Evolution and Divergence of the MADS-Box Gene Family Based on Genome-Wide Expression Analyses

Rumiko Kofugi,* Naomi Sumikawa,† Misuzu Yamasaki,‡ Kimihiko Kondo,* Kunihiko Ueda,* Motomi Ito,* and Mitsuyasu Hasebe†§

*Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan; †National Institute for Basic Biology, Okazaki, Japan; ‡Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan; §Department of Molecular Biomechanics, The Graduate University for Advanced Studies, Okazaki, Japan

MADS-box genes encode transcription factors involved in various important aspects of development and differentiation in land plants, metazoans, and other organisms. Three types of land plant MADS-box genes have been reported. MIKC- and MIKC*-type genes both contain conserved MADS and K domains but have different exon/intron structures. M-type genes lack a K domain. Most MADS-box genes previously analyzed in land plants are expressed in the sporophyte (diploid plant body); few are expressed in the gametophyte (haploid plant body). Land plants are believed to have evolved from a gametophyte (haploid)-dominant ancestor without a multicellular sporophyte (diploid plant body); most genes expressed in the sporophyte probably originated from those used in the gametophyte during the evolution of land plants. To analyze the evolution and diversification of MADS-box genes in land plants, gametophytic MADS-box genes were screened using macroarray analyses for 105 MADS-box genes found in the Arabidopsis genome. Eight MADS-box genes were predominantly expressed in pollen, the male gametophyte; all but one of their expression patterns was confirmed by Northern analyses. Analyses of the exon/intron structure of these seven genes revealed that they included two MIKC*-type, one M-type, and four MIKC*-type MADS-box genes. Previously, MIKC*-type genes have been reported only from a moss and a club moss, and this is the first record in seed plants. These genes can be used to investigate the unknown ancestral functions of MADS-box genes in land plants. The macroarray analyses did not detect expression of 56 of 61 M-type MADS-box genes in any tissues examined. A phylogenetic tree including all three types of Arabidopsis MADS-box genes with representative genes from other organisms showed that M-type genes were polyphyletic and that their branch lengths were much longer than for the other genes. This finding suggests that most M-type genes are pseudogenes, although further experiments are necessary to confirm this possibility. Our global phylogenetic analyses of MADS-box genes did not support the previous classification of MADS-box genes into type I and II groups, based on smaller scale analyses. An evolutionary scenario for the evolution of MADS-box genes in land plants is discussed.

Introduction

Developmental mechanisms at the molecular level have been studied extensively in some model organisms in both metazoans and flowering plants (Walpert et al. 1998, p. 23). Transcription factors play critical roles in development at key nodes of gene networks by interacting with and regulating other genes, and their evolution is tightly linked with that of developmental mechanisms. Among the major evolutionary processes of genes is the duplication and subsequent divergence of genes and the creation of gene families, in addition to the functional divergence of orthologous genes after speciation (reviewed in Otto and Yong 2002). MADS-box genes are involved in several aspects of development in plants, and their evolution has been fundamental for morphological diversification in plants, especially in the reproductive organs (Hasebe 1999; Theissen et al. 2000).

MADS-box genes are characterized by approximately 180-bp DNA sequences, termed the MADS-box, which encodes a DNA-binding and dimerization domain, the MADS domain (reviewed in Shore and Sharrocks 1995). The MADS domain is conserved across a diverse range of organisms, from fungi, slime mold, and metazoans to land plants; in addition, putative MADS domains have been found in bacteria (Mushegian and Koonin 1996). MADS-box genes in eukaryotes other than land plants are divided into two groups, SRF-type and MEF2-type, based on phylogenetic analyses (Theissen, Kim, and Saedler 1996; Hasebe and Banks 1997), and their functions vary, even within the two groups. The SRF-type MADS-box genes are involved mainly in cell growth and differentiation in mammalian cells and mating type differentiation in budding yeast, whereas the MEF2-type genes function in muscle cell differentiation (reviewed in Theissen et al. 2000). In land plants, MADS-box genes were first identified as floral homeotic selector genes (Sommer et al. 1990; Yanofsky et al. 1990); all floral homeotic genes except one were shown to be MADS-box genes (reviewed in Weigel and Meyerowitz 1994). Two groups of land plant MADS-box genes are recognized: MIKC- and MIKC*-type MADS-box genes (Henschel et al. 2002). MIKC*-type genes are composed of a MADS (M) domain, an intervening (I) domain, a keratin-like (K) domain, and a C-terminal (C) domain, and they have been reported from a variety of land plants, including seed plants, pteridophytes, and bryophytes (reviewed in Theissen et al. 2000; Henschel et al. 2002). MIKC*-type genes have been reported only from lower land plants, including a moss (Henschel et al. 2002) and a club moss (Svensson, Johannesson, and Engstrom 2000), which are distinguished by their different intron/exon structure in the I domain, as compared with MIKC*-type genes (Henschel et al. 2002). Another type of MADS-box genes that lack
the K domain (called M-type MADS-box genes here) was recently found during the Arabidopsis Genome Sequencing project (Alvarez-Buylla et al. 2000b), but their expression patterns have not yet been studied. Alvarez-Buylla et al. (2000b) divided SRF-, MEF2-, M-, MIKC- C, and MIKC*-type MADS-box genes into two groups, type I and II, according to the phylogenetic analyses and signature amino acid residues in the MADS domain. This hypothesis should be assessed using more data, such as all the MADS-box genes in the Arabidopsis (Arabidopsis thaliana) genome.

Although the functions of MIKC*- and M-type MADS-box genes are unknown, MIKC*-type MADS-box genes have diverse functions. In addition to the floral homeotic genes, MIKC*-type genes function in floral transition (Michaels and Amasino 1999; Sheldon et al. 1999; Borner et al. 2000; Hartmann et al. 2000; Lee et al. 2000), cell differentiation of fruits (Liljegren et al. 2000; Liljegren and Yanofsky 2000; Vrebalov et al. 2002), and root architecture (Zhang and Forde 1998). In addition to the roles of MIKC*-type MADS-box genes in sporophyte (diploid) generation, some angiosperm and fern MIKC*-type MADS-box genes are expressed in gametophyte (haploid) generation. The detailed patterns of expression in fern gametophytes are unknown (Münster et al. 1997; Hasebe et al. 1998), but DEFH125 of snapdragon, ZnmADS2 of maize, and At3g57390/AGL18 of Arabidopsis are expressed in mature pollen, the male gametophyte, although their functions are unknown (Zachgo, Saedler, and Schwarz-Sommer 1997; Alvarez-Buylla et al. 2000a; Hasebe, Löhr, and Dresselhaus 2000). At3g57390/ AGL18 is expressed in the female gametophyte in addition to the male gametophyte, but its function has not been analyzed (Alvarez-Buylla et al. 2000a). Land plants are of monophyletic origin from a freshwater ancestor believed to be a charophycean green alga, whose plant body is gametophytic, except for a unicellular zygote (Graham, Cook, and Busse 2000), indicating that gametophytic and/or zygote-specific MADS-box genes are likely ancestral. One of the possible hypotheses for the evolution of MADS-box genes is that they were involved in unknown functions in gametophytes and/or zygotes, and were subsequently recruited for sporophyte development during the course of land plant evolution as the sporophyte came to dominate the gametophyte. The characterization of MADS-box genes expressed in gametophytes in land plants is important to the understanding of the origin and evolution of MADS-box genes.

The genome of Arabidopsis thaliana Columbia-0 has been almost entirely sequenced (The Arabidopsis Genome Initiative 2000), and it is predicted to contain 82 MADS-box genes (Riechmann 2002). Phylogenetic analyses and expression analyses of all the MADS-box genes in the Arabidopsis genome were performed to examine the evolution and diversity of plant MADS-box genes, and especially to evaluate (1) the previous classification of type I and II MADS-box genes and their evolutionary relationships, (2) whether gametophyte-specific MADS-box genes in addition to At3g57390/AGL18 exist in the Arabidopsis genome, and (3) the evolution of expression patterns in the MADS-box gene family.

Materials and Methods

Macroarray Analyses

Arabidopsis thaliana Columbia-0 plants were grown in an environmental chamber at 22°C under a 16-h-light (70 μmol/m²/s) / 8-h-dark regime. A partial DNA fragment of each candidate MADS-box gene underwent polymerase chain reaction (PCR) amplification using gene-specific primers (table 1 in the Supplementary Material online) with Arabidopsis genomic DNA as the template, and used as a probe. The correct amplifications of targeted fragments were confirmed by sequencing. Genomic DNA was isolated from mature rosette leaves with the Phytopure Plant DNA Extraction Kit (Amersham Bioscience KK, Tokyo, Japan), and treated with 40 μg/ml RNaseA (Sigma-Aldrich, St. Louis, Mo.) at 37°C for 30 min. The PCR-amplified fragments were electrophoresed on 0.9% or 1.5% agarose gels and blotted onto Hybond N+ nylon membranes (Amersham Bioscience). A PCR-amplified 760-bp DNA fragment of the ROC1 gene (Lippuner et al. 1994) was used as an internal control. The primers used are shown in table 1 of the Supplementary Material online.

Total RNA for a target was extracted using an RNeasy Plant Mini Kit (QIAGEN, Valencia, Calif.) from roots collected from juvenile plants grown for 20 days after sowing on an agar plate, juvenile plants with a few young rosette leaves and roots grown for 8 days after sowing, mature rosette leaves, inflorescences containing flower buds at stages less than stage 12, flowers at stages 12 and 13, pollen, and siliques 0 to 7 days after flowering. The stages of flower development followed that devised by Smith, Bowman, and Meyerowitz (1990). The agar plates used for root cultivation contained 2.2 g/L Murashige and Skoog culture medium basal salt (Wako Pure Chemical Industries, Osaka, Japan), 1 ml/L 1,000× Gamborg’s vitamin solution (Sigma-Aldrich), 1% sucrose, and 0.8% agar. Pollen was collected by vortexing flowers in TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) for 30 s, removing the flower debris, centrifuging at 10,000× g for 1 min, and filtering through a 50-μm nylon mesh (Kyoshin Riko Inc., Tokyo, Japan). Poly-A+ RNA, purified using DynaBeads (Dynal, Oslo, Norway), was used for RNA extracted from siliques.

Total RNA was reverse-transcribed to cDNA in the presence of [32P]-dCTP (Amersham Bioscience). Five μg of total RNA or poly-A+ RNA, purified using 5 μg of total RNA, and 10 pmol of adaptor primer were mixed, heated to 70°C for 10 min, and cooled on ice. Four μl of 5× first strand buffer (Invitrogen Japan, Tokyo, Japan), 2 μl of 0.1 M DTT, 0.5 μl of dNTPs (20 mM each dATP, dGTP, and dTTP), 0.48 μl of 100 mM dCTP, 4 μl of [32P]-dCTP, and 200 units of SuperScript II RNaseH reverse transcriptase (Invitrogen) were mixed and incubated at 42°C for 50 min. The remaining RNA was digested with 2 units of RNaseH (Invitrogen) at 37°C for 20 min. Labeled cDNA was further purified using a QiAquick PCR Purification Kit (QIAGEN).

The membranes were hybridized with labeled target cDNA at 65°C for 16 h with Church phosphate buffer (0.5 M Na2HPO4, 1 mM EDTA, 7% SDS), and washed twice with Church wash buffer (40 mM Na2HPO4, 1 mM EDTA, 1% SDS) for 20 min at 65°C (Church and Gilbert 1984).
Cloning and Northern Hybridization of cDNA of MADS-Box Genes Expressed in Pollen

Complementary DNA of pollen-specific MADS-box genes was amplified by PCR with a poly-T primer (Adaptor primer; Invitrogen), and two nested gene-specific primers located in the MADS domain, using pollen cDNA, and cloned into pPAM1 vector (Invitrogen). These clones were sequenced on an ABI PRISM 3100 sequencer using BigDye-terminator chemistry (Applied Biosystems, Foster City, Calif.).

Total RNA extracted from roots collected from juvenile plants grown for 20 days after sowing on an agar plate, juvenile plants with few young rosette leaves without roots grown for 8 days after sowing, mature rosette leaves, inflorescences containing flower buds at stages less than stage 12, flowers at stage 12 and 13, pollens, siliques 0 to 7 days after flowering, and ovules of stages less than stage 12, flowers at stage 12 and 13, and Kishino (1994). Gene names and their accession numbers used for phylogenetic analyses are shown in table 2 of the Supplementary Material online.

Results
Screening Arabidopsis MADS-Box Genes from the DNA Database

A translated BLAST (blastn) search (Altschul et al. 1997), using the 56 amino acid residues of the MADS domain, corresponding to positions 18–73 from the initiation Methionine codon of AGAMOUS (AG) (Riechmann, Ito, and Meyerowitz 1999) as a query, identified 197 entries from the non-redundant Arabidopsis database in the National Center for Biotechnology Information (NCBI; retrieved on 12 March 2002). Duplicate and fragmental sequences were eliminated, and 105 putative MADS domains were identified. Complementary DNA sequences had previously been determined for 36 of them and classified into 34 MIKCC- and 2 M-type MADS-box genes. Of the remaining 70 putative MADS domains, 65 were predicted to be genes in ATGV (Arabidopsis thaliana; microarray)

Phylogenetic Analyses

Alignment was performed using ClustalW version 1.8 (Thompson, Higgins, and Gibson 1994). Phylogenetic inferences were made using the Protdist, Neighbor, Consense and ClustalW programs in the PHYLIP version 3.572c software package (Felsenstein 1997) for the phylogenetic analysis.

The Neighbor-Joining (NJ) analyses. The evolutionary distances were calculated with the Protdist program using the Dayhoff and PAM matrix model (Dayhoff 1978), and a gene tree was constructed by the NJ method (Saitou and Nei 1987) using the program Neighbor. Statistical support for internal branches was estimated by bootstrap analyses with 100 replicates using the programs Seqboot and Consense.

For the maximum likelihood (ML) analyses, the NJdist and ProtML programs in MOLPHY, version 2.3b3, were used (Adachi and Hasegawa 1996). An NJ tree was obtained with NJdist, based on the ML distance under the JTT model (Jones, Taylor, and Thornton 1992), and used as the starting tree for a local rearrangement search using the program ProtML (Adachi and Hasegawa 1996). The local bootstrap probability of each branch was estimated by the resampling-of-estimated-log-likelihood (RELL) method (Kishino, Miyata, and Hasegawa 1990; Hasegawa and Kishino 1994). Gene names and their accession numbers used for phylogenetic analyses are shown in table 2 of the Supplementary Material online.
Fig. 1. (Continued)
thaliana genome view: http://www.ncbi.nlm.nih.gov/mapview/map_search?chr¼arabid.inf), and the AGI (Arabidopsis Genome Initiative) codes were assigned as their gene names. Five putative MADS domains were not predicted to be genes in the program, and were named AtMADS1 through AtMADS5.

The tblastn search using the K domain of AG, corresponding to positions 91–155 from the start codon (Riechmann, Ito, and Meyerowitz 1999) as a query, predicted a K domain in four of the 70 putative MADS-box genes, but not in the other 66 genes. The three genes (At5g51860, At5g51870, and At5g65060) with a K domain were classified as MIKCC-type genes, based on their predicted exon/intron structure by ATGV (fig. 1). At3g30270 was predicted to encode a MADS domain without a K domain. At3g20260 located 39 to At3g30270 was predicted to encode only a K domain, suggesting that these two genes are likely parts of a single MADS-box gene. We tentatively named them At3g30270*, which we classified as a MIKC*-type gene. In total, 38 MIKCC-type genes were found in the Arabidopsis genome. The remaining 62 MADS-box genes were classified as M-type MADS-box genes, together with the 5 AtMADS sequences (fig. 1). Six of the 67 M-type genes were revealed to be MIKC*-type genes, as mentioned below, and these six genes are shown as MIKC*-type genes instead of M-type genes in figure 1. The map positions of these putative Arabidopsis MADS-box genes were specified based on ATGV (fig. 2). The 105 putative MADS-box genes are distributed so that chromosomes I to V contain 33, 14, 12, 10, and 36, respectively.

Expression Patterns of all the Arabidopsis MADS-Box Genes

The spatial patterns of all the Arabidopsis MADS-box genes were assessed by macroarray analyses, in which PCR-amplified DNA fragments of each MADS-box gene were blotted onto nylon membranes, and mRNA extracted from tissues at different developmental stages was labeled and hybridized to the membranes. To prevent cross-hybridization, the PCR primers were located in regions without a MADS domain, except for four genes for which it was difficult to design appropriate primers (AtMADS1, AtMADS2, AtMADS5, and At5g27050) (fig. 1). When cross-hybridization was likely because of high similarity to other gene(s), a probe for one of the genes was designed: At1g31640 for At1g33070; AtMADS4 for At4g11250 and At5g65330; At5g26870 for At5g27580; and At5g26575/AGL34 for At5g27960. When the nucleotide sequences were close to 100% identical, a single primer set was designed for the genes, which should amplify these genes together: the primer set for At1g28460 should amplify At1g28450 and At1g29960, and that for At1g59810 should amplify At1g60040.
Signals for 44 of the 105 putative MADS-box genes were detected in some of the tissues examined, but no signals were detected for the remaining 61 genes (fig. 1). For the latter genes, except MIKC*-type genes, other probes covering the MADS domain were amplified and used for other macroarray analyses to exclude the possibility that the original probe region was not transcribed. The signals of five additional genes were detected with the new probes, and those of the other 56 were not detected (fig. 1). The cDNA sequences of two M-type genes (At5g27130/AGL39 and At4g36590/AGL40) were reported in the EST database. Our macroarray analyses did not detect the expression of either gene even after several repeated experiments.

Characterization of MADS-Box Genes Expressed in Pollen

MADS-box genes expressed in gametophytic tissue are candidates for MADS-box genes with ancestral functions because the ancestors of land plants are inferred to be gametophytic without multicellular sporophytic tissue; sporophytic generation was acquired during land plant evolution (Graham, Cook, and Busse 2000). Female gametophytic tissue, the embryo sac, is embedded in sporophyte tissue and is difficult to isolate, whereas male gametophytes are easy to isolate as pollen grains. The expression of eight MADS-box genes (At2g03060/AGL30, At1g18750, At1g77980, At1g22130, At2g34440/AGL29, At4g11880/AGL14, At3g57390/AGL18, and At5g26950/AGL26) was detected in pollen (fig. 1). Complementary DNA sequences of the previously isolated At4g11880/AGL14 and At3g57390/AGL18 genes were obtained from the DNA database, and the cDNA of other genes, except At5g26950/AGL26, was isolated by 3'-RACE, sequenced, and compared with genomic nucleotide sequences (accession numbers: AB094114, AB094116, AB094115, and AB094117). Complementary DNAs of the genes At1g69540 and At1g77950 were also isolated and sequenced (accession numbers: AB094112 and AB094113), because these genes were sisters to At2g03060/AGL30 and At1g77980, respectively, in preliminary phylogenetic analyses (data not shown). We could not isolate At5g26950/AGL26 cDNA, after several attempts, probably because there are genes closely related to At5g26950/AGL26 with more than 90% nucleotide sequence similarity, which prevented PCR amplification during the 3'-RACE analysis.

The At2g03060/AGL30, At1g18750, At1g22130, At1g69540, At1g77950, and At3g77950 genes in the database were predicted to be M-type MADS-box genes, and we did not detect K domains for these genes in the tblastn search using the AG K domain as a query. However, each cDNA, except At1g77950, encodes a weakly conserved K domain based on sequences obtained by 3'-RACE and has an exon/intron structure specific to MIKC*-type MADS-box genes (fig. 1). This is the first record of MIKC*-type genes in seed plants. The stop codon of the At2g03060/AGL30 gene is located in its fifth exon, which encodes the K domain. The At1g77950 gene is composed of a single exon encoding a MADS domain, and it lacks a K domain.

To confirm the expression patterns of these genes, Northern analyses were performed (fig. 3), and the results were concordant with those of the macroarray analyses. Strong signals were detected in pollen for the At2g03060/AGL30, At1g18750, At1g77980, At1g22130, At2g34440/AGL29, and At3g57390/AGL18 genes. Signals of At4g11880/AGL14 mRNA of different sizes were detected in roots and pollen. The mRNA detected in pollen was longer than that detected in root tissue. Weak expression of At3g57390/AGL18 was detected in ovules, which is concordant with previous in situ hybridization studies (Alvarez-Buylla et al. 2000a). Expression of ROC1 (Lippuner et al. 1994), used as an internal control, was...
detected in all the tissues tested, but the expression was much weaker in pollen than in other tissues. Similar amounts of ribosomal RNA from pollen and other tissues were blotted, indicating that ROC1 mRNA expression is repressed in pollen, and that equal amounts of RNA were blotted.

Phylogenetic Analysis of MADS-Box Genes

MADS-box genes were obtained from the nr data sets at NCBI, using the program tblastn (Altschul et al. 1997) on 6 October 2002, using the 56 residues of the MADS domain corresponding to amino acids 18–73 from the initiation methionine codon of AG (Riechmann, Ito, and Meyerowitz 1999) as a query; 805 MADS-box genes were obtained. To reduce the calculation time, flowering plant MADS-box genes were eliminated from the phylogenetic analyses, except for all the Arabidopsis and rice genes, and two genes reported to be expressed in pollen (ZnMADS2: Heuer, Lörz, and Dresselhaus 2000; DEFH125: Zachgo, Saedler, and Schwarz-Sommer 1997). The remaining 253 MADS-box genes were aligned (see fig. 1 of the online Supplementary Material), and genes containing more than five amino acid deletions in their MADS domains were not used for further analysis. Genes with more than one amino acid deletion in their MADS domain were also excluded when there was a sister gene supported with more than 80% bootstrap probability in a preliminary phylogenetic analysis (data not shown). These elimination processes were essential to increase the total number of amino acid residues usable for phylogenetic analyses. The alignment of 50 amino acid residues covering the MADS domain of the remaining 215 genes, consisting of 88 Arabidopsis, 31 rice, 1 maize, 1 snapdragon, 40 gymnosperm, 8 fern, 1 club moss, 7 moss, 31 metazoan, 1 slime mold, and 6 fungus MADS-box genes (see fig. 2 of the online Supplementary Material), was used for NJ analysis (Saitou and Nei 1987). Computation of the ML tree using MOLPHY did not finish within 30 days, so we could not obtain a ML tree.

The NJ tree of the MADS-box gene family (fig. 4) showed that SRF-type and MEF2-type MADS-box genes form distinct groups. RLM1 and SMP1 of budding yeast MADS-box genes formed a sister group, but they were not sister to either SRF-type or MEF2-type genes. Previously reported moss MIKC*-type MADS-box genes (Henschel et al. 2002) formed a group with the Arabidopsis MIKC*-type genes found in this study (fig. 1). The club moss MIKC*-type gene LAMB1 did not form a group with other MIKC*-type genes, MIKC*-type genes except LAMB1, were included in MIKC*-type genes, although bootstrap support was low. M-type MADS-box genes were polyphyletic, although some formed monophyletic groups with high bootstrap support.

MIKC*-type and MIKC*-type MADS-box genes contain a K domain in addition to the MADS domain; the phylogenetic relationships of these genes were analyzed further. From the original alignment, the MIKC*-type and MIKC*-type MADS-box genes were selected (see fig. 3 of the online Supplementary Material). The K domain of At2g03060/AGL30 was not well aligned with other MIKC*-type and MIKC*-type genes. At1g69540 has a 10-amino-acid deletion in its K domain. These two genes were eliminated from the analysis. Neighbor-joining and ML trees were calculated using 114 amino acid residues, and the topologies of both trees were mostly concordant; the ML tree is shown in fig. 5. Although the trees are unrooted, each MIKC*-type and MIKC*-type gene group formed a cluster. As previously reported (Theissen et al. 2000), MIKC*-type genes were divided into several subgroups supported with high bootstrap values. Arabidopsis MIKC*-type genes were classified into 12 monophyletic groups, but two orphan genes (At5g13790/AGL15 and At3g57390/AGL18) did not cluster with the other MIKC*-type genes. Each group diversified before the divergence of monocots and eudicots, and the monophony of each group was supported by more than 89% bootstrap probability. Each group is represented by the name of an Arabidopsis gene member (fig. 5). Of the new MIKC*-type genes found in the search of the Arabidopsis genome (At5g62165, At5g51860, At5g51870, and At5g65060), the first three genes clustered in the SOC group and the last one in the FLF group.

Discussion

MIKC*-Type MADS-Box Genes as New Members of Vascular Plant MADS-Box Genes

MIKC*-type MADS-box genes have previously been reported only from a moss and a club moss, and they were thought to be absent in ferns and seed plants, because no MIKC*-type genes were found in a computer prediction of all the Arabidopsis MADS-box genes (Henschel et al. 2002). The cloning of MADS-box genes, specifically expressed in pollen, and comparison of these cDNA and genomic sequences revealed that four Arabidopsis MADS-box genes (At2g03060/AGL30, At1g18750, At1g77980, and At1g22130) have the typical exon/intron structure of MIKC*-type genes, which is characterized by additional exons in the I region, as compared to MIKC*-type genes (fig. 1). These four genes formed a clade with the moss MIKC*-type genes in the phylogenetic analyses of all the Arabidopsis MADS-box genes, together with representa- tive genes from land plants and other organisms, thus supporting the postulate that these four genes are MIKC*-type genes (fig. 4). The phylogenetic analyses also suggest that At1g69540 is clustered with the four genes and likely MIKC*-type genes, although expression of this gene was not detected in pollen (fig. 1 and fig. 3). At1g69540 is expressed in inflorescences, and cloning of its cDNA, as well as comparison with the genomic sequence, confirmed that it is of the MIKC* type (fig. 1). At1g77980 formed a sister group with At1g77980 in preliminary analyses, but its expression was not detected in the macroarray or Northern analyses. A small cDNA fragment (320 bp) with a poly-A signal was isolated by 3’-RACE, however, indicating that At1g77980 is expressed at very low levels. As At1g77980 cDNA contains only a MADS domain, this gene is likely a M-type MADS-box gene. However, comparison of the 3’ flanking genomic region of At1g77980 cDNA with At1g77980 identified seven regions similar to the 2nd to 8th exons of At1g77980...
FIG. 4.—The NJ tree of the 215 MADS-box genes, including 88 Arabidopsis, 31 rice, 1 maize, 1 snapdragon, 40 gymnosperm, 8 fern, 1 club moss, 7 moss, 31 metazoan, 1 slime mold, and 6 fungus genes, using 50 amino acid residues. Because the amino acid sequences in the 50 residues are the same in the following genes, the gene names shown in this tree are changed as follows: MEF2A for CcaMEF2A, DreMEF2A, HsaMEF2A, MmuMEF2A, GgaMEF2A; MEF2C-1 for CcaMEF2C and DreMEF2C; MEF2C-2 for HsaMEF2C and MmuMEF2C; MEF2D for MmuMEF2D, RnuMEF2D, HsaMEF2D, XlaSL1, gMEF2D, and DreMEF2D; SRF-1 for HsaSRF, MmuSRF, GgaSRF, XlaSRF, and DreSRF; SRF-2 for JcoSRF and DmesSRF. This is an unrooted tree. The horizontal branch length is proportional to the estimated evolutionary distance. Branches with more than 80% bootstrap probability support are indicated as thicker lines. Arabidopsis genes whose expression was detected in the macroarray analyses are indicated in boldface. MIKC\(^{-}\)-type, MIKC\(^{+}\)-type, MEF2-type, and SRF-type MADS-box genes are indicated. GGM6 and LAMB1, which did not cluster with other members, are classified as MIKC\(^{-}\)- and MIKC\(^{+}\)-genes, respectively. Other unspecified genes are M-type MADS-box genes.
Fig. 5.—The maximum likelihood tree for the 111 MIKCC-type and 7 MIKC*-type MADS-box genes found by a local rearrangement search using 114 amino acid residues. The local bootstrap probability is shown on branches where available. Bootstrap values in the NJ tree using the same data set are shown below branches that exist in the NJ tree. This tree is an unrooted tree. The horizontal branch length is proportional to the estimated evolutionary distance. The genus from which the gene was isolated is indicated after the gene name, except for Arabidopsis. The brackets on the right indicate MIKC^-type and MIKC*-type MADS-box genes and the different subfamilies in the MIKC^-type MADS-box genes.
(83%, 71%, 79%, 69%, 74%, 84%, and 72%, respectively). This indicates that At1g77950 was originally a MIKC*-type MADS-box gene, which recently lost its I, K, and C domains by the acquisition of a poly-A' signal adjacent to the 3' region of the MADS domain. It is difficult to decide whether At1g77950 is functional, but its low level of expression suggests that it is likely a pseudogene.

Function of MIKC*-Type Genes

MIKC*-type genes have been reported in a moss and a club moss. Mosses branched from the common ancestor of vascular plants around 400 MYA (Kenrick and Crane 1997). The club moss lineage branched from other extant vascular plants, and then ferns with horsetails (Equisetales) and whisk ferns (Pteridales) diverged from the seed plant lineage (Pryer et al. 2001). Although mosses are an outgroup of Arabidopsis and club mosses, the club moss MIKC*-type gene LAMB1 clustered more basal to Arabidopsis and the moss Physcomitrella MIKC*-type genes (fig. 5) or did not cluster with other MIKC*-type genes in the analysis that used only MADS domain amino acid residues (fig. 4). Although the exon/intron structure of LAMB1 is similar to that of other MIKC*-type genes, LAMB1 likely belongs to another subgroup and is therefore omitted from the following discussion.

All the Arabidopsis MIKC*-type genes except At1g69540 and At1g77950 are specifically expressed in mature pollen, which is gametophytic. Although the expression patterns of moss MIKC*-type genes were not analyzed, these genes were isolated from gametophytic tissues (Henschel et al. 2002), indicating that these moss genes are expressed in gametophyte tissues. An Arabidopsis pollen grain consists of three cells, a vegetative cell and two sperm cells, whereas the Physcomitrella gametophyte is multicellular and forms a shoot-like structure with leaves and stems together with hypha-like protonemata. As angiosperm male gametophytes, such as pollen, are much reduced and the body plans and morphology of the gametophytes of angiosperms and mosses differ, it is difficult to speculate about a common function of MIKC*-type genes in gametophytes of Arabidopsis and Physcomitrella. Analyses of disruptive and over-expressor phenotypes in both Arabidopsis and Physcomitrella are in progress in our laboratories, and the results can be expected to provide further insight into the functions of MIKC*-type genes. At3g57390/AGL18 is a MIKC*-type gene, and it is expressed in both male and female gametophytes (Alvarez-Buylla et al. 2000a), which suggests that this gene has different functions from other MIKC*-type genes.

Classification of MADS-Box Genes Based on Their Phylogeny

Alvarez-Buylla et al. (2000b) inferred the phylogenetic relationships of 45 Arabidopsis MADS-box genes with 9 MEF2-type and 8 SRF-type MADS-box genes, and they found that MADS-box genes are divided into two groups: type I, including all SRF-type genes and 12 M-type genes (At1g65360/AGL23, At1g01530/AGL28, At1g34440/AGL29, At1g30360/AGL30, At1g26320/AGL33, At5g26757/AGL34, At5g26645/AGL35, At5g26625/AGL36, At1g65330/AGL37, At1g65300/AGL38, At5g27130/AGL39, and At4g36590/AGL40), and type II, including all MEF2-type genes and all MIKC*-type genes. They also found some conserved amino acids to support this grouping as synapomorphic characters and stated that an ancestral gene duplication before the divergence of plants and animals gave rise to these two types. In this study, we inferred a phylogenetic tree using 214 MADS-box genes, including all the putative Arabidopsis MADS-box genes (fig. 4); the result necessitates the revision of the type I and II classification. Both MIKC*- and MIKC*-type genes have K domains, indicating the monophyletic relationship of these groups, which is concordant with the phylogenetic tree in figure 4. MEF2-type genes do not form a sister group with MIKC*-type and MIKC*-type genes in figure 4, although the bootstrap support is insufficient and cannot reject the monophyly of MEF2-, MIKC*, and MIKC*-type genes with statistical confidence. The conserved amino acid residues were also used to support the monophyly of MEF2-type and MIKC*-type genes, and we examined whether these amino acid residues are still conserved in the large data set. The eight amino acid residues R (10th from the start methionine of AP1), Q (18th), V (19th), L (28th), V (37th), D (40th), F (48th), and L (54th) were conserved 8.0 ± 0.2 and 7.2 ± 1.1 in MEF2-type and MIKC*-type genes, respectively. However, these amino acid residues were not well conserved in MIKC*-type genes (5.8 ± 1.6). These results imply that we need to be careful to accept the monophyly of MEF2-, MIKC*, and MIKC*-type MADS-box genes.

The type I MADS-box genes are probably not monophyletic, according to the phylogenetic tree (fig. 4). At5g26757/AGL34, At5g26625/AGL36, At1g65300/AGL38, and At5g27090 had deletions in the MADS domain and were excluded from the phylogenetic analysis. These genes were closely related to At5g27960, At5g48670, At1g65330/AGL37, and At5g26950/AGL26, respectively, in preliminary phylogenetic analyses. At2g03060/AGL30 has been classified as a type I gene, but our comparison of its cDNA to genomic sequences showed that At2g03060/AGL30 is a MIKC*-type gene, even though there is a stop codon in the 5th exon, suggesting that AGL30 protein lacks the K and C domains, if translated. We could not assign I, K, and C domains for these genes in the analysis that used only MADS domain amino acid residues (fig. 4), indicating that the grouping of type I MADS-box genes is inappropriate.

Because the original grouping of type II MADS-box genes did not include MIKC*-type genes, and because the monophyly of type I genes is not supported, it is better to use the original classification based on the domain structure: SRF-, MEF2-, MIKC*, MIKC*, and M-types. Each of the first four types likely has a single origin, but the last does not. The phylogenetic relationships among the first four groups were not resolved with statistical confidence, and future studies are necessary. RLM1 and SMP1 of yeast MADS-box genes did not cluster with...
either SRF- or MEF2-type genes, suggesting that these genes form a new group of MADS-box genes.

Origin and Evolution of M-Type MADS-Box Genes

The most unexpected result of this study was that most M-type MADS-box genes were not expressed in any tissues examined. We used tissues that covered most developmental stages; most M-type genes probably do not function in regular development in Arabidopsis. Furthermore, the branch lengths of most M-type MADS-box genes are relatively longer than those of other types of genes, suggesting that the functional constraints on most M-type genes are not as strong as in other genes, and that M-type genes are pseudogenes. We should be careful in regarding all M-type genes as pseudogenes and non-functional, because some M-type genes are expressed (fig. 1), and the amino acids are conserved in some clades (e.g., the clade including \( At4g56590/AGL40 \) in fig. 4).

The conspicuous shared character of M-type genes is that these genes are intron-less, or have few introns, based on the computer prediction in the MIPS site (fig. 1), which suggests that M-type genes are functional retrogenes or processed retrosequences (Weiner, Deininger, and Efratistiadi 1986). Retrogenes are reverse-transcribed from mRNA and usually have poly-A sequences at their 3' region. Such poly-A sequences were not found within 1 kb of the predicted stop codons, except in \( At5g49490 \) (data not shown), indicating that most M-type genes are probably not retrosequences or retrogenes. There are 25 adenine nucleotides at the 183rd site from the stop codon of \( At5g49490 \), suggesting that it is a retrogene. Another possible origin of M-type MADS-box genes is that they are pseudogenes that originated from MIKCC- or MIKC*-type genes. The K domain is characterized by regular spacing of conserved hydrophobic residues and is inferred to form amphipathic helices (Shore and Sharrocks 1995). The MADS domain is composed of ca. 60 conserved amino acid residues, and might still be distinguished after substantial mutations accumulate. The MADS domain is distinguishable even after a stop codon is formed in the MADS domain as in \( At3g32313, At3g32316, TSN9.10, At3g32371, \) and \( At3g31902 \). By contrast, approximately 10 amino acid residues characterized only by their hydrophobicity, which are spaced throughout the K domain, are weakly conserved, and it would be difficult to recognize a K domain with the same number of accumulated mutations. In other words, once MIKCC-type or MIKC*-type genes become pseudogenes, the K domain would be lost faster than the MADS-domain. \( At1g77950 \) might be an example of this process; it was transcribed as an M-type gene, but the exon/intron structure of MIKC*-type genes in region I remained in the genomic sequence. No K domain was found, even with a careful search of the amino acid sequences in the region corresponding to the K domain of \( At1g77950 \) (data not shown). \( At1g77950 \) probably originated in a duplication of its sister gene, \( At1g77980 \), which is expressed and is likely functional. If M-type MADS-box genes are pseudogenes that originated by duplications of functional MIKCC-type or MIKC*-type genes, it is curious that most M-type MADS-box genes do not have corresponding sister genes, which should form sister groups in the phylogenetic tree (fig. 4). One possible explanation is that after substantial mutations randomly accumulate in a pseudogene, following gene duplication, the pseudogene does not form a sister group with its sister gene, and likely clusters with other pseudogenes as a result of long-branch-attraction (Felsenstein 1978). Note that some M-type MADS-box genes are duplicated repeatedly, and the numbers increase after the original pseudogenes fully diverge from their ancestor (e.g., the clade including \( At5g49490 \)). Once MIKCC-type or MIKC*-type genes become pseudogenes, how long can we recognize these pseudogenes as MADS-box genes? There are approximately two synonymous nucleotide substitutions per site between eudicots and monocots (Wolfe, Sharp, and Li 1989). When nucleotides are substituted under a Poisson distribution (Sokal and Rohlf 1995, p. 84), approximately 86% of their nucleotides should be substituted more than once after the divergence of eudicots and monocots, approximately 200 MYA, implying that most amino acids in the MADS domain were probably substituted more than once. The BLAST search that we used in this study obtained MADS-box genes containing more than 20% amino acid sequences identical to the \( At4g18960/AG \) gene. A rough calculation indicates that a pseudogene that originated at the time of eudicot/monocot divergence would not be included in this study. If the M-type genes detected in this study are pseudogenes, they originated after the divergence of eudicots and monocots.

The expression patterns of previously characterized MADS-box genes (reviewed in Johansen et al. 2002) are concordant with the results shown in figure 1. This indicates that our method is highly reliable. However, the sensitivity of our method of detecting the expression of mRNA is not as high as that of reverse-transcription-PCR (RT-PCR), and we cannot deny the possibility that M-type genes whose expression was not detected in this study may be expressed at very low levels and be functional. We performed RT-PCR for all genes whose expression was not detected in figure 1, and detected the expression of the following genes (data not shown): \( At1g17310, At1g28460, At1g31140, At1g31630, At1g46408, At1g47760, At1g48150, At1g54760, At1g59810, At1g60880, At1g60920, At1g65300, At1g65330, At1g65360, At2g24840, At2g26320, At2g26880, At2g28700, At2g40210, At3g05860, At3g06656, AtMADS3, AtMADS4, At4g36590, AtMADS5, At5g06500, At5g26625, At5g26645, At5g27130, At5g27580, At5g27960, At5g38620, At5g39750, At5g40120, At5g40210, At5g48670, At5g49420, At5g49490, At5g55690, At5g58890, At5g60440, and At5g65330. Further studies using loss-of-function mutants of these genes are necessary to analyze whether these genes are functional. It should also be noted that several M-type MADS-box genes might be induced by environmental factors.

Evolution of the Expression Patterns of MIKCC-Type MADS-Box Genes

The expression patterns of all the MADS-box genes in the Arabidopsis genome were analyzed using the
macroarray technique (fig. 1), and those of eight genes were confirmed by Northern analysis (fig. 3). The expression patterns of most MIKC\(^C\)-type genes are concordant with those previously reported (Alvarez-Buylla et al. 2000a), except AGL19. Weak expression of AGL19 was detected in rosette leaves, but not in roots, although Alvarez-Buylla et al. (2000a) detected AGL19 expression in roots in Northern and in situ RNA analyses. This difference is probably caused by the probe used in the previous study, which was a whole cDNA, including the MADS domain. The nucleotide sequence of the AGL19 MADS-box is 90% identical to that of At4g11880/AGL14, which is expressed in roots, suggesting cross-hybridization of the probe to At4g11880/AGL14 mRNA in roots.

Our phylogenetic analyses confirmed a previous result, which was that MIKC\(^C\)-type MADS-box genes are divided into several monophyletic groups, and that members of each group are similarly expressed (reviewed in Theissen et al. 2000). The evolution of the expression patterns should be discussed according to the phylogenetic relationships of genes with reliable statistical confidence. In the case of a MADS-box gene family, inter-subfamily relationships are not inferred with statistical support, whereas the monophyletic relationship of each subfamily is usually supported with high bootstrap values (Theissen et al. 2000; Johansen et al. 2002). Some authors have stated that MADS-box genes expressed in vegetative organs predate those in reproductive organs; in other words, reproductive MADS-box genes originated from duplicated vegetative MADS-box genes (Purugganan 1997; Alvarez-Buylla et al. 2000a). However, the phylogenetic relationships used in those studies have not been reproducible in larger-scale analyses (Theissen et al. 2000; Johansen et al. 2002), including this study, and we must state that the evolution of expression patterns in MIKC\(^C\)-type genes is still ambiguous. Further analyses including a larger number of genes from other taxa may give more robust phylogenetic trees to infer the evolution of expression patterns in the MADS-box gene family.

Evolution of Gametophytic and Sporophytic MADS-Box Genes

Based on fossil data, land plants emerged approximately 480 MYA (Kenrick and Crane 1997), and both morphological and molecular data support the sister relationship between extant land plants and the freshwater green algae, charophytes (Graham, Cook, and Busse 2000; Graham and Wilcox 2000, p. 497; Karol et al. 2001). Land plants are composed of four monophyletic groups: vascular plants, mosses, liverworts, and hornworts. The phylogenetic relationships of these groups are still controversial, because of discrepancies among the phylogenetic relationships inferred from morphological and molecular data and among the different molecular data sets used (Nishiyama and Kato 1999; Qiu and Palmer 1999; Nickrent et al. 2000). However, there is consensus that bryophytes, which include the last three groups, are more basal than vascular plants in the land plant phylogenetic tree. Bryophytes are characterized by their dominant gametophytic plant body and their sporophytes epiphytic to gametophytes. The sporophytes of bryophytes do not differentiate shoots and roots, unlike vascular plants, but form only a sporangium with tissue connecting them to the gametophytes. In charophytes, the first cell division of the zygote is meiosis, and their plant body is gametophytic. A sister group of the charophyte-land plant clade is the zygnematalean green algae, which have a life cycle similar to that of charophytes (Graham and Wilcox 2000, p. 503). These features suggest that the common ancestor of charophytes and land plants had a life cycle similar to that of charophytes, and that the common ancestor of bryophytes and vascular plants was gametophyte-dominant. According to this evolutionary scenario of land plants, all the genes used in the development of the diploid plant body should have originated from those used in gametophytes. Expression analyses of all the Arabidopsis MADS-box genes showed that most MIKC\(^C\)-type genes, except At3g57390/AGL18, were not expressed, or were very weakly expressed, in the male gametophyte, whereas all MIKC*-type genes, except At1g69540, were specifically expressed in the male gametophyte. This suggests that At3g57390/AGL18 and the MIKC*-type genes retain their original function in gametophytes from the common ancestor of land plants and charophytes. The characterization of these genes will give further insight into the original function and subsequent evolution of MADS-box genes.

DEFH125 of Antirrhinum and ZmMADS2 of Zea MADS-box genes are expressed in pollen (Zachgo, Saedler, and Schwarz-Sommer 1997; Heuer, Lörz, and Dresselhaus 2000). Phylogenetic analyses including these taxa showed that both genes formed a clade with Antirrhinum, Zea, and other MADS-box genes (Purugganan 1997; Alvarez-Buylla et al. 2000a). The phylogenetic relationships used in those studies have not been reproducible in larger-scale analyses (Theissen et al. 2000; Johansen et al. 2002), including this study, and we must state that the evolution of expression patterns in MIKC\(^C\)-type genes is still ambiguous. Further analyses including a larger number of genes from other taxa may give more robust phylogenetic trees to infer the evolution of expression patterns in the MADS-box gene family.

Supplementary Material

The following information is available online: A list of the primers used for the macroarray analyses (table 1 of
the online Supplementary Material), a list of genes, accession numbers, and organism names used for phylogenetic analyses (table 2 of the online Supplementary Material), the alignments of MADS-box genes collected from the DNA database (fig. 1 of the online Supplementary Material), and the alignments used for the phylogenetic analyses in fig. 4 (fig. 2 of the online Supplementary Material) and fig. 5 (fig. 3 of the online Supplementary Material).

Acknowledgments

We thank Tomomichi Fujita, Tomoaki Nishiyama, and members of Hasebe laboratory for their valuable comments on the manuscript. C. S. Gasser kindly provided the ROC1 clone. The computations were done on a SGI Origin 2000 in the Computer Lab of National Institute for Basic Biology (NIBB). The NIBB Center for Analytical Instruments provided the sequence facilities. This research was supported in part by grants from the Ministry of Education, Culture, Sports, Science and Technology—MET (M.H.), and Japan Society for the Promotion of Science—JSPS (M.H., R.K., K.U., M.I.).

Literature Cited


