Journal of Integrative Plant Biology 2007, **49** (6): 927–939

Duplication and Divergence of Floral MADS-Box Genes in Grasses: Evidence for the Generation and Modification of Novel Regulators

Guixia Xu[1, 2] and Hongzhi Kong[1*]

([1]*State Key Laboratory of Systematic and Evolutionary Botany*, *Institute of Botany*, *the Chinese Academy of Sciences*, Beijing 100093, China;
[2]*Graduate University of the Chinese Academy of Sciences*, Beijing 100039, China)

Abstract

The process of flowering is controlled by a hierarchy of floral genes that act as flowering time genes, inflorescence/floral meristem identity genes, and/or floral organ-identity genes. The most important and well-characterized floral genes are those that belong to the MADS-box family of transcription factors. Compelling evidence suggests that floral MADS-box genes have experienced a few large-scale duplication events. In particular, the pre-core eudicot duplication events have been considered to correlate with the emergence and diversification of core eudicots. Duplication of floral MADS-box genes has also been documented in monocots, particularly in grasses, although a systematic study is lacking. In the present study, by conducting extensive phylogenetic analyses, we identified pre-Poaceae gene duplication events in each of the *AP1*, *PI*, *AG*, *AGL11*, *AGL2/3/4*, and *AGL9* gene lineages. Comparative genomic studies further indicated that some of these duplications actually resulted from the genome doubling event that occurred 66–70 million years ago (MYA). In addition, we found that after gene duplication, exonization (of intron sequences) and pseudoexonization (of exon sequences) have contributed to the divergence of duplicate genes in sequence structure and, possibly, gene function.

Key words: divergence; *Oryza sativa*; gene duplication; genome doubling event; grasses; MADS-box genes.

Xu G, Kong H (2007). Duplication and divergence of floral MADS-box genes in grasses: Evidence for the generation and modification of novel regulators. *J. Integr. Plant Biol.* **49**(6), 927–939.

Available online at www.blackwell-synergy.com/links/toc/jipb, www.jipb.net

Flowers, the specialized reproductive structures of angiosperms, are presumably the most important morphological innovations of flowering plants. Developmental and genetic studies have shown that the process towards forming a flower is controlled by a complex network of genes (nodes) and interactions (lines; Zhao et al. 2001; Soltis et al. 2002; Kaufmann et al. 2005). Among the genes in this network, the best known are MADS-box genes, the protein products of which are MADS domain-containing transcription factors (Yanofsky et al. 1990; Ma et al. 1991; Becker and Theissen 2003). By regulating the expression of other genes, these MADS-box genes can act as flowering time genes, inflorescence/floral meristem identity genes, and/or floral organ identity genes (Zhao et al. 2001; Soltis et al. 2002; Kaufmann et al. 2005). In particular, studies of floral organ identity genes in several model plants (such as *Arabidopsis thaliana* and *Antirrhinum majus*) have led to the proposal of the famous “ABC model” for floral development (Coen and Meyerowitz 1991). According to this model, the development of the four different floral organ types is determined genetically by three classes of gene function (A, B, and C): A

Received 9 Jan. 2007 Accepted 12 Mar. 2007

Supported by the National Natural Science Foundation of China (30530090, 30470116 and 30121003) and Institute of Botany, the Chinese Academy of Sciences.

Publication of this paper is supported by the National Natural Science Foundation of China (30624808).

*Author for correspondence.

Tel: +86 (0)10 6283 6489;

Fax: +86 (0)10 6259 0843;

E-mail: <hzkong@ibcas.ac.cn>.

doi: 10.1111/j.1672-9072.2007.00502.x

function alone specifies sepals in the first whorl; A and B specify petals in the second whorl; B and C specify stamens in the third whorl; and C alone specifies carpels in the fourth whorl (Coen and Meyerowitz 1991; Ma and dePamphilis 2000). Later, a fourth (D) gene function was recognized, which is indispensable for the formation of ovules within the carpel (Angenent et al. 1995; Colombo et al. 1995; Angenent and Colombo 1996). In *Arabidopsis*, A-function genes are represented by *APETALA1* (*AP1*) and *APETALA2* (*AP2*), B-function genes by *APETALA3* (*AP3*) and *PISTILLATA* (*PI*), C-function genes by *AGAMOUS* (*AG*) and miR172, and D-function genes by *SHATTERPROOF1* and *2* (*SHP1*, *2*) and *SEEDSTICK* (*STK*) (=*AGAMOUS-LIKE11* (*AGL11*)) (Coen and Meyerowitz 1991; Zahn et al. 2006a). Except for *AP2* and miR172, all these genes are MIKC-type MADS-box genes, the proteins of which can form quaternary complexes together with the protein of the SEPALATA1-4 genes (*SEP1-4*; also known as *AGL2*, *-4*, *-9*, and *-3*; these are the so-called E-function genes): sepals are controlled by the "AP1-AP1-SEP-SEP", petals by the "AP1-AP3-PI-SEP", stamens by the "AP3-PI-AG-SEP", carpels by the "AG-AG-SEP-SEP", and ovules by the "AG-STK-SHP-SEP" complexes (Theissen 2001; Theissen and Saedler 2001; Melzer et al. 2006).

Phylogenetic studies have suggested that floral MADS-box genes were derived from a single ancestral gene approximately 650 million years ago (MYA) and were the products of repeated gene duplications (Purugganan 1997; Nam et al. 2003). These duplications, usually accompanied by modifications in coding and/or regulatory regions, have also been shown to correlate with the occurrence of major plant groups (Irish 2003; Irish and Litt 2005; Kramer and Hall 2005; Zahn et al. 2005). For example, just before the occurrence of angiosperms, an ancient duplication event has occurred in each of the *AP3*/*PI*, *AG*/*AGL11*, and *SEP* clades to generate the *AP3* and *PI*, *AG* and *AGL11*, and *AGL2/3/4* and *AGL9* gene lineages, respectively (Kramer et al. 1998, 2004; Kim et al. 2004; Zahn et al. 2005, 2006b). Similarly, within each of the *AP1*, *AP3*, *AG* and *AGL2/3/4* lineages, additional gene duplications have occurred before the diversification of extant core eudicots to create the eu*FUL*, eu*AP1* and *AGL79*, eu*AP3* and *TM6*, eu*AG* and *PLE*, and *AGL2*, *AGL3* and *FBP9* lineages, respectively (Kramer et al. 1998, 2004; Litt and Irish 2003; Kim et al. 2004; Stellari et al. 2004; Zahn et al. 2005; Shan et al. 2007). Expression and functional analyses further indicated that phylogenetically closely related paralogs from each species tend to have similar but differentiated expression patterns, suggesting that they perform related but distinct functions (Kramer et al. 1998; Lamb and Irish 2003; Vandenbussche et al. 2003). More interestingly, several floral organ identity genes, such as *Arabidopsis AP1* and *AP3*, appear to be novel genes generated in the pre-core eudicot duplication events because, due to frameshift mutations, the C-termini of their proteins are no longer homologous to that of the paleoAP1 and paleoAP3 proteins, respectively (Kramer and Irish 2000; Litt and Irish 2003). Considering that the core eudicots are a successful angiosperm group with very elaborate and highly derived floral structures, many people have suggested that the origin of the core eudicot-specific floral structures may have been caused by the inclusion of more regulatory genes (nodes) and interactions (lines) into the already well-organized regulatory network for floral development in basal eudicots (Irish 2003, 2006; Kramer and Hall 2005; Zahn et al. 2005; Kramer and Zimmer 2006).

Duplications of floral MADS-box genes have also been documented in monocots, an important group that comprises approximately 22% of angiosperm species (Litt and Irish 2003; Zahn et al. 2005). For example, in the *AP1* subfamily, at least two large-scale duplication events have been recognized, one before the split of the Asparagales and commelinids and the other within the Poales, probably prior to the origin of the Poaceae, a family that contains rice, maize, wheat, and other grasses (Shan et al. 2007). As a result, although the latest common ancestor (LCA) of monocots and eudicots possessed only one *AP1* subfamily member, most (if not all) grass species have three types (i.e. *OsMADS14*, *OsMADS15*, and *OsMADS18*) of *AP1* genes (Litt and Irish 2003; Whipple and Schmidt 2006). Similar situations were found in the *PI*, *AG*, *AGL11*, *AGL2/3/4*, and *AGL9* lineages; the multiple grass genes in each of these lineages seem to have been derived from a single ancestral copy in the LCA of monocots and eudicots (Kim et al. 2004; Kramer et al. 2004; Zahn et al. 2005, 2006b; Figure 1). However, due to the lack of a detailed analysis, it is still not known when the duplication events in each subfamily happened, nor is it clear whether these events correspond to the genome-wide duplication event that occurred before the origin of the Poaceae 66-70 MYA (Vandepoele et al. 2003; Paterson et al. 2004; Wang et al. 2005; Yu et al. 2005). In the present study, by performing extensive phylogenetic analyses on floral MADS-box genes from monocots, we reveal pre-Poaceae gene duplication events in each of the *AP1*, *PI*, *AG*, *AGL11*, *AGL2/3/4*, and *AGL9* lineages. In addition, we found that a few of these duplications can be explained by the genome doubling event, and that after gene duplication, exonization (of intron sequences) and pseudoexonization (of exon sequences) contributed to the divergence of floral MADS-box genes in both structure and function.

Results

Duplication of floral MADS-box genes in grasses

The 16 rice genes included in the present study belong to seven major lineages: four in the *AP1* lineage, three in the *AGL2*/*3*/*4* lineage, two in each of the *PI*, *AG*, *AGL11* and *AGL9* lineages,

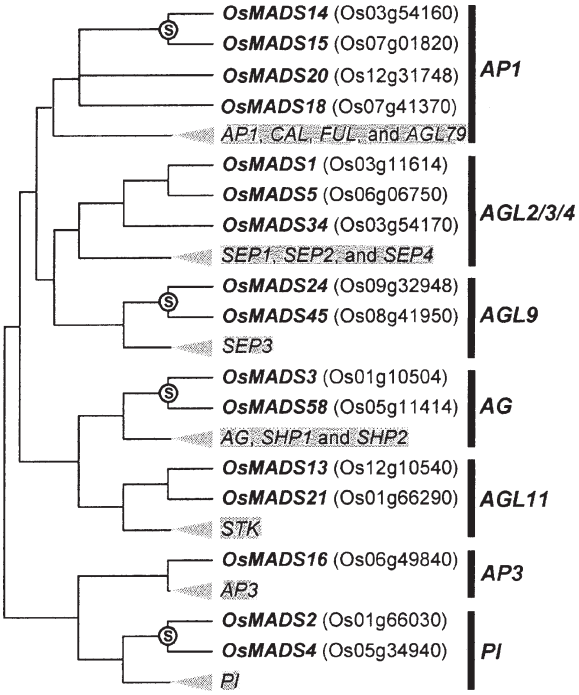


Figure 1. Summary of phylogenetic relationships of floral MADS-box genes from rice and *Arabidopsis*.

Shaded triangles represent the *Arabidopsis* genes. Segmental duplications are indicated by circled "S".

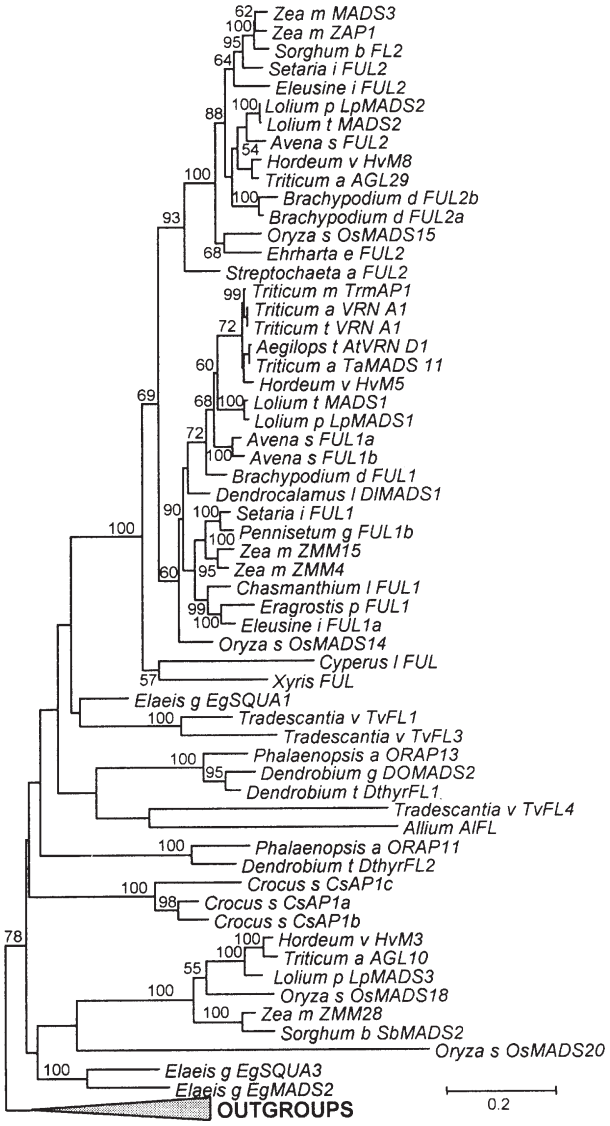


Figure 2. Maximum-likelihood (ML) tree of *AP1*-like genes from monocots.

This tree was based on the DNA sequence analysis. Bootstrap values greater than 50% are shown at nodes.

and one in the *AP3* lineage. Thus, except for the *AP3* lineage, all other gene lineages have expanded during the evolution of monocots. In particular, phylogenetic analyses for each of these gene lineages suggest that one (or two) pre-Poaceae gene duplication(s) must have happened in each of the *AP1*, *PI*, *AG*, *AGL11*, *AGL2/3/4* and *AGL9* lineages. This can be seen from the fact that many rice genes tend to form grass-specific clades with their putative orthologs from other grass species (Figures 2–5). For example, the two *AP1*-lineage members, *OsMADS14* and *OsMADS15*, appear to have been generated through a relatively recent gene duplication that occurred before the diversification of the extant Poaceae but after the split between the Poaceae-Restionaceae-Flagellariaceae clade and the Xyridaceae-Juncaceae-Cyperaceae clade, because each gene forms a separate clade with its putative orthologs from other grasses such as *Zea*, *Sorghum*, *Setaria*, *Avena*, *Lolium*, and *Hordeum*, with genes from *Xyris* (Xyridaceae) and *Cyperus* (Cyperaceae) resolved as the outgroups of both clades. Similarly, within the *PI*, *AG*, *AGL11*, *AGL2/3/4* and *AGL9* lineages, the duplication events that gave rise to *OsMADS2* and *OsMADS4*, *OsMADS3* and *OsMADS58*, *OsMADS13* and *OsMADS21*, *OsMADS1* and *OsMADS5*, and *OsMADS24* and *OsMADS45*, respectively, all seem to have occurred before, and just before, the diversification of the Poaceae.

In addition to the pre-Poaceae duplications, the present study revealed two earlier gene duplication events, one in each of the *AP1* and *AGL2/3/4* lineages. In the *AP1* lineage, although the exact position of *OsMADS20* is still controversial, the duplication event that gave rise to the *OsMADS14/15* and *OsMADS18* clades may have happened before the diversification of extant

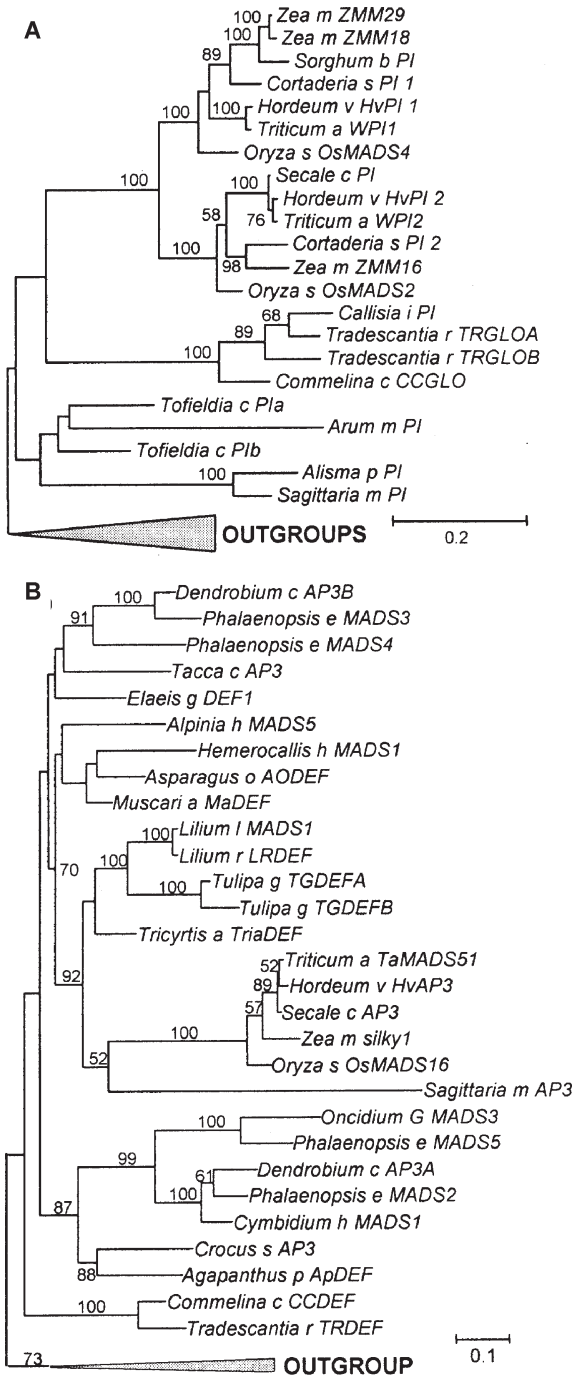


Figure 3. ML trees for **(A)** *AP3*-like genes and **(B)** *PI*-like genes from monocots.

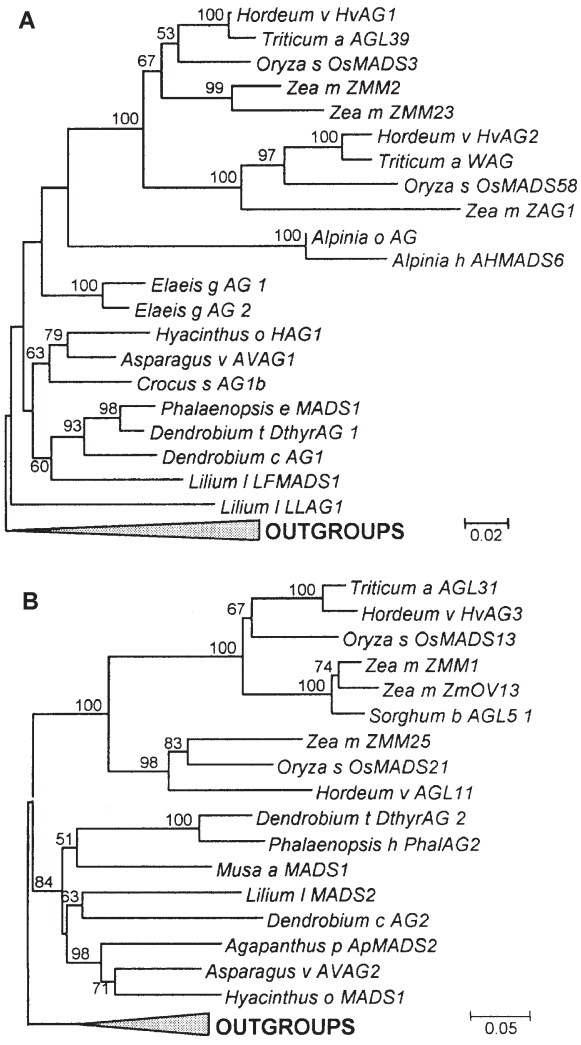
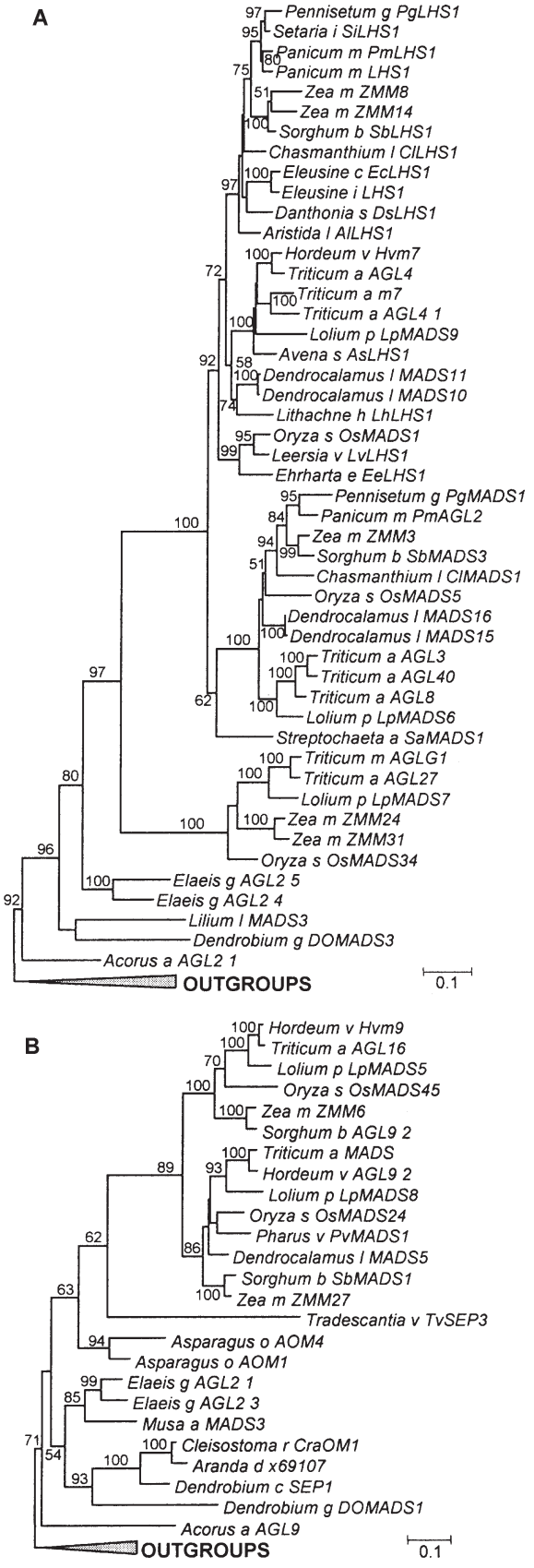


Figure 4. ML trees for **(A)** *AG*-like genes and **(B)** *AGL11*-like genes from monocots.

commelinids (Figure 2). This result, although not well supported, is consistent with our recent phylogenetic studies on the *AP1* subfamily, which suggests that the *OsMADS14/15-OsMADS18* duplication may have happened before the divergence of commelinids (including grasses) from the Asparagales (Shan et al. 2007). The *OsMADS1/5-OsMADS34* duplication in the *AGL2/3/4* lineage, however, seems to have occurred within the commelinids clade, probably after the split of Arecales from other commelinids, if the positions of the two *Elaeis* genes are taken into consideration.

Because the pre-Poaceae duplication events can be inferred in six of the seven floral MADS-box gene lineages, we wonder whether these duplication events correspond to a single genome doubling event. To understand this, we checked to see whether any of the rice gene pairs were actually generated through segmental duplications. Indeed, we found that the duplication events that gave rise to *OsMADS14* and *OsMADS15*, *OsMADS24* and *OsMADS45*, *OsMADS3* and *OsMADS58*, and



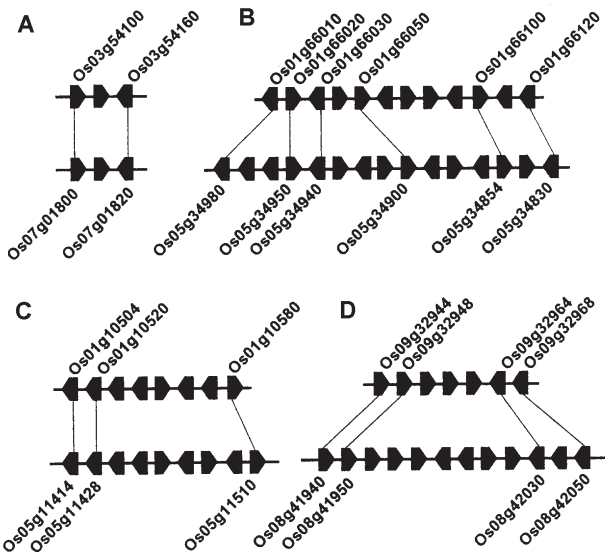


Figure 6. Segmental duplications of the **(A)** *OsMADS14* (Os03g54160) and *OsMADS15* (Os07g01820); **(B)** *OsMADS2* (Os01g66030) and *OsMADS4* (Os05g34940); **(C)** *OsMADS3* (Os01g10504) and *OsMADS58* (Os05g11414), and **(D)** *OsMADS24* (Os09g32948) and *OSMADS45* (Os08g41950) genes.

Putative paralogous genes are linked with thin lines.

OsMADS2 and *OsMADS4*, respectively, can be explained by segmental duplication (Figure 6). More interestingly, the synonymous distance (K_S) values between these duplicate genes (0.681±0.123 for *OsMADS14* and *OsMADS15*, 0.612±0.108 for *OsMADS24* and *OsMADS45*, 0.845±0.149 for *OsMADS3* and *OsMADS58*, and 0.726±0.129 for *OsMADS2* and *OsMADS4*) are very close to each other, suggesting that the four independent duplication events may have happened simultaneously, or nearly so. In addition, because these K_S values are also close to the mean K_S values (0.631 to 0.688) estimated for the duplicated segments in the rice genome (Yu et al. 2005), we believe that the aforementioned duplication events may have been caused by the genome doubling event before the origin of the Poaceae 66-70 MYA (Vandepoele et al. 2003; Paterson et al. 2004; Wang et al. 2005; Yu et al. 2005).

Divergence between paralogous genes

To understand the evolutionary fate of the duplicate genes, we have also compared the exon/intron structure of the aforementioned

←

Figure 5. ML trees for **(A)** *AGL2/3/4*-like genes and **(B)** *AGL9*-like genes from monocots.

duplicate gene pairs. To our surprise, we discovered obvious differences between such paralogous gene pairs as *OsMADS3* and *OsMADS58*, *OsMADS13* and *OsMADS21*, and *OsMADS5* and *OsMADS1*. *OsMADS3* and *OsMADS58*, the two *AG* lineage members in rice, are 65% and 79% identical at the protein and DNA levels, respectively. At the protein level, OsMADS3 shares with many other *AG* lineage members the highly conserved "AG II" motif ("YAHQLQPTTLQLG"; Kramer et al. 2004) at the C-terminal region, whereas OsMADS58 has a partially different C-terminal end that is not homologous to those of the others. By comparing the structure of the two genes, we found that *OsMADS3* has nine exons, whereas *OsMADS58* has eight (Figure 7A). The eighth exon of *OsMADS3* is 176 bp long and matches very well to the first half (1–172 of 231 bp) of the eighth exon of *OsMADS58*. However, the ninth exon of *OsMADS3*, which is 55 bp in length, does not match to the second half (173–231 of 231 bp) of the eighth exon of *OsMADS58*, but shares significant similarity to a 56 bp long intergenic region downstream of the eighth exon of *OsMADS58* (Figure 7B). This observation, together with the fact that the second half of the eighth exon of *OsMADS58* matches very well to the 68 bp long intronic region following the eighth exon of *OsMADS3* (Figure 7C), strongly suggests that *OsMADS58* may have been generated from an *OsMADS3*-like ancestral gene through the exonization of the first 68 bp of intron 8 and the pseudoexonization of exon 9.

The differences between the two *AGL11*-like genes in rice (i.e. *OsMADS13* and *OsMADS21*) are also conspicuous, although they are 56% and 84% identical at the protein and DNA levels, respectively. Both genes contain seven exons and six introns (Figure 8A); the first six exons are highly conserved, whereas the seventh exon contains many out-of-frame insertions/deletions (Figure 8B). As a result, the peptides that are homologous between OsMADS13 and OsMADS21 proteins are

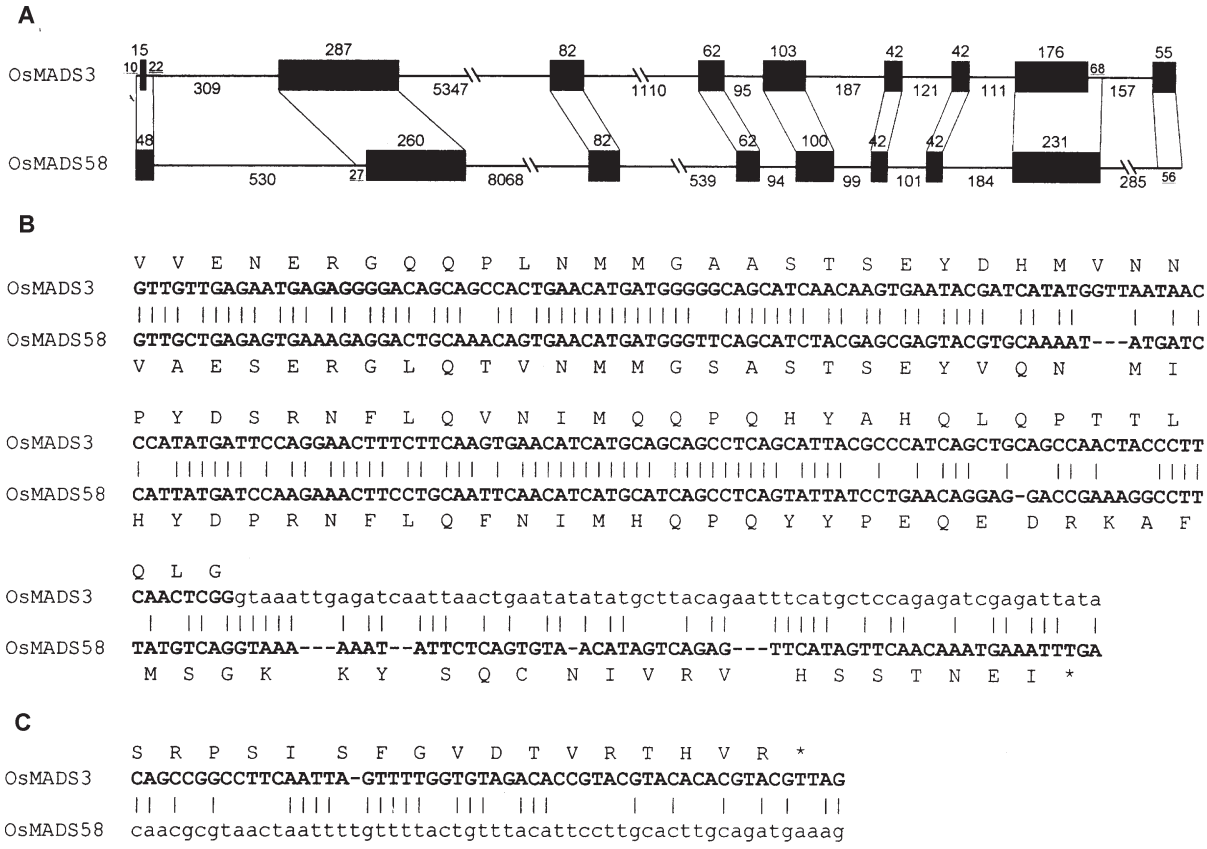


Figure 7. Comparison of the exon/intron structure of *OsMADS3* and *OsMADS58*.

(A) Schematic representations of exons (in black boxes) and introns. The lengths of each exon and intron are largely proportional to the real lengths. Regions (especially exons) that can match to each other are connected with thin lines.

(B) Alignment for the eighth exon of the two genes. Uppercase bold letters denote exon sequence, and lowercase, intron sequence. Vertical lines indicate nucleotides identical between the two genes. Amino acid sequences are given above and below the exons.

(C) Alignment for the ninth exon of *OsMADS3* and the intergenic region downstream the eighth exon of *OsMADS58*.

only occasionally found within this region. Nevertheless, the AG I motif (Kramer et al. 2004) seems to be present in both proteins, although they are quite different in sequence (i.e. "LDMKCFFPLNLFE" in OsMADS13 and "FDTREYYQPAPPV" in OsMADS21). However, due to out-of-frame mutations, the AG II motif (Kramer et al. 2004) of OsMADS13 is only partially homologous to that of OsMADS21 and other AGL11-like proteins. More interestingly, compared with other AGL11-like proteins, both OsMADS13 and OsMADS21 possess elongated C-terminal ends, suggesting that mutational changes may have also converted the otherwise conserved stop codon into a sense codon.

OsMADS5 and *OsMADS1*, the two *AGL2*-like genes that are 71% and 88% identical at the protein and DNA levels, respectively, both contain eight exons and seven introns (Figure 9A). The first six exons are highly conserved in both length and sequence, whereas the seventh exon of *OsMADS5* contains several in-frame deletions compared with that of *OsMADS1*, so that the SEP I motif (Zahn et al. 2005) of the OsMADS5 protein is no longer intact (Figure 9B). More strikingly, the eighth exon of *OsMADS5* is only 34 bp, much shorter than that of OsMADS1, which is 115 bp. Sequence comparison further suggests that the eighth exon of *OsMADS5* matches very well to the middle part of the eighth exon of *OsMADS1* (Figure 9C). However, an insertion of a cytosine (C) in *OsMADS5* seems to have led to the occurrence of a premature stop codon so that the remaining 21 amino acids at the C-terminal end of OsMADS1, which contains part of the SEP II motif (Zahn et al. 2005; same as the ZMM3 motif defined by Vandenbussche et al. 2003) and the whole OsMADS1 motif (Vandenbussche et al. 2003), are missing in OsMADS5.

Discussion

Parallel duplications of floral MADS-box genes

Previous studies have indicated that floral MADS-box genes duplicated frequently during the evolution of flowering plants. In particular, the origin of angiosperms, as well as that of core

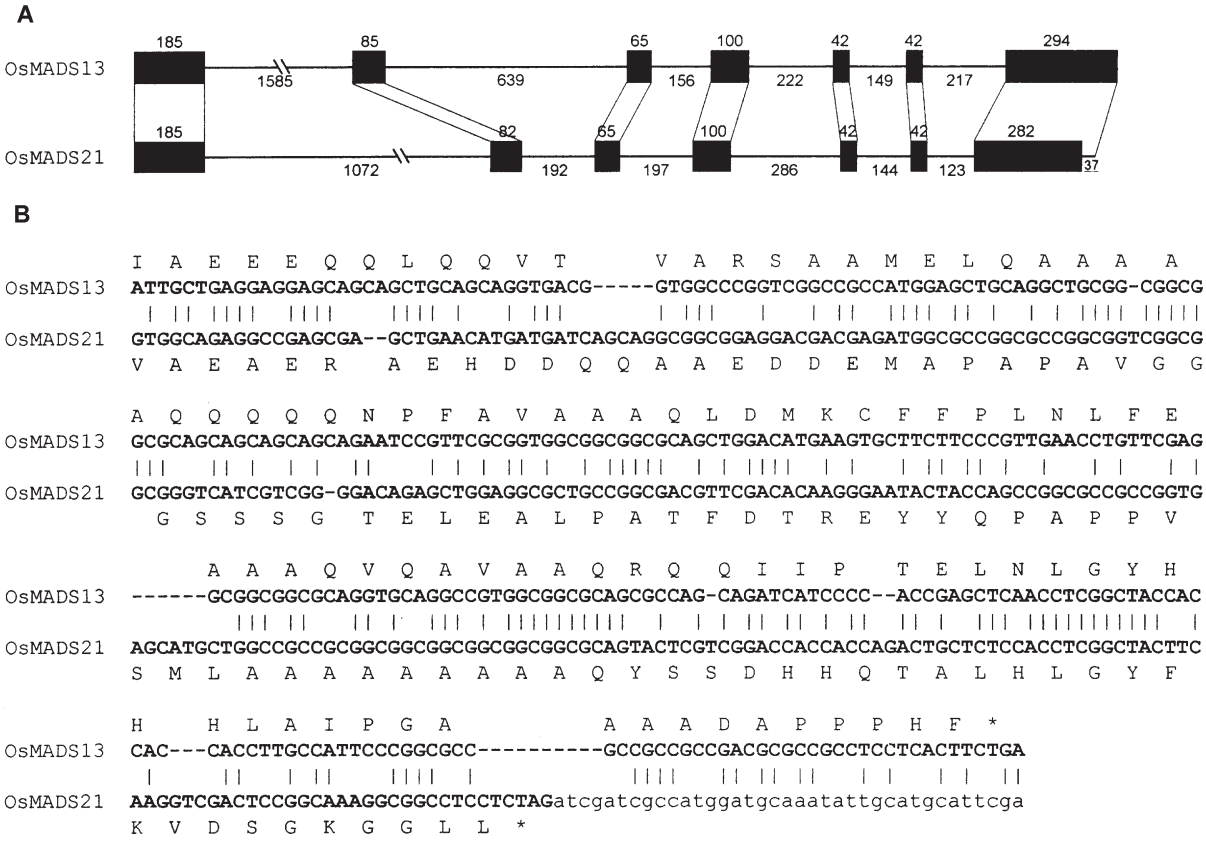


Figure 8. Comparison of the exon/intron structure of *OsMADS13* and *OsMADS21*.

(A) Schematic representations of exons and introns.

(B) Alignment for the seventh exon of the two genes.

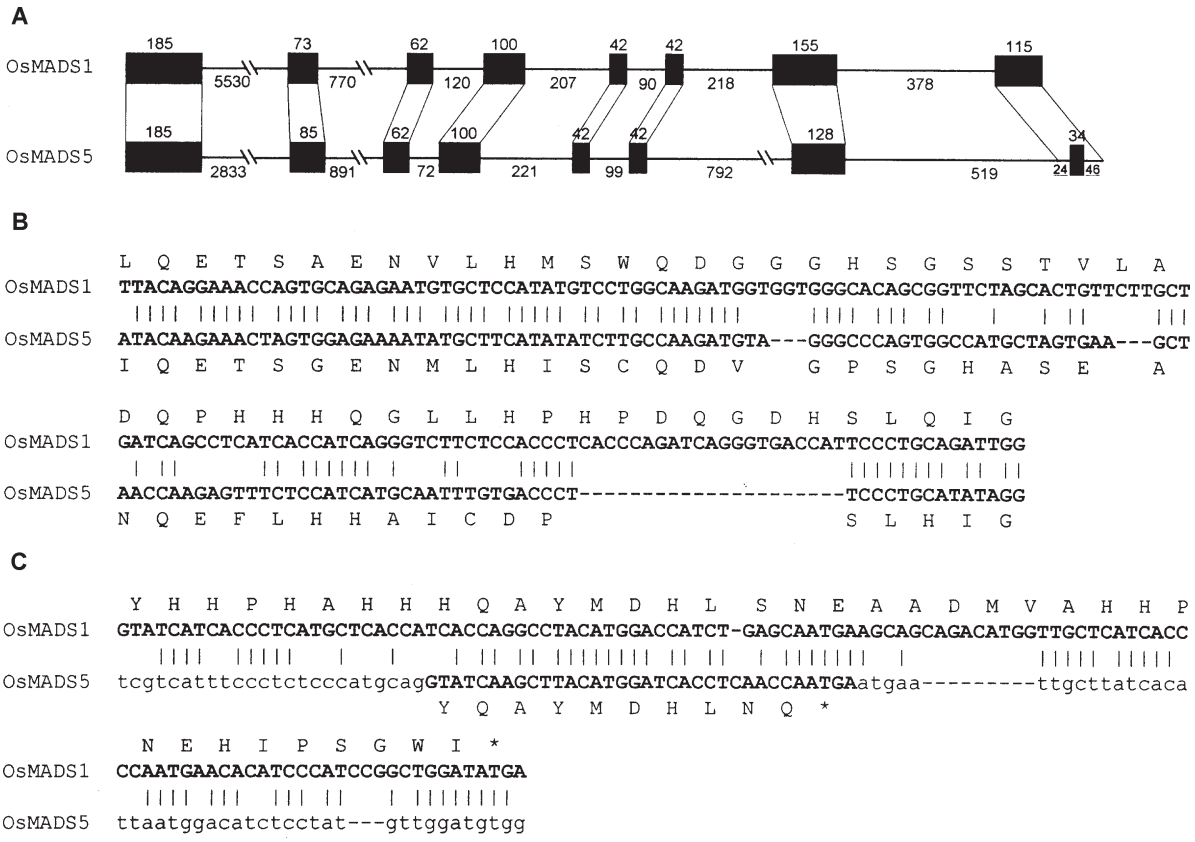


Figure 9. Comparison of the exon/intron structure of *OsMADS1* and *OsMADS5*.

(A) Schematic representations of exons and introns. Note that the eighth exon of *OsMADS5* matches very well to the middle part of the eighth exon of *OsMADS1.*

(B) Alignment for the seventh exon of the two genes.

(C) Alignment for the eighth exon of the two genes.

eudicots, seems to coincide with the duplication of floral genes. In the present study, by conducting extensive phylogenetic analyses, we identified pre-Poaceae gene duplication events in six of the seven lineages of floral MADS-box genes. This observation, together with the fact that members of the Poaceae usually have quite distinct floral structures from other monocots and other angiosperms, strongly suggests that the duplication of floral genes may have contributed to the formation of a more complex regulatory network for floral development and thus led to the origin/diversification of more advanced systems. In addition, the similar phenomena observed in the core eudicots and the Poaceae, the two highly derived plant groups in eudicots and monocots, respectively, suggest that the modification of the already well-organized regulatory network can happen independently in different organismal lineages.

It has also been suggested that the pre-core eudicot duplication in each of the *AP1*, *AP3*, *AG* and *AGL2/3/4* lineages arose through a genome duplication event (Irish 2003; Zahn et al. 2005). In the present study, we found evidence that four pairs of rice genes very likely resulted from segmental duplications that correspond to the doubling of the rice genome 66-70 MYA (Vandepoele et al. 2003; Paterson et al. 2004; Wang et al. 2005; Yu et al. 2005). If this is true, then it suggests that, after genome duplication, most duplicated floral MADS-box genes tend to be preferentially retained in the genome. This is easy to understand, because the genes generated this way usually possess the regulatory elements needed and, thus, have a high probability to survive unless extra copies can result in a dosage effect that is detrimental to the plant. In addition, because duplicated floral MADS-box genes tend to have overlapping (or redundant) as well as differentiated expression patterns and/or functions (Ferrandiz et al. 2000a, 2000b; Pelaz et al. 2000), the probability for both copies to be retained increases. In spite of this, several lines of evidence suggest that elevation

of the expression level of some floral MADS-box genes can result in obvious morphological, physiological, and/or biochemical changes. For example, overexpression of an *AP1* lineage member (such as *AP1*, *CAL*, or *FUL*) in *A. thaliana* can cause early flowering and other phenotypes (Blazquez and Weigel 2000; Ferrandiz et al. 2000a, 2000b). This suggests that the increase in gene number may be harmful to the plant, and that the mechanism to selectively preserve or abandon a duplicate gene copy may be rather complex.

Conservation and divergence of duplicated genes

Several studies have proposed that duplicated genes can be retained in the genome either by subfunctionalization (the process in which the function of the ancestral gene is partitioned between two duplicate genes) or by neofunctionalization (the process in which one or both duplicate genes acquire novel functions; Force et al. 1999; Kramer et al. 2004). In plants, many duplicate genes have been shown to perform related but distinguishable functions. In other words, they can accumulate differences in both coding and regulatory regions. Changes, especially out-of-frame insertions and/or deletions, in coding regions may result in the generation of a protein with different functions, whereas mutations in regulatory regions may sometimes lead to a shift in expression patterns (Moore and Purugganan 2005).

In the present study, we observed differences in the coding regions of duplicate genes (*OsMADS3* and *OsMADS58*, *OsMADS13* and *OsMADS21*, and *OsMADS1* and *OsMADS5*). Because these differences have caused changes in protein sequences, it is reasonable to assume that the ability of these duplicate genes to interact with their potential partners may have also changed. However, owing to a lack of functional analyses (especially protein-protein interaction assays), it is still hard to know the real situation. Nevertheless, there is evidence that the aforementioned duplicate genes all perform partially redundant and partially divergent functions. For example, although both *OsMADS3* and *OsMADS58* are initially expressed in the floral meristems, at the later stages the transcripts of *OsMADS3* were detected in the ovule primordia, whereas those of *OsMADS58* were detected in the stamens and carpels (Kang et al. 1995; Kyozuka et al. 2000; Kater et al. 2006; Yamaguchi et al. 2006). Functional studies further suggested that the former gene plays more crucial roles in the development of lodicules and stamens, whereas the latter contributes more to floral meristem determinacy and carpel identities (Kang et al. 1998; Kater et al. 2006; Yamaguchi et al. 2006). Similarly, within the *AGL2/3/4* lineage, *OsMADS1* is more likely to be an E-function gene because it can influence the development of palea, lemma, lodicules, stamens, and carpels (Prasad et al. 2005; Kater et al. 2006). However, *OsMADS5* seems to have little effect on flower development, because the *osmads5* mutant almost exhibits no obvious phenotypes (Agrawal et al. 2005; Kater et al. 2006).

Materials and Methods

Data retrieval

Floral MADS-box genes (i.e. members of the *AP1*, *AP3*, *PI*, *AG*, *AGL11*, *AGL2/3/4*, and *AGL9* lineages) used in the present study were retrieved by BLAST searches (Altschul et al. 1997) against the NCBI (http://www.ncbi.nlm.nih.gov), TAIR (http://www.arabidopsis.org), and TIGR (http://www.tigr.org) databases. Sequences from the same species were regarded as alleles if they were over 95% identical at the DNA level. Only one such allele was included, whereas other alleles, as well as sequences with poor quality, were excluded from further analyses, leaving a total of 322 genes from 100 species in our original data set. Because the rice genome has been completely sequenced, we included the full collection of the rice genes in the data set. The synonym(s) and locus number of the rice genes are as follows: *OsMADS14* (*RAP1B* or *FDRMADS6*; Os03g54160; Moon et al. 1999b; Jia et al. 2000; Kyozuka et al. 2000), *OsMADS15* (*RAP1A*; Os07g01820; Moon et al. 1999b; Kyozuka et al. 2000), *OsMADS18* (*FDRMADS7*; Os07g41370; Moon et al. 1999b; Jia et al. 2000), *OsMADS20* (Os12g31748; Lee et al. 2003), *OsMADS16* (*SPW1*; Os06g49840; Moon et al. 1999a; Nagasawa et al. 2003), *OsMADS2* (*NMADS1*; Os01g66030; Chung et al. 1995; Yuan et al. 2000), *OsMADS4* (Os05g34940; Chung et al. 1995), *OsMADS3* (*RAG*; Os01g10504; Kang and Hannapel 1995; Kyozuka et al. 2000), *OsMADS58* (Os05g11414; Yamaguchi et al. 2006), *OsMADS13* (Os12g10540; Lopez-Dee et al. 1999), *OsMADS21* (Os01g66290; Sasaki et al. 2002), *OsMADS1* (*LHS1*; Os03g11614; Chung et al. 1994; Jeon et al. 2000), *OsMADS5* (*FDRMADS2*; Os06g06750; Kang and An 1997; Jia et al. 2000), *OsMADS34* (*OsMADS19*; Os03g54170; Shinozuka et al. 1999; Malcomber and Kellogg 2004), *OsMADS24* (*OsMADS8*; Os09g32948; Greco et al. 1997; Kang and An 1997), and *OsMADS45* (*FDRMADS1* or *OsMADS7*; Os08g41950; Greco et al. 1997; Kang and An 1997; Jia et al. 2000).

Sequence alignment and phylogenetic analysis

Protein sequences in each of the *AP1*, *AP3*, *PI*, *AG*, *AGL11*, *AGL2/3/4*, and *AGL9* gene lineages were first aligned with CLUSTALX 1.83 (Thompson et al. 1997) and then adjusted manually in Gendoc (Nicholas and Nicholas 1997). Because some parts of the C-terminal regions were too divergent to be aligned confidently, a preliminary tree for each lineage was produced based on the analyses of the conserved M-, I-, K-domain regions. Then, the order of the sequences was adjusted according to

the phylogenetic relationships so that closely related sequences were listed together. At this time, the alignment for the less-conserved C-terminal regions became much easier and a new alignment was produced. A DNA version of each protein alignment was also generated with the help of the publicly available software aa2dna (http://www.bio.psu.edu/People/Faculty/Nei/Lab/software.htm).

Phylogenetic trees for each gene lineage were estimated on the basis of both protein and DNA matrices. To assure the reliability of the phylogenetic estimates, only relatively conserved regions (such as the M-, I-, and K-domain regions) were used because the alignment in the less-conserved regions (such as the C-terminal regions) is still problematic. In particular, due to the occurrence of frameshift mutations, the C-terminal ends of some proteins are no longer homologous to those of the others (see below). During phylogenetic analyses, these non-homologous regions should be excluded, as suggested in previous studies (Zahn et al. 2005; Shan et al. 2007).

Phylogenetic estimates for each matrix were performed using maximum parsimony (MP) and maximum likelihood (ML) methods in PAUP* 4.0b10 (Swofford 2002) and PHYML version 2.4 (Guindon and Gascuel 2003), respectively. For the MP analyses, heuristic searches were conducted with 1 000 random addition replicates, with tree bisection-reconnection (TBR) branch swapping and saving all most parsimonious trees at each replicate (MulTree on). Support for each branch was assessed using bootstrap analyses with 250 bootstrap replicates, each with 50 stepwise additions and TBR branch swapping. For the ML analyses of protein matrices in PHYML, the default JTT model was chosen, with the proportion of invariable sites and the gamma distribution parameter optimized automatically and a BIONJ tree used as a starting point. For the ML analyses of DNA matrices, the most appropriate model, GTR+I+Γ, and other parameters were first obtained by running MODELTEST version 3.06 (Posada and Crandall 1998) and then applied in PHYML. Bootstrap analyses (200 replicates for protein matrices and 1 000 replicates for DNA matrices) were also performed to test the reliability of the phylogenetic trees. Because PHYML can give reasonable results in a relatively short time (Zahn et al. 2005; Kong et al. 2007; Shan et al. 2007), we based our descriptions and conclusions mainly on the ML trees generated by PHYML.

Identification of segmental duplication

To determine whether some paralogous genes in rice were actually generated through segmental duplications, we compared the 50-kb regions both upstream and downstream of the paired genes using the DotPlot function of the PipMaker program (Schwartz et al. 2000) at http://pipmaker.bx.psu.edu/pipmaker/. However, because this method may give false negative results when the genes compared contain small exons and large introns, we also tried to compare protein sequences of the candidate genes, as well as the 10 genes both upstream and downstream of the candidate genes, using the DotLet program (Junier and Pagni 2000) at http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html. Genomic regions containing candidate gene pairs were regarded as arising from segmental duplications if at least one additional gene pair could be identified in the corresponding positions. Segmental duplications were further regarded as corresponding to a genome doubling event if the synonymous distances (K_s) between duplicate gene pairs are close to each other. The K_s values between duplicate genes were calculated in MEGA3.1 (http://www.megasoftware.net/), using the Jukes-Cantor model (Kumar et al. 2004).

Detection of sequence divergences between duplicate genes

Differences between duplicate genes were first observed from the alignments of protein sequences. Then, detailed comparisons of the exon/intron sequences between closely related paralogous genes were conducted to understand the mechanism by which a diverged C-terminal region was generated. During this process, the exon of one gene was aligned with its candidate counterpart (i.e. the exon that was located at the same position) and the two adjacent (both upstream and downstream) introns of the other gene, and *vice versa*, to identify the best matches.

Acknowledgements

The authors thank Drs Hong Ma (Department of Biology and the Huck Institute of Life Sciences, Pennsylvania State University, USA) and Hongyan Shan (Institute of Botany, the Chinese Academy of Sciences, Beijing, China), and Yang Liu, Jian Zhang, and Jin Hu (Institute of Botany, the Chinese Academy of Sciences, Beijing, China) for their critical reading of the manuscript and their valuable comments. The authors also thank Dr Yang Zhong (School of Life Sciences, Fudan University) for helpful suggestions.

References

Agrawal GK, Abe K, Yamazaki M, Miyao A, Hirochika H (2005). Conservation of the E-function for floral organ identity in rice revealed by the analysis of tissue culture-induced loss-of-function mutants of the *OsMADS1* gene. *Plant Mol. Biol.* **59**, 125–135.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*

25, 3389–3402.

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

Angenent GC, Franken J, Busscher M, van Dijken A, van Went JL, Dons HJ et al. (1995). A novel class of MADS box genes is involved in ovule development in petunia. *Plant Cell* **7**, 1569–1582.

Becker A, Theissen G (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.* **29**, 464–489.

Blazquez MA, Weigel D (2000). Integration of floral inductive signals in *Arabidopsis*. *Nature* **404**, 889–892.

Chung YY, Kim SR, Finkel D, Yanofsky MF, An G (1994). Early flowering and reduced apical dominance result from ectopic expression of a rice MADS box gene. *Plant Mol. Biol.* **26**, 657–665.

Chung YY, Kim SR, Kang HG, Noh YS, Park MC, Finkel D et al. (1995). Characterization of two rice MADS box genes homologous to *GLOBOSA*. *Plant Sci.* **109**, 45–56.

Coen ES, Meyerowitz EM (1991). The war of whorls: Genetic interactions controlling flower development. *Nature* **353**, 31–37.

Colombo L, Franken J, Koetje E, van Went J, Dons HJ, Angenent GC et al. (1995). The *Petunia* MADS box gene *FBP11* determines ovule identity. *Plant Cell* **7**, 1859–1868.

Ferrandiz C, Gu Q, Martienssen R, Yanofsky MF (2000a). Redundant regulation of meristem identity and plant architecture by *FRUITFULL*, *APETALA1* and *CAULIFLOWER*. *Development* **127**, 725–734.

Ferrandiz C, Liljegren SJ, Yanofsky MF (2000b). Negative regulation of the *SHATTERPROOF* genes by *FRUITFULL* during *Arabidopsis* fruit development. *Science* **289**, 436–438.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.

Greco R, Stagi L, Colombo L, Angenent GC, Sari-Gorla M, Pe ME (1997). MADS box genes expressed in developing inflorescences of rice and sorghum. *Mol. Gen. Genet.* **253**, 615–623.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol*. **52**, 596–704.

Irish VF (2003). The evolution of floral homeotic gene function. *BioEssays* **25**, 637–646.

Irish VF (2006). Duplication, diversification and comparative genetics of angiosperm MADS-box genes. *Adv. Bot. Res.* **44**, 129–161.

Irish VF, Litt A (2005). Flower development and evolution: Gene duplication, diversification and redeployment. *Curr. Opin. Genet. Dev.* **15**, 454–460.

Jeon JS, Jang S, Lee S, Nam J, Kim C, Lee SH et al. (2000). *leafy hull sterile1* is a homeotic mutation in a rice MADS box gene affecting rice flower development. *Plant Cell* **12**, 871–884.

Jia H, Chen R, Cong B, Cao K, Sun C, Luo D (2000). Characterization and transcriptional profiles of two rice MADS-box genes. *Plant Sci.* **155**, 115–122.

Kang HG, An G (1997). Isolation and characterization of a rice MADS box gene belonging to the *AGL2* gene family. *Mol. Cell* **7**, 45–51.

Kang HG, Noh YS, Chung YY, Costa MA, An K, An G (1995). Phenotypic alterations of petal and sepal by ectopic expression of a rice MADS-box gene in tobacco. *Plant Mol. Biol.* **29**, 1–10.

Kang HG, Jeon JS, Lee S, An G (1998). Identification of class B and class C floral organ identity genes from rice plants. *Plant Mol. Biol.* **38**, 1021–1029.

Kang SG, Hannapel DJ (1995). Nucleotide sequences of novel potato (*Solanum tuberosum* L.) MADS-box cDNAs and their expression in vegetative organs. *Gene* **166**, 329–330.

Kater MM, Dreni L, Colombo L (2006). Functional conservation of MADS-box factors controlling floral organ identity in rice and *Arabidopsis*. *J. Exp. Bot.* **57**, 3433–3444.

Kaufmann K, Melzer S, Theissen G (2005). MIKC-type MADS-domain proteins: Structural modularity, protein interactions and network evolution in land plants. *Gene* **347**, 183–198.

Kim S, Yoo MJ, Albert VA, Farris JS, Soltis PS, Soltis DE (2004). Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implication of a 260-million-year-old duplication. *Am. J. Bot.* **91**, 2102–2118.

Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, dePamphilis CW (2007). Patterns of gene duplication in the plant *SKP1* gene family in angiosperms: Evidence for multiple mechanisms of rapid gene birth. *Plant J.* (in press).

Kramer EM, Hall JC (2005). Evolutionary dynamics of genes controlling floral development. *Curr. Opin. Plant Biol.* **8**, 13–18.

Kramer EM, Irish VF (2000). Evolution of the petal and stamen developmental program: Evidence from comparative studies of the lower eudicots and basal angiosperms. *Int. J. Plant Sci.* **161** (Suppl.), S29–S40.

Kramer EM, Zimmer EA (2006). Gene duplication and floral developmental genetics of basal eudicots. *Adv. Bot. Res.* **44**, 353–384.

Kramer EM, Dorit RL, Irish VF (1998). Molecular evolution of genes controlling petal and stamen development: Duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. *Genetics* **149**, 765–783.

Kramer EM, Jaramillo MA, Di Stilio VS (2004). Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. *Genetics* **166**, 1011–1023.

Kumar S, Tamura K, Nei M (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**, 150–163.

Kyozuka J, Kobayashi T, Morita M, Shimamoto K (2000). Spatially and temporally regulated expression of rice MADS box genes with similarity to *Arabidopsis* class A, B and C genes. *Plant Cell Physiol.* **41**, 710–718.

Lamb RS, Irish VF (2003). Functional divergence within the *APETALA3*/*PISTILLATA* floral homeotic gene lineages. *Proc. Natl. Acad. Sci. USA* **100**, 6558–6563.

Lee S, Jeon JS, An K, Moon YH, Chung YY, An G (2003). Alteration of floral organ identity in rice through ectopic expression of *OsMADS16*. *Planta* **217**, 904–911.

Litt A, Irish VF (2003). Duplication and diversification in the *APETALA1*/*FRUITFULL* floral homeotic gene lineage: Implications for the evolution of floral development. *Genetics* **165**, 821–833.

Lopez-Dee ZP, Wittich P, Enrico Pe M, Rigola D, Del Buono I, Gorla MS et al. (1999). *OsMADS13*, a novel rice MADS-box gene expressed during ovule development. *Dev. Genet.* **25**, 237–244.

Ma H, dePamphilis C (2000). The ABCs of floral evolution. *Cell* **101**, 5–8.

Ma H, Yanofsky MF, Meyerowitz EM (1991). *AGL1–AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Dev. Genes Evol.* **5**, 484–495.

Malcomber ST, Kellogg EA (2004). Heterogeneous expression patterns and separate roles of the *SEPALLATA* gene *LEAFY HULL STERILE1* in grasses. *Plant Cell* **16**, 1692–1706.

Melzer R, Kaufmann K, Theissen G (2006). Missing links: DNA-binding and target gene specificity of floral homeotic proteins. *Adv. Bot. Res.* **44**, 209–236.

Moon YH, Jung JY, Kang HG, An G (1999a). Identification of a rice *APETALA3* homologue by yeast two-hybrid screening. *Plant Mol. Biol.* **40**, 167–177.

Moon YH, Kang HG, Jung JY, Jeon JS, Sung SK, An G (1999b). Determination of the motif responsible for interaction between the rice *APETALA1/AGAMOUS-LIKE9* family proteins using a yeast two-hybrid system. *Plant Physiol.* **120**, 1193–1204.

Moore RC, Purugganan MD (2005). The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* **8**, 122–128.

Nagasawa N, Miyoshi M, Sano Y, Satoh H, Hirano H, Sakai H et al. (2003). *SUPERWOMAN1* and *DROOPING LEAF* genes control floral organ identity in rice. *Development* **130**, 705–718.

Nam J, dePamphilis CW, Ma H, Nei M (2003). Antiquity and evolution of the MADS-box gene family controlling flower development in plants. *Mol. Biol. Evol.* **20**, 1435–1447.

Nicholas KB, Nicholas HB (1997). *Genedoc*: *A Tool for Editing and Annotating Multiple Sequence Alignments*. www. psc.ed/biomed/genedoc.

Paterson AH, Bowers JE, Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA* **101**, 9903–9908.

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.

Posada D, Crandall KA (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Prasad K, Parameswaran S, Vijayraghavan U (2005). *OsMADS1*, a rice MADS-box factor, controls differentiation of specific cell types in the lemma and palea and is an early-acting regulator of inner floral organs. *Plant J.* **43**, 915–928.

Purugganan MD (1997). The MADS-box floral homeotic gene lineages predate the origin of seed plants: Phylogenetic and molecular clock estimates. *J. Mol. Evol.* **45**, 392–396.

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y et al. (2002). The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J et al. (2000). PipMaker— A web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586.

Shan H, Zhang N, Liu C, Xu G, Zhang J, Chen Z et al. (2007). Patterns of gene duplication and functional diversification during the evolution of the *AP1*/*SQUA* subfamily of plant MADS-box genes. *Mol. Phylogenet. Evol.* (in press).

Shinozuka Y, Kojima S, Shomura A, Ichimura H, Yano M, Yamamoto K et al. (1999). Isolation and characterization of rice MADS-box gene homologues and their RFLP mapping. *DNA Res.* **6**, 123–129.

Soltis DE, Soltis PS, Albert VA, Oppenheimer DG, dePamphilis CW, Ma H et al. (2002). Missing links: The genetic architecture of flower and floral diversification. *Trends Plant Sci.* **7**, 22–31.

Stellari GM, Jaramillo MA, Kramer EM (2004). Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. *Mol. Biol. Evol.* **21**, 506–519.

Swofford DL (2002). *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods).* Version 4.0b10. 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.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T (2003). Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Res.* **31**, 4401–4409.

Vandepoele K, Simillion C, Van de Peer Y (2003). Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* **15**,

2192–2202.

Wang X, Shi X, Hao B, Ge S, Luo J (2005). Duplication and DNA segmental loss in the rice genome: Implications for diploidization. *New Phytol.* **165**, 937–946.

Whipple CJ, Schmidt RJ (2006). Genetics of grass flower development. *Adv. Bot. Res.* **44**, 385–424.

Yamaguchi T, Lee DY, Miyao A, Hirochika H, An G, Hirano HY (2006). Functional diversification of the two C-class MADS box genes *OSMADS3* and *OSMADS58* in *Oryza sativa*. *Plant Cell* **18**, 15–28.

Yanofsky M, 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, Wang J, Lin W, Li S, Li H, Zhou J et al. (2005). The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3**, e38.

Yuan ZQ, Yang JS, Liu J (2000). Cloning and characterization of two cDNAs encoding rice MADS-box protein. *Prog. Nat. Sci.* **5**, 357–363.

Zahn LM, Kong H, Leebens-Mack J, Kim S, Soltis PS, Landherr LL et al. (2005). The evolution of the *SEPALLATA* subfamily of MADS-box genes: A pre-angiosperm origin with multiple duplications throughout angiosperm history. *Genetics* **169**, 2209–2223.

Zahn LM, Feng B, Ma H (2006a). Beyond the ABC model: Regulation of floral homeotic genes. *Adv. Bot. Res.* **44**, 163–207.

Zahn LM, Leebens-Mack J, Arrington JM, Hu Y, Landherr LL, dePamphilis CW et al. (2006b). Conservation and divergence in the *AGAMOUS* subfamily of MADS-box genes: Evidence of independent sub- and neofunctionalization events. *Evol. Dev.* **8**, 30–45.

Zhao D, Yu Q, Chen C, Ma H (2001). Genetic control of reproductive meristems. In: McManus MT, Veit B, eds. *Meristematic Tissues in Plant Growth and Development*. Sheffield Academic Press, Sheffield. pp. 89–142.

(Handling editor: Hong Ma)