3 repeat are highly conserved in some subgroups (Figure 1, black boxes). We designated these extension motifs, E1, E2 and E3. Subgroups with an extension motif contain few or zero motifs in their carboxy-terminal coding regions when compared to those subgroups without extension motifs. One exception is subgroup G03, which contains motif E1 and two other carboxy-terminal motifs - 1 and 2 (Figure 1). In subgroup G08, a short conserved segment following E1 is termed E2. The three extension motifs are relatively small, ranging from 8 to 13 residues, but they are much more conserved than other motifs (Table 1). In the group of three extension motifs, 28 (of 33) sites are occupied by a single residue in more than 50% of the Myb proteins, and this value is greater than twice the relative frequency of the second most frequent residue.

To test the reliability of the motif predictions, the similarity scores were calculated over the motif plus its flanking regions. The similarity plots produced much higher scores in the motif region than in the flanking regions (Figure 3a), thus supporting the identified motifs. Similar results were observed in the nonsynonymous (dN) substitution analysis, which is a typical way of examining the degree of functional constraints on proteins using evolutionary comparisons [37]. The results indicated that motif regions were less frequently subject to substitution than flanking regions. The distribution of dN values showed that most dN values are equal to or less than 0.6 in motif regions, and greater than 1 in flanking regions (Figure 3, Table 2). Interestingly, there are seven other motifs identified by Stracke *et al.* [1] that had a high dN value and did not pass this test (see Additional data file 3). The presence of carboxy-terminal motifs could reflect either the long-term conservation of critical sequences from antiquity or more recent gene duplications. The low dN values in the motif regions compared with the flanking regions suggest that the motifs are ancient sequences which have been conserved over

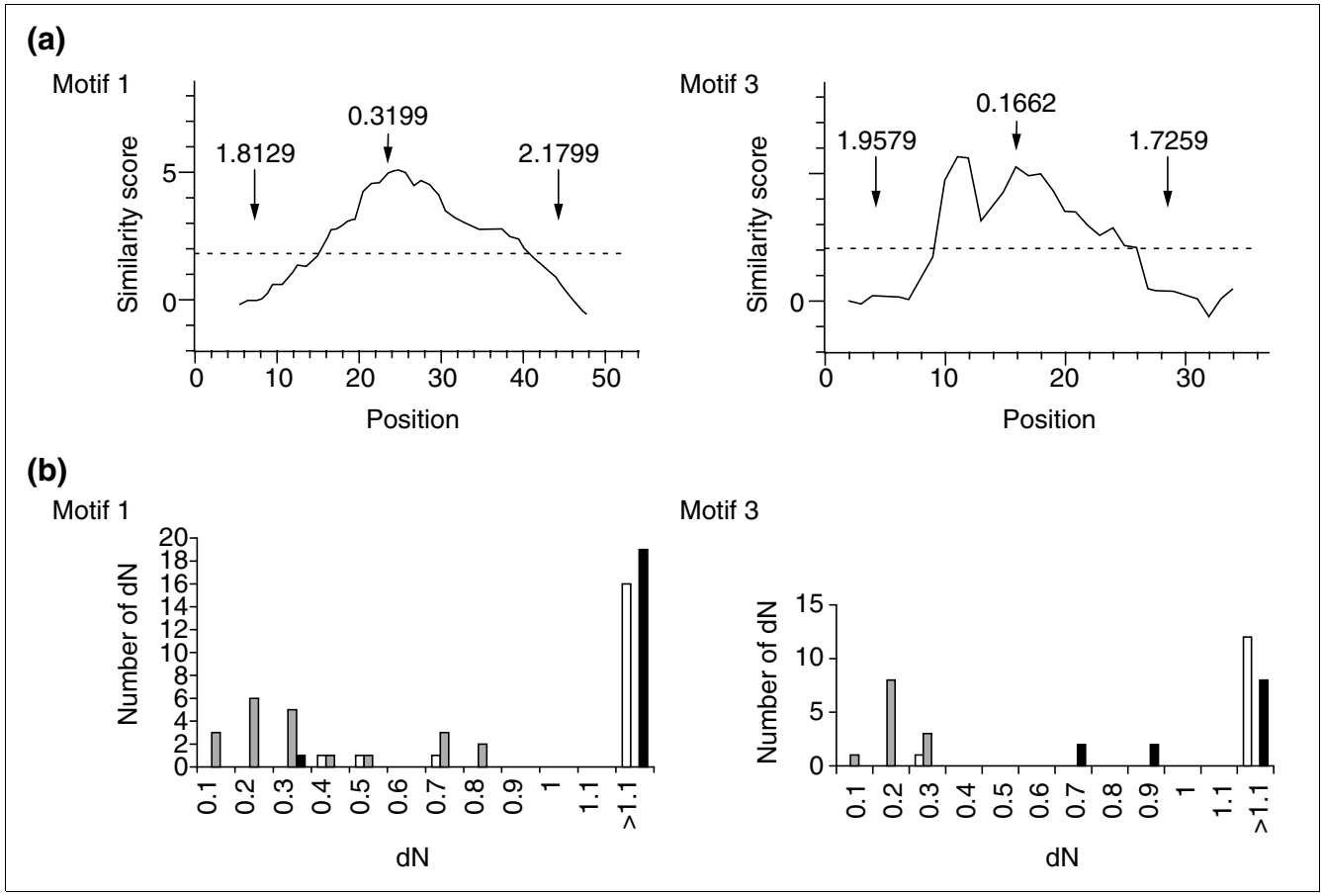

Figure 3
Similarity scores and dN values. **(a)** Similarity scores and the average dN values of motifs 1 and 3 plus 10-residue flanking fragments. The curve indicates the similarity scores along the upstream flanking region, the motif region (peak), and the downstream flanking region. The dashed line shows the average similarity value for the entire alignment. The vertical axis is the score obtained from the scoring matrix BLOSUM62, with scores ranging from -4 to 11. The horizontal axis indicates the position of the alignment. The three values indicate the nonsynonymous substitution (dN) values in the upstream flanking region, the motif region (peak), and the downstream flanking region, respectively. Diagrams for other motifs are given in Additional data file 9. **(b)** Distribution of pairwise dN values of motifs 1 and 3. Most dN values of motifs are equal to or less than 0.6. In contrast, the counterparts of flanking regions are equal to or greater than 1. This result indicates that sites in motif regions are highly conserved and less frequently subject to substitution than those in flanking regions. White boxes, amino-terminal regions; gray boxes, motif regions; black boxes, carboxy-terminal regions. Histograms for other motifs are given in Additional data file 10.

long periods of time, rather than being the result of more recent duplications.

Specificity of motifs to Myb genes

We wished to determine whether the detected motifs are specifically present in Myb genes, but are absent in non-Myb genes. Therefore, we used protein motif sequences as query sequences and performed Blastp searches in the Swiss-Prot database. For the motifs with size equal to or less than 15 amino acids, the homologous hits with 85% of the query motif length are all Myb domain containing proteins. For the motifs with size greater than 15 amino acids, the corresponding homologous hits with 70% of the query motif length contain Myb domains. The search result is described in Additional data file 4.

We obtained similar results in EST searches. When translated into proteins, all the 14 ESTs detected from an extension motif search also contain a Myb domain. Interestingly, we detected more ESTs from E1 than from the other extension motifs. Most probably this is due to the presence of E1 in more Myb genes than other motifs (Figure 1). The search result is described in Additional data file 5.

After comparing these two search results, we found that not all carboxy-terminal motifs detected homologous ESTs. This could be due to the low levels of expression of some Myb genes so that their EST sequences are not yet available. In some cases, these ESTs did not contain Myb domains; however, because the carboxy-terminal motifs are located downstream some distance from the Myb domains, the returned

Table 2

The average dN values of carboxy-terminal motifs and their flanking regions

Motif	Amino-terminal	Motif	Carboxy-terminal
E1	0.2516	0.3186	2.0265
E2	0.1624	0.4690	1.4686
E3	0.2309	0.3326	1.9216
M1	1.8129	0.3199	2.1799
M2	1.9477	0.5149	1.9044
M3	1.9579	0.1662	1.7259
M4	1.7544	0.5062	2.1131
M5	2.3692	0.3601	2.2258
M6	2.2981	0.4278	
M7	1.9431	0.3928	1.6694
M8	2.2754	0.2326	1.6436
M9	1.8157	0.5401	2.2248
M10	2.1319	0.3695	1.8878
M11	2.1358	0.3209	
M12	1.5835	0.3525	1.6608
M13	1.5503	0.4099	1.9772
M14	1.9661	0.2701	1.2742
M15	1.4746	0.4667	1.9047

Amino-terminal and carboxy-terminal indicate the upstream and downstream flanking regions, respectively. The upstream flanking regions of extension motifs E1, E2 and E3 are the carboxy-terminal ending fragments of the Myb R3 repeat, which is highly conserved. Therefore, their amino-terminal dN values are low. dN values for the carboxy-terminal motifs M6 and M11 could not be calculated because of their close proximity to the carboxyl terminus.

ESTs are probably too short to reach the Myb domains. However, alignment of ESTs with known Myb genes showed high identity not only in the motif sequence but also for considerable lengths in the flanking regions. This suggests that such ESTs are very likely from Myb genes.

Interestingly, we checked each carboxy-terminal motif in the 258 Myb proteins and found they are subgroup specific. For example, motif 1 from subgroup G03 was not detected in other subgroups.

Identification of regulatory elements in noncoding regions

In addition to the carboxy-terminal motifs detected in the predicted Myb proteins, we wanted to test whether any conserved DNA sequence motifs could be identified among the Myb gene subgroups. We applied motif-searching tools to detect conserved regulatory elements in the promoter region plus 5' UTR of the Myb genes and in intron regions. In contrast to the carboxy-terminal coding regions, no conserved DNA sequence motifs were identified in the Myb gene noncoding regions. This could be due to the fact that the Myb genes clustered in each subgroup are probably not orthologs or paralogs. In contrast, within the subgroup N14 (Figure 1) containing the maize *p1* and *p2* genes, and orthologs/paralogs from sorghum and rice, a highly conserved scheme of TATA-box, transcription start site sequences, and 5' UTR CA-box were found (data not shown). Otherwise, no significant regulatory elements were detected in noncoding regions of other Myb genes. However, it should be noted that segments of intron sequence closer to flanking exons are significantly more conserved than interior intron sequence. It has been reported that this level of intron sequence conservation may have a functional role in gene regulation [38]. Our results suggest that it will be difficult to directly identify regulatory motifs in noncoding regions using only existing computational techniques. The chance of identification of regulatory elements will be increased in orthologs/paralogs. Possibly, the identification of co-regulated genes using microarray analysis will assist in the identification of common regulatory elements.

Conclusions

The expansion of Myb genes in plants makes it one of the largest families of transcription factors known to date. However, the specific roles of Myb genes in regulating plant traits are still unclear. Here, we used overall sequence similarity to cluster Myb genes from *Arabidopsis* and rice into 42 subgroups. The subgroup designations were well supported by sequence similarity and exon-intron structure. In one subgroup, significant complementarity to a specific miRNA was also observed. Furthermore, we found that the splicing sites and the phase of introns are conserved in Myb domains within the same subgroup, but differ between subgroups. The phylogenetic topology of splicing patterns suggested that Myb domains may originally have been compact in size, and that introns were inserted and remained in place during evolution. Computational searches were used to identify conserved carboxy-terminal motifs present in the different subgroups. These motifs appear to be specific characteristics of the Myb subgroups. In contrast to the carboxy-terminal motifs specifically present in Myb genes, no conserved regulatory elements were identified in the noncoding regions.

Materials and methods

Myb proteins used in the analysis

At the initiation of our project, the international rice genome sequencing project was not finished. Two finished rice genomes - one by Monsanto, the other by Syngenta (*Oryza sativa* L. ssp. *japonica*) - were not available to the public. Only the rice (*indica*) genome sequenced by Beijing Genomics Institute was publicly available. The sequence-quality assessments through sequence-tagged site (STS) markers, UniGene clusters and nonredundant cDNAs showed that 92% of the functional sequences that encode genes, and their

immediate regulatory elements were present in the assembled sequences [9]. Therefore, we chose subspecies *indica* rather than *japonica* for this study.

Rice genome sequences (scaffold dataset) were obtained from Beijing Genomics Institute [39]. FGeneSH has been used successfully to predict genes in rice [9], and GenScan was used together with it to predict genes by taking rice genomic sequences as input. The two prediction results were combined as the complement of rice proteins. We performed Blastp and HMMER [40] searches to identify Myb genes from this rice protein dataset. For Blastp, we used a set of Myb R2R3 domains as query sequences. For HMMER, we used the Myb profile from Pfam. We parsed and combined the results of both searches, and obtained the final complement of rice Myb proteins with manual inspection of each sequence to confirm the identification of *bona fide* Myb genes. In the end, 85 typical Myb genes with complete R2R3 domains (one R0R1R2R3 and 84 R2R3) and 28 partial Myb genes were detected in the rice genome. Partial Myb genes contain a segment similar to one or a partial Myb repeat. The sequences of rice Myb genes are listed in Additional data file 6.

The complement of *Arabidopsis* proteins from GenBank were used to identify Myb proteins with complete R2R3 domains. The same methods as above were applied. We obtained 130 typical Myb proteins containing complete R2R3 domains (one R0R1R2R3 Myb, five R1R2R3 proteins, 124 R2R3 protein) and 11 partial Myb proteins. The results are consistent with previous findings [1].

To collect reference information on Myb gene functions, we used Blastp search against the nonredundant dataset in GenBank. The search yielded 43 plant Myb proteins with complete R2R3 domains; for most of these, some experimental information regarding functions or expression patterns was deposited by individual researchers.

Construction of phylogeny and subgroup designations

For phylogenetic analysis, the above 258 Myb proteins (130 *Arabidopsis*, 85 rice and 43 from various other plants) with complete R2R3 domains were included. The sequences were aligned by ClustalX (version 1.81). The phylogenetic tree was constructed by the neighbor-joining method using MEGA version 2.0 [41], with the setting of pairwise gap deletion and Poisson distance. Bootstrapping (1,000 replicates) was performed to evaluate the degree of support for a particular grouping in the neighbor-joining analysis. To enable the identification of motifs in the carboxy-terminal regions within each subgroup, we did not employ complete gap deletion as this may tend to exclude the contribution of carboxy-terminal residues because of their high divergence. The p-distance represents the simplest sort of genetic distance calculation and can be highly biased, so it was not used. In addition, attempts to use only the carboxy-terminal regions in construction of phylogeny were negative as a result of the high divergence. Therefore, we used the complete Myb proteins in clustering.

Three trees were constructed with the above settings, then taxa were classified into subgroups based on the topology of the phylogeny. Tree I used the 43 landmark Myb proteins with 130 *Arabidopsis* Myb proteins. The clustering result is consistent with the previous report [1]; that is, the taxa that were clustered as subgroups in Stracke *et al.*'s findings [1] are located within a subgroup in tree I. Tree II replaced the 130 *Arabidopsis* Myb proteins in tree I with 85 rice Myb proteins. Tree III used the total 258 Myb proteins. We found that the clustering result of Myb proteins from *Arabidopsis*, rice and the landmark Myb proteins was consistent among all three trees. Therefore, we used tree III as the representative in this study.

Motif identification

Within each subgroup, motifs were detected using MEME [42] with the following parameter settings: the distribution of motifs: zero or one per sequence; maximum number of motifs to find: 16; minimum width of motif: 6; maximum width of motif: 117, in order to identify long R2R3 domains; minimum number of sites for each motif: the number of sequences, i.e., the motif must be present in all members within the same subgroup. Other options used the default values. Only motifs with e-value <= 1e-10 were kept for further analysis.

Motif analysis: similarity scores and nonsynonymous (dN) substitution

To confirm the reliability of the 38 motif candidates identified by MEME, we used PlotSimilarity from GCG package from Genetics Computer Group, Inc. to calculate the similarity score of each motif plus its 10-residue flanking fragments (protein sequences). There were 33 motifs with values above the average score in the motif region and below the average score in the flanking regions, and these were tested further using the dN values. The program YN00 from PAML package [43] was applied to analyze the conservation of each motif plus its flanking regions (coding DNA sequences). The frequency of synonymous substitution is too high to detect the conservation. Therefore, nonsynonymous substitution value was calculated. Low dN values indicate conservation whereas high dN values indicate divergence. We detected 18 motifs with dN <0.5 in the motif region and >1 in flanking regions; these 18 motifs included 3 extension and 15 carboxy-terminal motifs.

Originally, MEME identified 38 motif candidates with e-value <= 1e-10. Then five motifs were removed in the similarity score test. Later, 15 motifs were discarded in the nonsynonymous substitution test. This result suggests that the similarity score test is not sufficiently powerful to determine the reliability of motif candidates, and may be safely ignored in the future.

Specificity of motifs

To test whether the carboxy-terminal motifs are specific to Myb genes, motif sequences were used to perform homology search in Swiss-Prot database and EST data set from GenBank. The latter can also provide information on the expression pattern of Myb genes. Low complexity was turned off for optimal short sequence search in both homology searches. In addition, in EST search, for motifs less than 15 residues 10 downstream residues were appended, and this elongated sequence was used as query sequence to perform EST search. The corresponding Myb R2 repeats were used in a tblastn EST search as an internal positive control.

Additional data files

The following additional files are available: Additional data file 1 gives the mapping relations of subgroups; Additional data file 2 gives the distribution of intron sizes in Myb genes; Additional data file 3 gives the previously identified carboxy-terminal motifs not included in this study; Additional data file 4 gives the Blastp search results for homologous Myb genes; Additional data file 5 gives the homologous EST search results; Additional data file 6 (a .FAS file) gives the sequences of all rice Myb genes; Additional data file 7 is a tree of the relationship of 130 *Arabidopsis*, 85 rice Myb proteins and 43 'landmark' Myb proteins; Additional data file 8 gives the intron-exon structure of all R2R3 domains; Additional data file 9 gives similarity scores and average dN values for all motifs; Additional data file 10 gives the distribution of pairwise dN values of motifs and flanking regions.

References

1. Stracke R, Werber M, Weisshaar B: **The R2R3-MYB gene family in *Arabidopsis thaliana*.** *Curr Opin Plant Biol* 2001, **4:**447-456.
2. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, *et al.*: ***Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes.** *Science* 2000, **290:**2105-2110.
3. Rabinowicz PD, Braun EL, Wolfe AD, Bowen B, Grotewold E: **Maize R2R3 Myb genes: sequence analysis reveals amplification in the higher plants.** *Genetics* 1999, **153:**427-444.
4. Ogata K, Morikawa S, Nakamura H, Sekikawa A, Inoue T, Kanai H, Sarai A, Ishii S, Nishimura Y: **Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices.** *Cell* 1994, **79:**639-648.
5. Klempnauer KH, Gonda TJ, Bishop JM: **Nucleotide sequence of the retroviral leukemia gene v-*myb* and its cellular progenitor c-*myb*: the architecture of a transduced oncogene.** *Cell* 1982, **31:**453-463.
6. Martin C, Paz-Ares J: **MYB transcription factors in plants.** *Trends Genet* 1997, **13:**67-73.
7. Rosinski JA, Atchley WR: **Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin.** *J Mol Evol* 1998, **46:**74-83.
8. Petroni K, Tonelli C, Paz-Ares J: **The MYB transcription factor family: from maize to *Arabidopsis*.** *Maydica* 2002, **47:**213-232.
9. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).** *Science* 2002, **296:**79-92.
10. Kranz HD, Denekamp M, Greco R, Jin H, Leyva A, Meissner RC, Petroni K, Urzainqui A, Bevan M, Martin C, *et al.*: **Towards functional characterisation of the members of the R2R3-MYB gene family from *Arabidopsis thaliana*.** *Plant J* 1998, **16:**263-276.
11. Meissner RC, Jin H, Cominelli E, Denekamp M, Fuertes A, Greco R, Kranz HD, Penfield S, Petroni K, Urzainqui A, *et al.*: **Function search in a large transcription factor gene family in *Arabidopsis*: assessing the potential of reverse genetics to identify insertional mutations in R2R3 MYB genes.** *Plant Cell* 1999, **11:**1827-1840.
12. Jiang C, Gu J, Chopra S, Gu X, Peterson T: **Ordered origin of the typical two- and three-repeated Myb genes.** *Gene* 2004, **326:**13-22.
13. Braun EL, Grotewold E: **Newly discovered plant c-*myb*-like genes rewrite the evolution of the plant myb gene family.** *Plant Physiol* 1999, **121:**21-24.
14. **Supplemental materials: index of */~czjiang/Myb*** [http://www.public.iastate.edu/~czjiang/Myb]
15. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP: **Prediction of plant microRNA targets.** *Cell* 2002, **110:**513-520.
16. Gilbert W: **The exon theory of genes.** *Cold Spring Harb Symp Quant Biol* 1987, **52:**901-905.
17. Fedorova L, Fedorov A: **Introns in gene evolution.** *Genetica* 2003, **118:**123-131.
18. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Curr Biol* 2003, **13:**1512-1517.
19. Lawrence CJ, Malmberg RL, Muszynski MG, Dawe RK: **Maximum likelihood methods reveal conservation of function among closely related kinesin families.** *J Mol Evol* 2002, **54:**42-53.
20. Toledo-Ortiz G, Huq E, Quail PH: **The *Arabidopsis* basic/helix-loop-helix transcription factor family.** *Plant Cell* 2003, **15:**1749-1770.
21. Athma P, Grotewold E, Peterson T: **Insertional mutagenesis of the maize *P* gene by intragenic transposition of *Ac*.** *Genetics* 1992, **131:**199-209.
22. Menssen A, Hohmann S, Martin W, Schnable PS, Peterson PA, Saedler H, Gierl A: **The *En*/*Spm* transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a *dSpm* element in the *A2* gene.** *EMBO J* 1990, **9:**3051-3057.
23. Waites R, Selvadurai HR, Oliver IR, Hudson A: **The *PHANTASTICA* gene encodes a MYB transcription factor involved in growth and dorsoventrality of lateral organs in *Antirrhinum*.** *Cell* 1998, **93:**779-789.
24. Tsiantis M, Schneeberger R, Golz JF, Freeling M, Langdale JA: **The maize *rough sheath2* gene and leaf development programs in monocot and dicot plants.** *Science* 1999, **284:**154-156.
25. Oppenheimer DG, Herman PL, Sivakumaran S, Esch J, Marks MD: **A *myb* gene required for leaf trichome differentiation in *Arabidopsis* is expressed in stipules.** *Cell* 1991, **67:**483-493.
26. Noda K, Glover BJ, Linstead P, Martin C: **Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor.** *Nature* 1994, **369:**661-664.
27. Lee MM, Schiefelbein J: **WEREWOLF, a MYB-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning.** *Cell* 1999, **99:**473-483.
28. Paz-Ares J, Ghosal D, Wienand U, Peterson PA, Saedler H: **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* 1987, **6:**3553-3558.
29. Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L: **The *Arabidopsis TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed.** *Plant Cell* 2001, **13:**2099-2114.
30. Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, Koes R: **Molecular analysis of the *anthocyanin2* gene of petunia and its role in the evolution of flower color.** *Plant Cell* 1999, **11:**1433-1444.
31. Grotewold E, Athma P, Peterson T: **Alternatively spliced products of the maize P gene encode proteins with homology to the DNA-binding domain of myb-like transcription factors.** *Proc Natl Acad Sci USA* 1991, **88:**4587-4591.
32. Zhang P, Chopra S, Peterson T: **A segmental gene duplication generated differentially expressed myb-homologous genes in maize.** *Plant Cell* 2000, **12:**2311-2322.
33. Aharoni A, De Vos CH, Wein M, Sun Z, Greco R, Kroon A, Mol JN, O'Connell AP: **The strawberry FaMYB1 transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tobacco.** *Plant J* 2001, **28:**319-332.
34. Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C: **Activation tagging identifies a conserved MYB regulator of phenylpropanoid**

biosynthesis. *Plant Cell* 2000, **12:**2383-2394.
35. Burns CG, Ohi R, Krainer AR, Gould KL: **Evidence that Myb-related CDC5 proteins are required for pre-mRNA splicing.** *Proc Natl Acad Sci USA* 1999, **96:**13789-13794.
36. Ogata K, Hojo H, Aimoto S, Nakai T, Nakamura H, Sarai A, Ishii S, Nishimura Y: **Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core.** *Proc Natl Acad Sci USA* 1992, **89:**6428-6432.
37. Dias AP, Braun EL, McMullen MD, Grotewold E: **Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication.** *Plant Physiol* 2003, **131:**610-620.
38. Hare MP, Palumbi SR: **High intron sequence conservation across three mammalian orders suggests functional constraints.** *Mol Biol Evol* 2003, **20:**969-978.
39. **Rice GD** [http://btn.genomics.org.cn:8080/rice]
40. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6:**361-365.
41. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.
42. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
43. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Mol Biol Evol* 2000, **17:**32-43.
44. Joshi CP, Zhou H, Huang X, Chiang VL: **Context sequences of translation initiation codon in plants.** *Plant Mol Biol* 1997, **35:**993-1001.