J Mol Evol (2003) 56:573–586
DOI: 10.1007/s00239-002-2426-x

JOURNAL OF MOLECULAR EVOLUTION

Genomewide Structural Annotation and Evolutionary Analysis of the Type I MADS-Box Genes in Plants

Stefanie De Bodt,[1] Jeroen Raes,[1] Kobe Florquin,[1] Stephane Rombauts,[1] Pierre Rouzé,[1,2] Günter Theißen,[3] Yves Van de Peer[1]

[1]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, K.L Ledeganckstraat 35, B-9000 Gent, Belgium
[2]Laboratoire Associé de I'lnstitut National de la Recherche Agronomique (France), Ghent University, B-9000 Gent, Belgium
[3]University of Jena, Lehrstuhl for Genetics, Philosophenweg 12, D-07743 Jena, Germany

Received: 1 August 2002 / Accepted: 18 November 2002

Abstract. The type I MADS-box genes constitute a largely unexplored subfamily of the extensively studied MADS-box gene family, well known for its role in flower development. Genes of the type I MADS-box subfamily possess the characteristic MADS box but are distinguished from type II MADS-box genes by the absence of the keratin-like box. In this *in silico* study, we have structurally annotated all 47 members of the type I MADS-box gene family in *Arabidopsis thaliana* and exerted a thorough analysis of the C-terminal regions of the translated proteins. On the basis of conserved motifs in the C-terminal region, we could classify the gene family into three main groups, two of which could be further subdivided. Phylogenetic trees were inferred to study the evolutionary relationships within this large MADS-box gene subfamily. These suggest for plant type I genes a dynamic of evolution that is significantly different from the mode of both animal type I (SRF) and plant type II (MIKC-type) gene phylogeny. The presence of conserved motifs in the majority of these genes, the identification of *Oryza sativa* MADS-box type I homologues, and the detection of expressed sequence tags for *Arabidopsis thaliana* and other plant type I genes suggest that these genes are indeed of functional importance to plants. It is therefore even more intriguing that, from an experimental point of view, almost nothing is known about the function of these MADS-box type I genes.

Key words: Structural annotations — Type I MADS-box gene family — *Arabidopsis* — Rice — Classification

Correspondence to: Yves Van de Peer; *email:* yvdp@gengenp.rug.ac.be

Introduction

The MADS-box gene family encodes a family of transcription factors involved in diverse aspects of plant development and has been designated by an acronym (Schwarz-Sommer et al. 1990) after a few of its earliest members, namely, ***M****CM1*, found in yeast (Passmore et al. 1988), ***A****GAMOUS*, in *Arabidopsis thaliana* (Yanofsky et al. 1990), ***D****EFICIENS*, in *Antirrhinum majus* (Sommer et al. 1990; Schwarz-Sommer et al. 1992), and ***S***RF, in human (Norman et al. 1988). All MADS-box genes encode a strongly conserved MADS domain—found in the N-terminal region—that is responsible for DNA binding to $CC(A/T)_6GG$ boxes in the regulatory region of their target genes (Shore and Sharrocks 1995). Recent analyses have shown that this large gene family can be divided into two major lineages, called type I and type II (Alvarez-Buylla et al. 2000b). Since both type I and type II genes are found in plants, animals, and fungi, both types of MADS-box genes are assumed to

have originated by duplication before the divergence of these kingdoms. Based on the structure of the MADS domain, type I and type II genes are also referred to as MADS SRF-like and MADS MEF2-like genes, respectively (Alvarez-Buylla et al. 2000b).

In animals, type I genes are involved in response to growth factors, while type II genes are involved in muscle development (Norman et al. 1988; Yu et al. 1992). Besides the highly conserved MADS domain, animal type I (SRF-like) and type II (MEF2-like) genes contain an additionally conserved region, the SAM and MEF2 domain, respectively (Shore and Sharrocks 1995, Riechmann and Meyerowitz 1997; Alvarez-Buylla et al. 2000b). The same is true for Fungi.

Plant type II MADS-box genes possess a strongly conserved MEF2-like MADS box, followed by a weakly conserved I (intervening) box, a K (keratin-like) box, and a C box and are therefore termed the MIKC-type (MIKC) genes (Münster et al. 1997). The moderately conserved K domain has been shown to be important for protein–protein interactions and probably forms a coiled-coil structure. The poorly conserved carboxyl-terminal (C) region may function as a trans-activation domain (Riechmann and Meyerowitz 1997). Plant type II MADS-box genes have been extensively studied during the last decade and are best known for their role in flower development (see, e.g., Riechmann and Meyerowitz 1997; Pelaz et al. 2000; Theißen et al. 2000; Ng and Yanofsky 2001; Theißen 2001; Theißen and Saedler 2001). Besides this role, MADS-box genes also have an important function in the development of other plant organs such as fruit (Liljegren et al. 1998, 2000), roots (Zhang and Forde 2000; Alvarez-Buylla et al. 2000a; Burgeff et al. 2002), and ovules (Angenent and Colombo 1996). The type II MADS-box transcription factors provide an excellent genetic toolkit to study the evolution of plant development. Alterations in the expression of genes coding for transcriptional regulators, such as MADS-box genes, are emerging as a major source of the diversity and change that underlie evolution and can be linked to changes in plant body plan or the generation of evolutionary novelties (Riechmann et al. 2000; Theißen 2001).

Unlike the type II MADS-box genes in plants, the type I subfamily has remained largely unexplored. Plant type I MADS-domain proteins are characterized by an SRF-like MADS domain but the C-terminal region of these genes is still not well defined and is of variable length. Furthermore, type I genes are characterized by the absence of the well-defined K box. Based on phylogenetic tree inference, Alvarez-Buylla et al. (2000b) concluded that this K box arose in plant type II genes after the divergence of plants and animals and fungi. Hitherto, only a few members of this subfamily have been identified by *in silico* prediction in *Arabidopsis thaliana*, whereas their function remains completely unknown (Alvarez-Buylla et al. 2000b). The recent discovery of this new subfamily of MADS-box genes in *Arabidopsis thaliana* and the lack of knowledge about their function urge upon the full characterization of this gene family in *Arabidopsis thaliana* and the identification of homologues in other plants. Moreover, further analysis of the type I MADS-box gene family may be very important in understanding the origin and evolution of the whole MADS-box gene family. In this respect, we have analyzed the size and the structural characteristics of the type I subfamily in *Arabidopsis thaliana* and have identified the first type I MADS-box genes in *Oryza sativa*. The completion of the *Arabidopsis thaliana* genome sequence (Arabidopsis Genome Initiative 2000) allows investigation of the full complement of MADS-box type I genes in this model plant. The structural annotation of the gene family was done in a semiautomated way, combining high-throughput gene prediction with a manual control step. By using this approach we tried to combine speed with accuracy because future research on these sequences depends on the correctness of their annotation. Additionally, we performed a phylogenetic analysis of the type I subfamily of MADS-box genes to study the evolutionary relationships between the newly annotated genes.

Methods

Structural Annotation of Type I MADS-Box Genes

The annotation of the type I MADS-box gene family in *Arabidopsis thaliana* was based on homology searches with the conserved part of the genes of the family. Hence, the MADS domain of the type I MADS-domain proteins identified by Alvarez-Buylla et al. (2000b) was used as a query sequence in BLAST (tblastn using default parameters) searches (Altschul et al. 1990) against the sequences of the *Arabidopis* genome. The E-value cutoff was initially set at 1e-10, where hits with higher E-values were selected manually, taking into account the conserved, possibly functionally important residues in the MADS domain. The genomic sequences containing putative type I MADS-box genes were subjected to gene prediction using GeneMark.hmm (Lukashin and Borodovsky 1998). A manual control step of the annotation involved the inspection of the exon–intron structure and the multiple alignment of the MADS-domain protein sequences using Artemis (Rutherford et al. 2000) and BioEdit (Hall 1999). Based on similarity to close relatives of the gene family, wrongly predicted exon borders and over- or under-prediction of exons were detected and corrected. To identify more distantly related proteins, we also constructed a HMMer profile (Eddy 1998) based on the already predicted and manually corrected genes. This profile was used to search a nonredundant database containing a collection of *Arabidopsis thaliana* proteins found through prediction with GeneMark.hmm (Lukashin and Borodovsky 1998) on the *Arabidopsis thaliana* genome (genome version of January 18 2001 [v180101], downloaded from the MIPS ftp-site at ftp://ftpmips.gsf.de/cress/). These gene predictions were then checked again manually.

Additionally, we searched for type I MADS-domain proteins in *Oryza sativa*. Based on the multiple sequence alignment of the

Table 1. *Arabidopsis thaliana* and *Oryza sativa* type I MADS-box genes

Locus name	Gene	Accession No.[a]	Start	Stop	Length	Strand	Chromosome	EST	Class
At1g28460		AC010155	35,082	35,630	182	–	1		M
At1g28450		AC010155_2	37,337	37,894	185	+	1		M
At1g60880		AC018908_2	24,777	25,352	191	–	1		M
At1g60920		AC018908_1	6,660	7,265	201	+	1		M
At3g04100		AC016829	84,782	85,405	207	+	3		M
At1g01530	AGL28	Y12776	6,766	7,788	247	+	1		M
At1g65360	AGL23	AC004512_2	47,399	48,213	226	+	1		M
At2g24840		AC006585	25,227	25,859	210	+	2		M
At5g60440		ABO11483	26,829	28,020	299	+	5		M
At4g36590	AGL40	AL161589	121,429	123,079	243	–	4		M
At5g38620		AB005231	463,826	464,875	349	–	5		M
At5g49420		AB023034	34,638	36,134	402	–	5		M
At2g34440	AGL29	AC004077	16,781	17,299	172	+	2		M
At1g48150		AC023673	497,767	498,738	323	+	1		M
At5g27130	AGL39	AF007271	71,901	75,618	435	–	5		M
At1g47760		AC012463	70,240	70,948	184	–	1		M
At3g66656		AC036106	29,224	29,760	178	–	3		M
At4g14530		AL161539	46,973	47,658	213	–	4		M
At5g49490		AB023033	10,587	11,330	247	+	5		M
At5g04640		AL162875	89,521	90,489	322	+	5		M
		Os_AP003951_1	28,733	29,365	633	–	6		M
		OS_AP003951_2	50,199	50,771	572	–.	6		M
		Os_AP003627	102,168	102,794	627	+	1		M
		Os_AP004093	73,268	73,128	861	+	2		M
		Os_Contig2417	5,705	9,967	210	+			M
		Os_Contig4095	1,453	2,109	218	+			M
		Os_Contig4276	6,289	6,921	210	+			M
		Os_Contig28459	1,540	2,078	141	–			M
		Os_Contig18609	465	1,049	194	+			M
At5g26580/ At5g26575[b]		AF058914	4,471	5,508	304	+	5		N
At5g26630/ At5g26625[b]		AF058914_2	40,425	47,737	315	–	5		N
At5g26650/ At5g26575[b]		AF058914_3	53,688	54,794	327	–	5	X	N
At1g65330		AC004512_1	32,543	33,382	279	–	1		N
At1g65300		AC004512_3	21,003	21,827	278	+	1		N
At3g05860		AC012393	56,275	57,224	260	–	3		N
At2g28700		AC007184	3,732	4,502	256	–	2		N
At5g27960		AC007627	64,277	65,368	363	–	5		N
At5g48670		AB015468	59,946	60,911	321	–	5		N
At1g31630		AC074360_2	59,951	60,970	339	+	1		N
At1g31640		AC074360_1	55,322	56,806	464	+	1		N
At2g40210		AC018721	40,137	42,106	402	–	2	X	N
At2g26880	AGL41	AC005168	51,386	52,188	260	–	2		N
		Os_AP002070_1	57,225	57,947	240	+	1		N
		Os_AP002070_2	71,207	72,127	306	+	1		N
		Os_Contig28311	1,167	1,580	138	–			N
		Os_Contig603	3,973	4,776	267	+			N
		Os_Contig23118	1,479	1,904	141	+			N
		Os_Contig18573	850	1,079	209	+			N
		Os_Contig119850	52	667	205	–			N
		Os_Contig31610	1,805	2,215	136	–			N
		Os_Contig18149	1,420	2,035	205	–			N
At2g03060[c]	AGL30	AC004138	81,852	83,914	364	+	2		O
At1g31140		AC004793	29,171	30,813	211	+	1		O
At1g22590		AC006551	24,033	24,524	163	–	1	X	O
At1g77950		AC009243	50,294	52,808	244	+	1	X	O
At1g72350		AC016529	73,282	73,956	224	+	1		O
At1g17310		AC026479	2,189	2,827	212	–	1		O
At5g26950	AGL26	AF007270	84,574	85,554	292	–	5		O
At2g26320	AGL33	AC004484	66,595	68,743	209	–	2		O
At1g18750[c]		AC011809	60,126	62,604	440	+	1		O
At1g22130		AC0690252	2,812,402	2,814,274	335	–	1		O

Continued

Table 1. Continued

Locus name	Gene	Accession No.[a]	Start	Stop	Length	Strand	Chromosome	EST	Class
At1g77980		AC009243_2	58,477	60,332	303	–	1		O
At1g69540		AC073178	88,592	90,480	359	–	1		O
At5g06500		AP002543	7,047	7,775	728	+	5		O
At5g58890	AGL43	AB016885	33,758	34,642	294	+	5		O
At5g55690		AB009050	40,372	41,205	277	–	5		O
		Os_AP000616	39,576	42,209	855	+	6		O
		Os_AP003104	53,129	54,943	1,815	+	1		O
		Os_AP003331_1	86,944	88,167	1,224	+	1		O
		Os_AP003331_2	89,653	90,947	975	+	1		O
		Os_AP003380	8,256	9,365	1,110	–	1		O
		Os_AP003436	171,451	172,890	1,440	–	1		O
		Os_AP003763	127,881	128,597	279	+	6		O
		Os_AP003742	63,104	64,429	645	–	7		O
		Os_AP004322	4,659	10,836	477	+	6		O
		Os_AP003331_3	95,818	98,653	1,188	+	1		O
		Os_Contig19550	853	1,324	375	+			O
		Os_Contig52002	790	?	?	+			O
		Os_Contig20368	?	405	?	–			O
		Os_Contig45237	1,180	?	?	+			O
		Os_Contig11428	5,081	?	?	+			O
		Os_Contig32902	2,555	?	?	–			
		Os_Contig2175	12,725	?	?	–			Unassigned[d]
		Os_Contig5668	?	5,842	?	–			
		Os_Contig21589	?	2,624	?	–			

[a] Rich genes are in boldface.
[b] Genes on BAC AF058914 have different locus names in MIPS and TIGR (http://www.tigr.org), respectively.
[c] Locus names of genes in MIPS (Schoof et al. 2002) that differ in their structural annotation from those presented here.
[d] These genes could not be classified unambiguously because the prediction was incomplete (see text for details).

Arabidopsis thaliana type I MADS-domain proteins, a HMMer profile (Eddy 1998) was built to search a rice protein database for type I MADS-domain proteins. This database contained 24,305 rice proteins predicted with GeneMark.hmm (Lukashin and Borodovsky 1998) on rice BAC sequences from the Rice Genome Project covering approximately 29% of the rice genome (*Oryza sativa* spp. *japonica* [Sasaki and Burr 2000; http://rgp.dna.affrc.go.jp/]). Furthermore, we screened the draft sequence of *Oryza sativa* spp. *indica* (Yu et al. 2002) for putative type I MADS-box genes using BLAST, with other type I genes as query sequences.

Duplicated blocks (i.e., large regions of colinearity) in the *Arabidopsis thaliana* genome were detected and dated as described earlier (Vandepoele et al. 2002; Raes et al. 2003).

Structural Analysis of the C-Terminal Region

All type I MADS-box genes possess the strongly conserved MADS box. However, the C-terminal region of these genes is much less conserved and has a variable length. We performed a motif search on all type I MADS-domain protein sequences using MEME (Multiple Expectation Minimization for Motif Elicitation), version 3.0 (Bailey and Elkan 1994). Based on the conserved motifs found by MEME, the type I MADS-box gene family was further subdivided into smaller subgroups, after which these subgroups were realigned, now taking into account additional sites that could be proven to belong to shared and conserved motifs.

A HMMer profile (Eddy 1998) was built from the different motifs identified by MEME (see Results). These profiles were scanned against our in-house *Arabidopsis thaliana* protein database (see Structural Annotation and Phylogenetic Analysis, below) and the MIPS protein database to search for other proteins that contain similar motifs. The InterPro database (release 4.0, November 2001 [Apweiler et al. 2001]) was also checked for the presence of the C-terminal motifs.

Table 2. List of MADS-like genes in *Arabidopsis thaliana*

Locus name	Accession No., BAC
At5g27090	AF170760
At5g27070	AF170670
At5g27580	AC007478
At5g26950	AF007270
At4g11250	AL096882
At5g65330	AB011479
At5g40220	AB010699
At5g39750	AB016876
At5g38740	AB011478
At5g40120	AB010699
At5g39810	AB016876
At5g41200	AB010072
At3g18650	AB026654
At5g27050	AF170670
At1g60040	AC005966
At1g59810	AC007258

To make sure that no type II MADS-domain proteins have been included in our data set, all sequences were analyzed for the presence of the type II-specific K domain using InterPro searches (release 4.0, November 2001 [Apweiler et al. 2001]) and Multicoil (Wolf et al. 1997) for coiled-coil prediction based on the presence of heptat-repeat signature motifs (abcdefg, where a and d are hydrophobic residues and are pointing to the core of the coiled-coil and b, d, e, f, and g are hydrophylic residues) in the sequences (Lupas 1996).

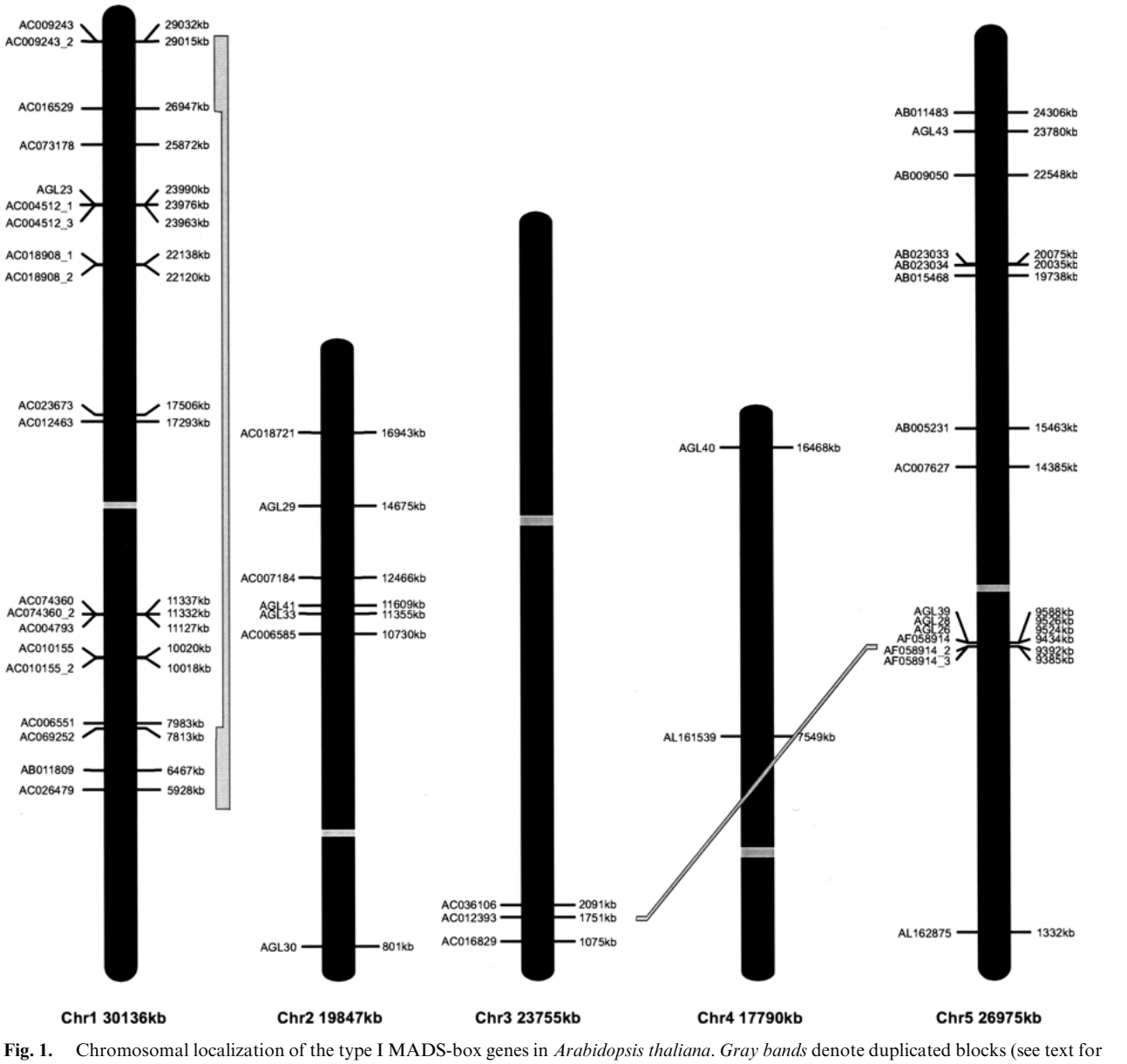

Fig. 1. Chromosomal localization of the type I MADS-box genes in *Arabidopsis thaliana*. *Gray bands* denote duplicated blocks (see text for details).

Phylogenetic Analysis of Type I MADS-Domain Proteins

The complete alignment of all type I MADS-domain proteins was edited and reformatted for phylogenetic analysis using BioEdit (Hall 1999) and ForCon (Raes and Van de Peer 1999), resulting in an alignment of the conserved residues (MAD domain + residues of shared motifs). Neighbor-joining (Saitou and Nei 1987) trees were constructed using TREECON (Van de Peer and De Wachter 1997) based on Poisson-corrected distances. To assess support for the inferred relationships, 500 bootstrap samples (Felsenstein 1985) were generated.

Maximum likelihood trees were constructed for type I MADS-box genes (see below) using TREE-PUZZLE 5.0 (Strimmer and von Haeseler 1996; Schmidt et al. 2002) and PAML (Yang 2000). In TREE-PUZZLE, the mutation probability matrix of Müller and Vingron (2000) was used, whereas the number of puzzling steps was set to 20,000. Bootstrapped maximum parsimony trees for class M and class N genes were constructed with PAUP* (Swofford 1998). Predicted sequences and multiple alignments are available at our web site, http://www.psb.ac.be/bioinformatics/MADS/.

Results

Structural Annotation and Phylogenetic Analysis

Based on a genomewide analysis, we identified 47 type I MADS-box genes in the genome of *Arabidopsis thaliana,* of which 14 correspond to genes previously described by Alvarez-Buylla et al. (2000b) and 33 are new (see Table 1). Additionally, we discovered the presence of a new group of MADS-like genes. These genes are different from type I (and also type II) MADS-box genes due to a highly divergent N-terminal region of the MADS box. Furthermore, al-

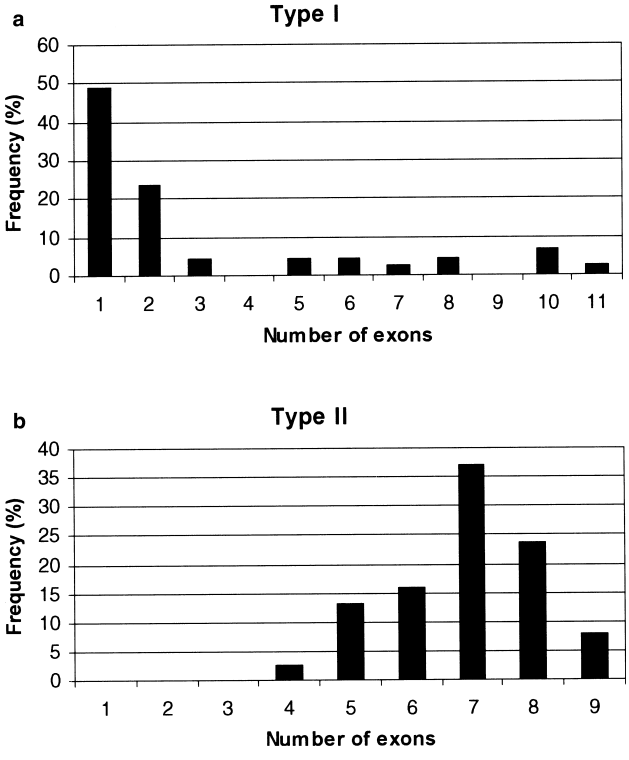

Fig. 2. Distribution of the number of exons in the type I (**a**) and type II (**b**) MADS-box gene family.

though most of these genes are overall strongly conserved, they do not possess the C-terminal conserved regions characteristic of type I (or type II) genes. For these reasons, we did not include these genes (listed in Table 2) in our analyses.

Figure 1 shows the distribution of the type I MADS-box genes on the different chromosomes. Seven genes could be linked to block duplications, namely, both the gene pairs AC016529 and AC026479 and the gene pairs AC009243_2 and AC069252, which are all located in an internally duplicated block that contains 172 duplicated genes on chromosome 1 (Simillion et al. 2002; Raes et al. 2003). Additionally, genes AC012393 and AF058914_2 (and its neighbor AF058914_3) belong to a smaller block of 13 genes duplicated between chromosome 3 and chromosome 5 (Fig. 1). The largest block has been dated to 69 ± 17 MYA, while the smaller block duplication was dated to 78 ± 29 MYA, which implies that they could both have originated during the same complete genome duplication event, estimated to have occurred at about that time (Lynch and Conery 2000; Simillion et al. 2002; Raes et al. 2003).

Figure 2a shows the distribution of the number of exons found in type I MADS-box genes. As can be observed, the majority of the type I genes consist of only one or two exons, which is quite different from type II MADS-box genes, where most genes consist of seven exons (Fig. 2b). In addition to the *Arabidopsis thaliana* type I genes 16 rice type I MADS-box genes were annotated on BAC sequences of the rice consortium (Sasaki and Burr 2000). Preliminary analysis of the draft sequence of rice resulted in the additional identification of 19 putative type I MADS-box genes. Six other genes were found through BLAST searches on the rice draft sequence but could not be ascribed unequivocally to the type I subfamily. Further analysis and manual annotation of these rice genes will be necessary to decide whether these are type I or type II genes. Furthermore, to improve gene prediction in rice, an assembly of the contigs of the draft sequence will be necessary because many MADS-box genes are located at the end of the contigs. We also searched the publicly available databases for type I MADS-box genes of other plants but could not find any other type I homologues. It should be noted that the sequencing and annotation of other plant sequences are still ongoing, which will probably result in the detection of many more type I MADS-domain proteins in the near-future.

The construction of reliable phylogenetic trees of the complete type I subfamily of MADS-domain proteins is very difficult due to the small size (60 amino acids) of the conserved MADS domain. Trees constructed on such a low number of residues often turn out to be unreliable and poorly supported by statistical analyses. As shown in Fig. 3, very few nodes are well supported and no conclusion can be drawn about possible subclasses present in the type I MADS-box gene family. Therefore, we applied alternative approaches to resolve the phylogeny of the gene family (see also Methods).

Detailed structural analysis using MEME (Bailey and Elkan 1994) enabled us to discover several conserved motifs in the C-terminal region of the type I MADS-domain proteins (summarized in Figs. 4 and 5). Two main distinct classes of type I MADS-domain proteins, which we designate class M and class N, can be identified, each of which can be further subdivided. Class M possesses three types of genes, *viz.*, type I M1 genes, which are characterized by motifs 1, 2, and 3; type I M2 genes, characterized by motifs 1 and 3; and type I M3 genes, which contain only motif 1 (Fig. 4). Class N possesses three types of genes, *viz.*, type I N1 genes, which are characterized by motifs 4, 5, 6, 7, sometimes 8, and 9; type I N2 genes, which possess motifs 4 and 5 and have a degenerated form of motif 6; and, finally, type I N3 genes, which contain only motifs 4 and 5 (Fig. 5). Next to class M and class N genes, there is a third class O of genes that do not possess the same conservation in the C-terminal region as the proteins in the other classes. Thus, although specific motifs could be identified for class M and N genes, it was not

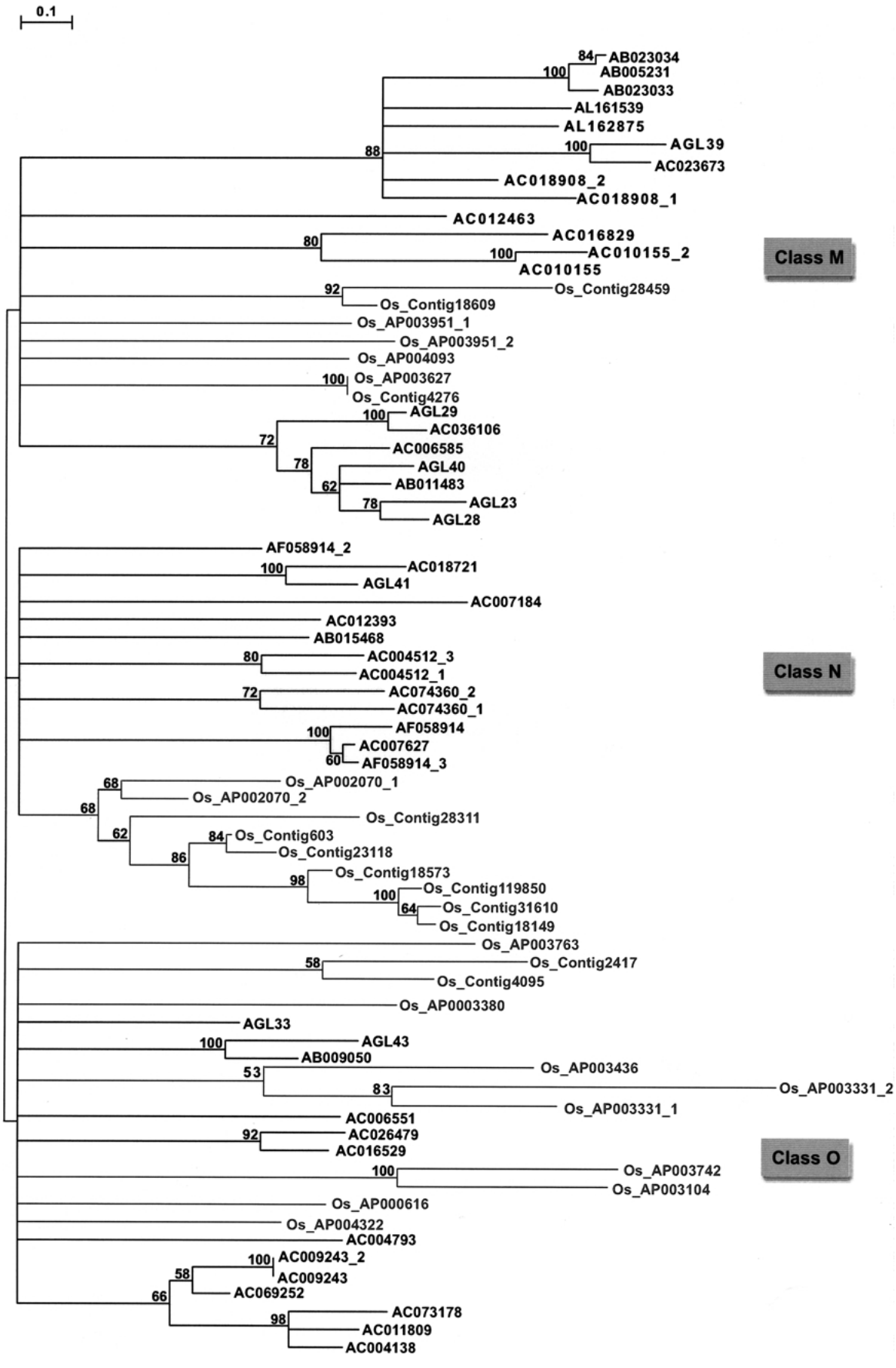

Fig. 3. Phylogenetic distance tree of all type I MADS-box proteins identified in *Arabidopsis thaliana* and *Oryza sativa*. Tree construction was based on only 47 conserved residues in the MADS domain. Five hundred bootstrap samples (Felsenstein 1985) were taken and branches are drawn as unresolved when supported by less than 50%. Based on the presence or absence of C-terminal motifs, genes were ascribed to class M, N, or O (see text for more details). Rice proteins are indicated in *gray*. The scale indicates 0.1 substitution per site.

Class M

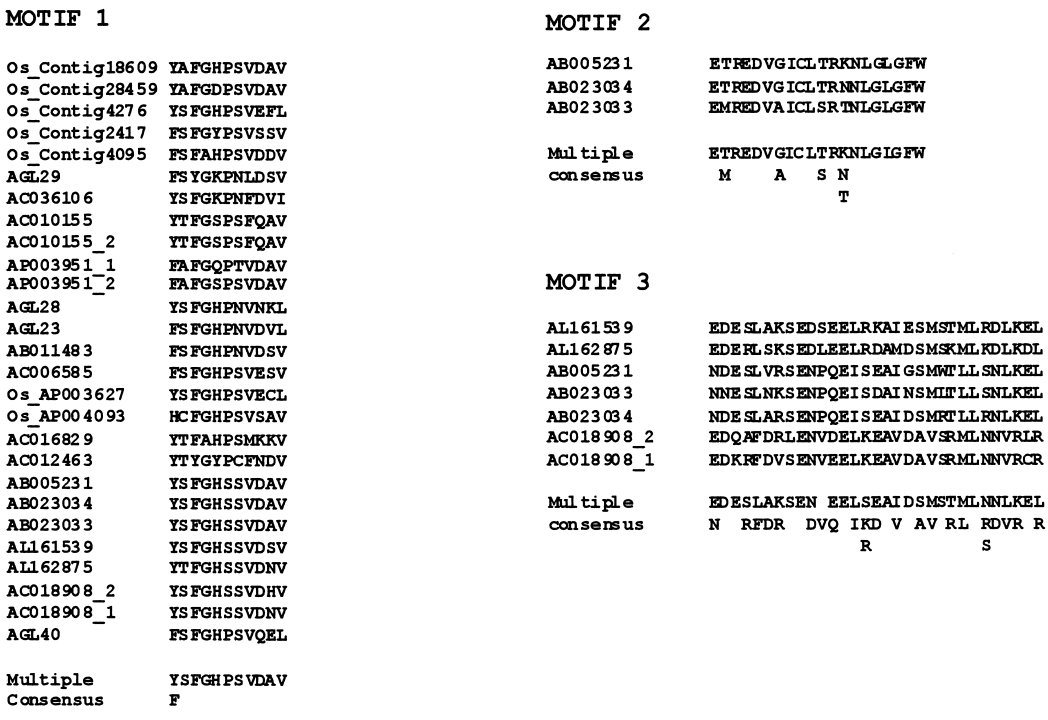

Fig. 4. Conserved motifs in the C-terminal region of class M proteins of the type I MADS-box gene family found by MEME (Bailey and Elkan 1994). Rice genes are preceded by the prefix Os. The multilevel consensus sequence is calculated from the motif position-specific probability matrix computed by MEME. For each column of the motif, the amino acid residues are sorted in decreasing order by the probability with which they are expected to occur at a certain position of the motif. The most probable amino acid is on top. Only amino acids with probabilities of 0.2 or higher at that position in the motif are listed.

possible to find any conserved motif for the proteins that we classified as belonging to class O. It should be noted that type I MADS-box genes of rice have been found for all three classes (see Fig. 1).

Classification of the type I MADS-box genes into classes M and N on the basis of the presence of certain conserved motifs allowed alignment of longer regions of the type I MADS-box genes. Therefore, a phylogenetic tree was constructed for genes belonging to class M from an alignment of 76 conserved residues, including the MADS domain and motif 1 (shared among all the genes belonging to class M), whereas a second tree for class N genes was constructed from an alignment of 116 conserved residues, based on the MADS domain and motifs 4 and 5. These trees are shown in Figs. 6 and 7, respectively. Both trees were artificially rooted based on the presence or absence of certain motifs. As expected, in general there is a clear correlation between the tree topology and the structural characteristics of a group of proteins. In other words, proteins with the same C-terminal motif composition seem to be more closely related. In a few cases, remnants of common ancestry can be found, but the conservation was too low to be picked up by MEME. For example, genes of type I N2 do not contain motif 6 according to MEME, but some residues of the consensus sequence of this motif can still be recognized in these proteins. Therefore, these motifs are represented by hatched boxes (Figs. 6 and 7).

The trees shown in Figs. 6 and 7 are neighbor-joining trees (Saitou and Nei 1987) based on Poisson-corrected distances computed with TREECON (Van de Peer and De Wachter 1997). Overall, maximum likelihood trees and maximum parsimony trees gave similar results and differences were observed only for nonsupported nodes. As expected, the resolution of the trees seems to be correlated with the number of residues that could be taken into account for tree inference. The tree of class M genes, shown in Fig. 6 and based on 76 alignment positions, is still not very well resolved, apart from one subgroup of sequences that also contain additional conserved motifs (type I M1 and type I M2). Although strong conclusions cannot be drawn regarding the rice genes, due to the uncertainty of most branching orders, it seems that none of the rice genes is specifically related to any of the *Arabidopsis thaliana* genes. This is also observed in the tree of the N genes, based on 116 alignment positions, where the rice genes clearly form a monophyletic group, which is well supported by bootstrap analysis and different methods of tree construction (Fig. 7).

Class N

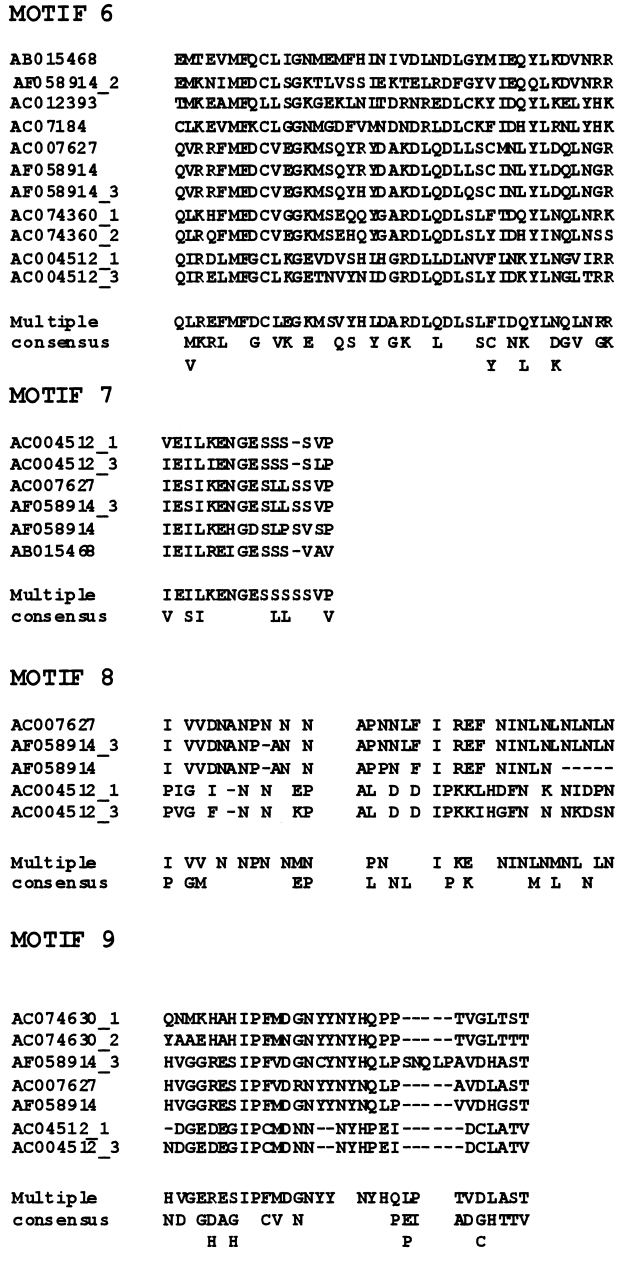

Fig. 5. Conserved motifs in the C-terminal region of class N proteins of the type I MADS-box gene family found by MEME. Interpretation is as in Fig. 4.

Functional Annotation

To assign a putative function to the type I MADS-box genes, we analyzed the C-terminal part of these genes in more detail. Genes that encode transcription factors often contain a transcription-activating domain. Three types of trans-activation domains are described in the literature: they are rich in acidic residues, proline residues, or glutamine residues but have low overall conservation on the primary structure level (Latchman 1998). Type I M1 and type I M2 proteins contain an acidic region in their characteristic motif 3. Class N proteins all contain a proline-rich region, starting from approximately position 160. This region shows low conservation on the primary sequence level and does not correlate with any particular C-terminal motif designated by MEME. However, as stated before, the abundance of prolines in this region might possibly refer to the trans-activation domain of these proteins (Latchman 1998). However, apart from these putative trans-activation domains, little can be said about the C-terminal region.

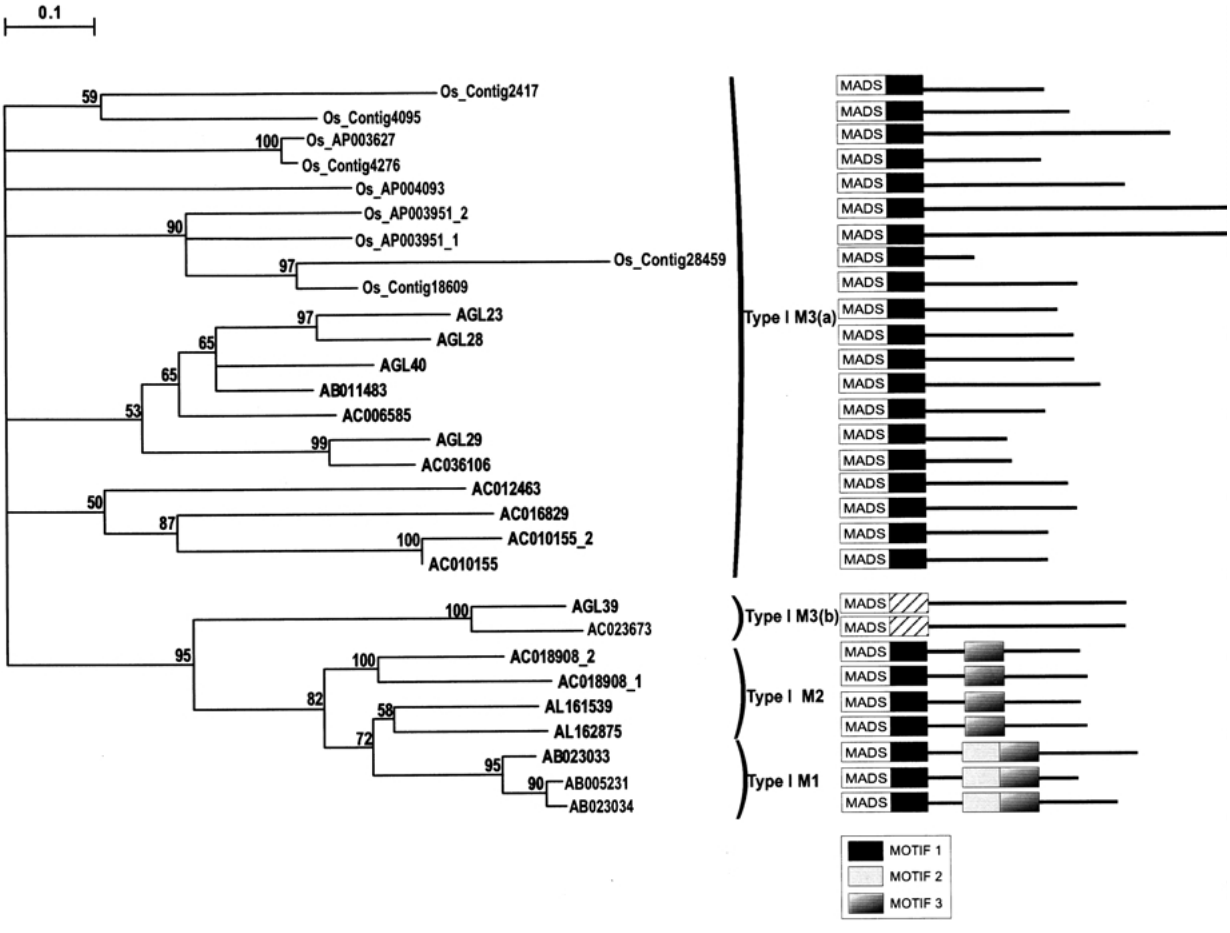

Fig. 6. Pairwise distance tree of the type I MADS-box genes belonging to class M (see text for details), inferred from a sequence alignment including sites of the MADS domain and motif 1. The motif composition of each gene is denoted by a *black line* (representing the length of the sequence) and *shaded boxes*. A *hatched box* denotes a degenerated form of the motif. Rice genes are preceded by the prefix Os. Interpretation of the scale is as in Fig. 3.

For example, no similarity could be found between the profiles inferred from the conserved motifs and any previously described motifs or domains (InterPro release 4.0, November 2001 [Apweiler et al. 2001]).

To get more information on the expression of type I MADS-box genes, and their possible functional annotation, we screened *Arabidopsis thaliana* ESTs, rice ESTs, and an EST collection containing all publicly available ESTs from diverse plant species. However, the number of ESTs corresponding to type I MADS-box genes of *Arabidopsis thaliana* was extremely low (see Table 3), in particular, in comparison with ESTs for type II genes, where on average four or five ESTs per gene could be identified. We found one EST (C99890) for type I gene AGL39 (type I M3[b]), which had also been identified previously by Alvarez-Buylla et al. (2000a) and ESTs for four other *Arabidopsis thaliana* genes. Some ESTs from other plant species could be found that were long enough to demonstrate unambiguously that they are ESTs from type I MADS-box genes (Table 3). ESTs of type I MADS-box genes are found in diverse plant species such as *Glycine max*, *Lycopersicon esculentum*, *Triticum aestivum*, and even *Ceratopteris richardii* (a fern) and *Physcomitrella patens* (a moss).

Discussion

Detailed structural and evolutionary analysis of the type I subfamily of MADS-box genes suggests that these genes are indeed of functional importance in plants. The type I subfamily possesses 47 members, which is more than the number of members of the very well-studied type II subfamily (unpublished results). Moreover, in a first preliminary analysis, 33 type I genes have already been identified in *Oryza sativa* spp. *japonica* on BAC sequences of the rice consortium (December 2001) and on the draft sequence of *Oryza sativa* spp. *indica*. Furthermore, *Arabidopsis thaliana* and rice type I proteins still have conserved common motifs in their C-terminal region (rice genes are present in the type I M3[a] and type I N3 classes). This conservation is most likely due to

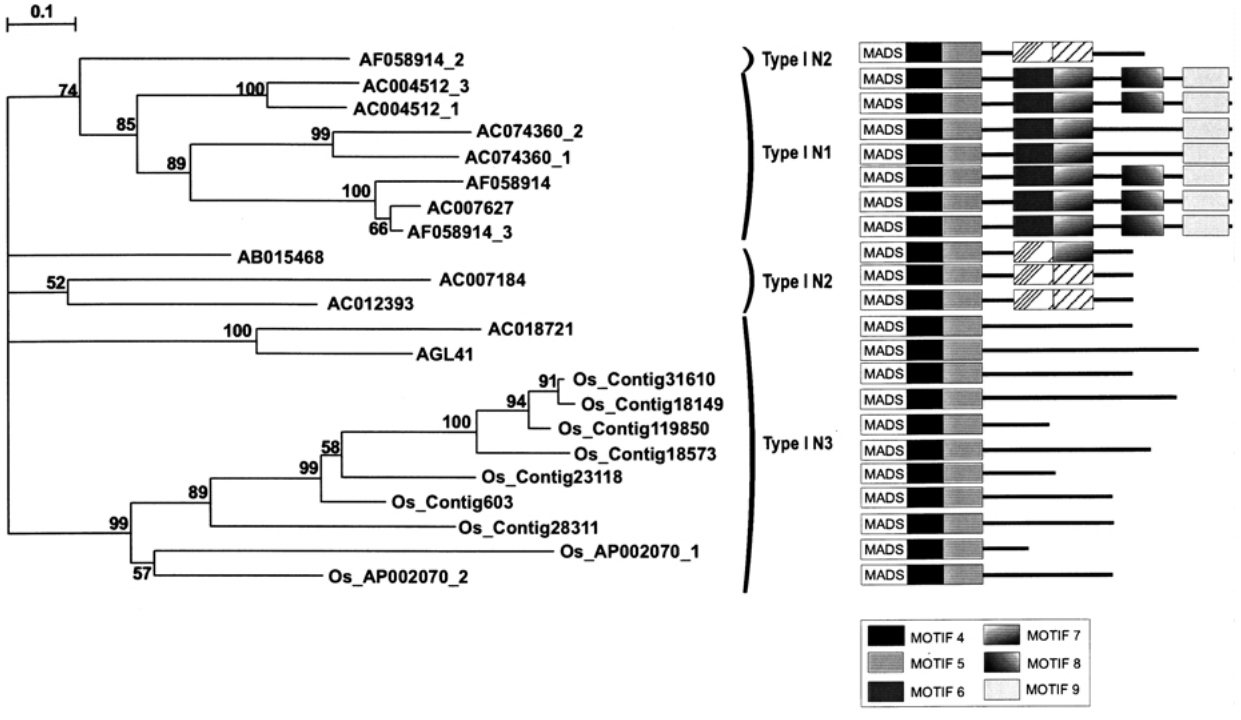

Fig. 7. Pairwise distance tree of the type I MADS-box genes belonging to class N, inferred from a sequence alignment including sites of the MADS domain and motifs 4 and 5. Interpretation is as in Fig. 6. Interpretation of the scale is as in Fig. 3.

functional constraints on the C-terminal region, although the overall functional constraint within the type I genes has probably been lower than that within the type II genes. This is, among other things, supported by the higher evolutionary distances between type I MADS-box genes (Alvarez-Buylla et al. 2000b; our own observations).

Unfortunately, based on *in silico* analyses, we cannot assign a putative function to the type I MADS-domain proteins. The low number of ESTs found for type I MADS-box genes of different plant species can probably be attributed to the fact that most of the type I genes have a very low expression level or that the genes are expressed under very specific conditions that are not yet monitored in EST-sequencing projects. Strikingly, nearly half of type I genes are intronless (Fig. 2). This gene structure could possibly be interpreted as a result of the evolutionary history of the type I genes through reverse transcription, with the possibility that many of them are inactive pseudogenes. However, it should be noted that gene AC006551, for which we found three ESTs, consists of only one exon, which argues that at least some of these genes are expressed and functional, and not pseudogenes as put forward by Ng and Yanofsky (2001). In maize, transposon-like elements have been identified that have recently hijacked AGAMOUS-like (type II) MADS boxes and distributed them through the maize genome (Fischer et al. 1995; Montag et al. 1995, 1996). To investigate whether this could have been the case for the *Arabidopsis thaliana* type I genes, we looked for characteristic transposon-like elements in the flanking and coding regions of the type I genes. To this end, we searched for similarity with known (retro)transposons and with proteins involved in their activity such as pol, gag, and RT (Bennetzen 2000). However, no evidence for the presence of transposable elements could be found in our analyses.

As stated previously, all the type I MADS-box class N rice genes form a well-supported monophyletic grouping, while a monophyletic origin of the rice class M genes cannot be ruled out on the basis of tree inference. If true, and provided that the root in Figs. 6 and 7 is placed correctly, this would suggest that the expansion of both the *Arabidopsis thaliana* and the rice class M and N type I MADS-box genes (nothing can be said about genes from class O) occurred after the divergence of these two plants, somewhere between 150 and 200 MYA (Wikstrom et al. 2001). This is in clear contrast with observations in MADS type II phylogenies, according to which the last common ancestor of extant gymnosperms and angiosperms already contained at least seven different MIKC-type MADS-box genes (Becker et al. 2000). If type I MADS-box genes were present in the most recent common ancestor of plants, animals, and fungi, as suggested by Alvarez-Buylla et al. (2000b), and our observations are correct, this would imply that type I MADS-box genes may have remained low-copy (or even single-copy) for many hundreds of millions of years, until the most recent common ancestor of

Table 3. List of ESTs found for type I MADS-box genes in different plant species

EST	Gene	Plant species	Expression[a]
AV558219	AF058914_3	*Arabidopsis thaliana*	Organ: green siliques
AV823886	AC018721	*Arabidopsis thaliana*	Developmental stage: in various developmental stages, from germination to mature seeds Treatment: dehydration and cold
AV787106	AC018721	*Arabidopsis thaliana*	Idem
AV787440	AC018721	*Arabidopsis thaliana*	Idem
AV788503	AC018721	*Arabidopsis thaliana*	Idem
AV784963	AC018721	*Arabidopsis thaliana*	Idem
AU238686	AC006551	*Arabidopsis thaliana*	Treatment: cold
Z37169	AC006551	*Arabidopsis thaliana*	Tissue type: green shoots
F13558	AC006551	*Arabidopsis thaliana*	Tissue type: green shoots
AU236968	AC009243	*Arabidopsis thaliana*	Organ: flowers and siliques
AV556667	AF058914	*Arabidopsis thaliana*	Organ: green siliques
BE610209		*Glycine max*	Tissue type: immature seed coats of greenhouse-grown plants
BE823841		*Glycine max*	From cDNA libraries from various tissues and stages of development of soybean that represent 2639 sequences from immature cotyledons, 1770 from immature seed coats, 3938 from flowers, and 869 from young pods
AW508033		*Glycine max*	From a cDNA library that was constructed from mRNA isolated from immature cotyledons of greenhouse-grown plants
BE054256		*Gossypium arboreum*	Tissue type: fibers isolated from bolls harvested 7–10 dpa
BE999756		*Medicago truncatula*	Tissue type: senescent root nodules Developmental stage: mixture of effective nodules from 40-day-old plants harvested 36 h post-shoot removal and nodules collected from 2-month-old plants at midpod stage
AW029842		*Lycopersicon esculentum*	Tissue type: callus Developmental stage: 25–40 days old
BI929334		*Lycopersicon esculentum*	Tissue type: flower Developmental stage: 3- to 8-mm buds
BG139571		*Lycopersicon pennellii*	Tissue type: pollen Developmental stage: pollen collected from open flowers
BJ247094		*Triticum aestivum*	Tissue type: spike at flowering date Developmental stage: Feekes' scale 10.5.1
BJ248139		*Triticum aestivum*	Idem
BJ218990		*Triticum aestivum*	Tissue type: spike at meiosis Developmental stage: Feekes' scale 9
BG525865		*Stevia rebaudiana*	Tissue type: leaf Developmental stage: field grown midsize
AW010840		*Pinus taeda*	Organ: shoot tips
BE643398		*Ceratopteris richardii*	Tissue type: gametophyte Cell type: spore Developmental stage: 20 h after germination initiation
BJ184681		*Physcomitrella patens* subsp. *patens*	Tissue type: mixture of chloronemata, caulonemata, and malformed buds

[a] Expression details (e.g., tissue or organ, condition) are as described in the EMBL entries.

Arabidopsis thaliana and rice, and then started to multiply independently, giving rise to high gene numbers in both *Arabidopsis thaliana* and rice. This seems highly unrealistic, given the evolutionary history of type II MADS-box genes (Becker et al. 2000; Krogan and Ashton 2000; Theißen et al. 2001). An alternative explanation is that the type I genes from animals, and plants are not monophyletic, i.e., that they originated two times independently in plants and animals, and, at least for plants, much more recently than previously suggested. In line with this, the type I genes from animals (SRF-like genes) have a structure which is significantly different from that of plant type I genes, and obvious sequence similarity between both gene types is restricted to the MADS domain anyway (Alvarez-Buylla et al. 2000b). Animal type I genes have an evolutionary history which is different from that of plant type I genes: while the gene number of the latter increased dramatically in the lineages that led to extant *Arabidopsis thaliana* and rice (this work), SRF seems to have remained a single-copy gene throughout the more than 500 million years of animal evolution and represents the evolutionarily most conserved subfamily of MADS-box genes (Escalante and Sastre 1998; Hoffmann and Kroiher 2001; Scheffer et al. 1997). As stated by Alvarez-Buylla et al. (2000b), the type I MADS-box clade in plants is defined by only one putative synapomorphy, while some synapomorphies are shared by all but one or a few sequences; this cannot be

considered strong proof for a monophyletic origin of type I MADS-box genes.

On the other hand, it is possible that there are orthologous type I genes in *Arabidopsis thaliana* and rice but that phylogeny reconstruction, due to the limited number of phylogenetically informative sites, is unable to identify them correctly. Probably, the identification of type I genes from other plants will be necessary to clarify this. This is not yet possible, however, due to the limited genomic data from other plant species.

Hopefully, as suggested by Riechmann and Ratcliffe (2000), *in silico* studies on the annotation and classification of specific gene families, such as the one described here, can guide future experimental work and enhance the functional characterization of genes.

Note Added in Proof

After acceptance, novel MADS-box genes were identified in *Physcomitrella* [Henschel K, Kofuji R, Hasebe M, Saedler H, Munster T, Theißen G (2002) Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. Mol Biol Evol 19:801–814]. By including the MIKC* (type II) genes (PPM3, PPM4, PPMADS2, and PPMADS3) in our analysis, some of the *Arabidopsis* genes that we denoted as being of type I clustered with the *Physcomitrella* genes. Although these *Arabidopsis* genes did not seem to possess a conserved K-box (the reason why they were included), a relic of this box could be identified through comparison with the very degenerated K-box found in *Physcomitrella*. Therefore, some of the genes (i.e., AC011809, AC073178, AC004138, AC069252, AC009243, AC009243_2, and AC004484; Fig. 1) should probably be classified as type II rather than type I genes in our study.

Acknowledgments. The authors want to thank Klaas Vandepoele and Cedric Simillion for technical help. S.D. and K.F. are indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) for a predoctoral fellowship. Annotated sequences have been submitted to the MAtDB (Schoof et al. 2002) and the TAIR (Huala et al. 2001) databases. Supplementary data are available at http://www.psb.rug.ac.be/bioinformatics/MADS/.

References

Altschul SF, Gish GW, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Alvarez-Buylla ER, Liljegren SJ, Pelaz S, Gold SE, Burgeff C, Ditta GS, Vergara-Silva F, Yanofsky MF (2000a) MADS-box gene evolution beyond flowers: Expression in pollen, endosperm, guard cells, roots and trichomes. Plant J 24:457–466

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS (2000b) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci USA 97:5328–5223

Angenent GC, Colombo L (1996) Molecular control of ovule development. Trends Plants Sci 1:228–232

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37–40

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:798–815

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp 28–36

Becker A, Winter K-U, Meyer B, Saedler H, Theißen G (2000) MADS-box gene diversity in seed plants 300 million years ago. Mol Biol Evol 17:1425–1434

Bennetzen JL (2000) Transposable elements contributions to plant gene and genome evolution. Plant Mol Biol 42:251–269

Burgeff C, Liljegren SJ, Tapia-Lopez R, Yanofsky MF, Alvarez-Buylla ER (2002) MADS-box gene expression in lateral primordia, meristems and differentiated tissues of *Arabidopsis thaliana* roots. Planta 214:365–372

Eddy SR (1998) Profile hidden Markov model. Bioinformatics 14:755–763

Escalante R, Sastre L (1998) A serum response factor homolog is required for spore differentiation in *Dictyostelium*. Development 125:3801–3808

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791

Fischer A, Baum N, Saedler H, Theißen G (1995) Chromosomal mapping of the MADS-box multigene family in *Zea mays* reveals dispersed distribution of allelic genes as well as transposed copies. Nucleic Acids Res 23:1901–1911

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hoffmann U, Kroiher M (2001) A possible role for the cnidarian homologue of serum response factor in decision making by undifferentiated cells. Dev Biol 236:304–315

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The *Arabidopsis* Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29:102–105

Krogan NT, Ashton NW (2000) Ancestry of plant MADS-box genes revealed by bryophyte (*Physcomitrella patens*) homologues. New Phytol 147:505–517

Latchman DS (1998) Eukaryotic transcription factors. Academic Press, San Diego, CA

Liljegren SJ, Ferrándiz C, Alvarez-Buylla ER, Pelaz S, Yanofsky MF (1998) *Arabidopsis* MADS-box genes involved in fruit dehiscence. Flower News Lett 25:9–19

Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL, Yanofsky MF (2000) *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. Nature 404:766–770

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 26:1107–1115

Lupas A (1996) Coiled coils: New structures and new functions. Trends Biochem Sci 21:375–382

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Montag K, Salamini F, Thompson RD (1995) ZEMa, a member of a novel group of MADS box genes, is alternatively spliced in maize endosperm. Nucleic Acids Res 23:2168–2177

Montag K, Salamini F, Thompson RD (1996) The ZEM2 family of maize MADS box genes possess features of transposable elements. Maydica 41:241–254

Müller T, Vingron M (2000) Modeling amino acid replacement. J Comput Biol 7:761–776

Münster T, Pahnke J, Di Rosa A, Kim JT, Martin W, Saedler H, Theißen G (1997) Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. Proc Natl Acad Sci USA 94:2415–2420

Norman C, Runswick M, Pollock R, Treisman R (1988) Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell 55:989–1003

Ng M, Yanofsky MF (2001) Function and evolution of the plant MADS-box gene family. Nat Rev Genet 2:186–195

Passmore S, Elble R, Tye BK (1989) A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. Genes Dev 3:921–935

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF (2000) B and C floral organ identity functions require SEPALLATA MADS-box genes, Nature 405:200–203

Raes J, Van de Peer Y (1999) ForCon, a tool to automatically convert sequence alignment formats. EMBnet.news 6(1); http://www.psb.rug.ac.be/~jerae/ForCon/index.html

Raes J, Vandepoele K, Saeys Y, Simillion C, Van de Peer Y (2003) Investigating ancient duplication events in the Arabidopsis genome. J Struct Funct Genom (in press)

Riechmann JL, Meyerowitz EM (1997) MADS domain proteins in plant development. Biol Chem 378:1079–1101

Riechmann JL, Ratcliffe OJ (2000) A genomic perspective on plant transcription factors. Curr Opin Plant Biol 3(5):423–434

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M, Barrell B (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16:944–945

Saitou N, Nei M (1987) The neighbour-joining method: A new method for constructing phylogenetic trees. Mol Biol Evol 4:406–425

Sasaki T, Burr P (2000) International Rice Genome Sequencing Project: The effort to completely sequence the rice genome. Curr Opin Plant Biol 3:138–141

Scheffer U, Krasko A, Pancer Z, Müller WEG (1997) High conservation of the serum response factor within Metazoa: cDNA from the sponge *Geodia cydonium*. Biol J Linn Soc 61:127–137

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF (2002) MIPS Arabidopsis thaliana Database (MAtDB): An integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res 30:91–93

Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H (1990) Genetic control of flower development by homeotic genes in *Antirrhinum majus*. Science 250:931–936

Schwarz-Sommer Z, Hue I, Huijser P, Flor PJ, Hansen R, Tetens F, Lonnig WE, Saedler H, Sommer H (1992) Characterization of the Antirrhinum floral homeotic MADS-box gene deficiens: Evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO J 11:251–263

Shore P, Sharrocks AD (1995) The MADS-box family of transcription factors. Eur J Biochem 229:1–13

Simillion C, Vandepoele K, Van Montagu M, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci USA 99:13627–13632

Sommer H, Beltrán J-P, Huijser P, Pape H, Lönnig W-E, Saedler H, Schwarz-Sommer Z (1990) *Deficiens,* a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus:* The protein shows homology to transcription factors. EMBO J 9:605–613

Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Redaschi N, Stoehr P, Tuli, MA, Tzouvara K, Vaughan R (2002) The EMBL Nucleotide Sequence Database. Nucleic Acids Res 30:21–26

Strimmer K, von Haeseler A (1996) Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (*And other methods). Version 4. Sinauer Associates, Sunderland, MA

Theißen G (2001) Development of floral organ identity: Stories from the MADS house. Curr Opin Plant Biol 4:75–85

Theißen G, Saedler H (2001) Floral quartets. Nature 409:469–471

Theißen G, Becker A, Di Rosa A, Kanno A, Kim JT, Münster T, Winter K-U, Saedler H (2000) A short history of MADS-box genes in plants. Plant Biol 42:115–149

Theißen G, Münster T, Henschel K (2001) Why don't mosses flower? New Phytol 150:1–8

Van de Peer Y, De Wachter R (1997) Construction of evolutionary distance trees with TREECON for Windows: Accounting for variation in nucleotide substitution rate among sites. Comput Appl Biosci 13:227–230

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y (2002) The Automatic Detection of Homologous Regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res 12:1792–1801

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: Calibrating the family tree. Proc R Soc Lond Ser B Biol Sci 268:2211–2220

Wolf E, Kim PS, Berger B (1997) MultiCoil: A program for predicting two- and three-stranded coiled coils. Protein Sci 6:1179–1189

Yang Z (2000) Phylogenetic Analysis by Maximum Likelihood (PAML), version 3.0. University College London, London

Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM (1990) The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature 346:35–39

Yu J, et al. (2002) A draft sequence of the rice genome (*Oryza sativa L. spp. ndica*). Science 296:79–92

Yu YT, Breitbart RE, Smoot LB, Lee Y, Mahdavi V, Nadal-Ginard B (1992) Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. Genes Dev 6:1783–1798

Zhang H, Forde BG (2000) Regulation of *Arabidopsis* root development by nitrate availability. J Exp Bot 51:51–59