# MIKC type genes of the MADS-box family in wheat: molecular and phylogenetic analysis

**Paolacci AR[1], Tanzarella OA[1], Porceddu E[1], Ciaffi M[1]**

*[1] DABAC, Università della Tuscia, Via S. Camillo de Lellis, 01100 Viterbo, Italy*

## INTRODUCTION

The MADS-box family of transcription factors play key roles in the control of development and signal transduction in eukaryotes (1). The MADS family contains a DNA-binding domain (MADS box) and includes two main lineages, type I and type II, both of which are present in plants, animals and fungi (2). Type II MADS-box proteins of plants,also named as MIKC-type proteins, possess three more domains than Type I MADS-box proteins: intervening (I) domain, keratin-like (K) domain and C-terminal (C) domain. In plants MIKC-type genes are involved in several important developmental processes, such as flower morphogenesis, ovule development, vegetative growth, embryogenesis and fruit formation, and through phylogenetic analysis based on sequence comparison, they have been classified into 13 subfamilies (3). The MIKC transcription factors controlling floral organ identity are the best characterized. Analysis of homeotic floral mutants of *Arabidopsis* resulted in the formulation of the ABCDE genetic model, which explains how the combined functions and organ-specific expression of five gene classes (A, B, C, D and E) specify the identity of the floral organs (sepals, petals, stamens and carpels), forming the four concentric whorls, and ovules (4,5). The most comprehensive cloning and characterization of MIKC genes of grasses have been carried out in maize and rice (6,7). In this study we extend a detailed characterization of the diversity and complexity of MIKC-type genes to hexaploid wheat.

## CLONING AND CHARACTERIZATION OF MIKC-TYPE MADS-BOX GENES IN WHEAT

The 34 available sequences of MIKC-type genes of rice (7) and 37 of *Arabidopsis* (3) were exploited to BLAST search the public databases of wheat ESTs: TIGR wheat gene index database (TaGI, version 10), HarvEST wheat (version 1.13) and NCBI. The 29 identified non-redundant MIKC-type consensus sequences were used as templates for 5' and 3' RACE extensions. Full-length cDNAs of the 29 putative wheat MIKC-type genes from various plant tissues were obtained using primer pairs specific for the very 5' and 3' ends, The products of five independent RT-PCR reactions for each of the 29 primer pairs were cloned and sequenced. For 15 primer pairs the five full-length cDNAs exhibited the same sequence, whereas the five clones obtained by each of the remaining 14 primer pairs showed either two or three similar but not identical sequences. A code of Ta, the first two letters of *Triticum aestivum* was assigned to each sequence, followed by the name of its orthologous gene of *Arabidopsis* identifying the corresponding MIKC phylogenetic subfamily. Multiple sequences clustering into the same subfamily were designed by an additional number (1-6) and the different clones obtained by the same primer pair were distinguished by an additional letter (A-C) (Table 1). Search for conserved motifs using the Pfam HMMs, InterPro and SMART databases revealed that proteins encoded by the 45 full-length wheat cDNAs showed the modular structure typical of plant MIKC-type MADS-box proteins (Table 1). Four wheat cDNAs showed N-terminal extensions upstream their MADS-box domain: TaAG-1 (36 aa), TaSOC1-2 (33 aa), TaAG-2A (37 aa) and TaAG-2B (37 aa). On the basis of sequence homology three of them (TaAG-1, TaAG-2A and TaAG-2B) were assigned to the subfamily including the gene *Agamous* of *Arabidospsis*, whose members are characterized by N-terminal extensions. The presence of the N-terminal extension is less obvious for TaSOC1-2,

**Table 1. Characteristics of nucleotide and deduced amino acid sequences of 45 full-length cDNAs of wheat MIKC-type genes.**

| Clone | Sequence length (nt) | | | | Protein | | Domains (aa) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Tot. | 5'UTR | 3'UTR | ORF | (aa) | 5'ext. | M | I | K | C |
| TaSOC1-1A | 1000 | 84 | 226 | 690 | 230 | - | 60 | 12 | 101 | 57 |
| TaSOC1-1B | 999 | 84 | 228 | 687 | 229 | - | 60 | 12 | 100 | 57 |
| TaAG-1 | 1101 | 18 | 276 | 807 | 269 | 36 | 60 | 13 | 100 | 60 |
| TaAG-3A | 1071 | 148 | 167 | 756 | 252 | - | 60 | 14 | 101 | 77 |
| TaAG-3B | 1077 | 148 | 167 | 762 | 254 | - | 60 | 14 | 101 | 79 |
| TaSEP-1 | 1076 | 15 | 317 | 744 | 248 | - | 60 | 14 | 97 | 77 |
| TaSEP-2A | 976 | 75 | 187 | 714 | 238 | - | 60 | 14 | 98 | 66 |
| TaSEP-2B | 976 | 75 | 226 | 675 | 225 | - | 60 | 14 | 98 | 53 |
| TaAP1-1 | 1163 | 144 | 287 | 732 | 244 | - | 60 | 14 | 100 | 70 |
| TaAP1-2 | 1111 | 116 | 194 | 801 | 267 | - | 60 | 15 | 100 | 92 |
| TaAP1-3 | 1255 | 140 | 293 | 822 | 274 | - | 60 | 14 | 100 | 100 |
| TaAGL6-A | 1144 | 126 | 241 | 777 | 259 | - | 60 | 13 | 99 | 87 |
| TaAGL6-B | 1141 | 126 | 241 | 774 | 258 | - | 60 | 13 | 99 | 86 |
| TaAGL6-C | 1158 | 134 | 250 | 774 | 258 | - | 60 | 13 | 99 | 86 |
| TaSEP-3A | 1226 | 113 | 357 | 756 | 252 | - | 60 | 14 | 102 | 76 |
| TaSEP-3B | 1233 | 113 | 364 | 756 | 252 | - | 60 | 14 | 102 | 76 |
| TaSEP-4 | 1033 | 67 | 228 | 738 | 246 | - | 60 | 14 | 102 | 70 |
| TaAGL12 | 890 | 124 | 94 | 672 | 224 | - | 60 | 9 | 105 | 50 |
| TaAP3 | 970 | 104 | 179 | 687 | 229 | - | 60 | 10 | 100 | 59 |
| TaPI-1 | 1035 | 217 | 194 | 624 | 208 | - | 60 | 10 | 100 | 38 |
| TaPI-2 | 953 | 24 | 302 | 627 | 209 | - | 60 | 10 | 100 | 39 |
| TaWM16 | 986 | 118 | 280 | 588 | 196 | - | 60 | 11 | 100 | 25 |
| TaSOC1-2 | 967 | 32 | 158 | 777 | 259 | 33 | 60 | 17 | 100 | 49 |
| TaSEP-5A | 1202 | 214 | 280 | 708 | 236 | - | 60 | 14 | 100 | 62 |
| TaSEP-5B | 1186 | 197 | 281 | 708 | 236 | - | 60 | 14 | 100 | 62 |
| TaSEP-6 | 1028 | 91 | 256 | 681 | 227 | - | 60 | 14 | 98 | 55 |
| TaSOC1-3A | 937 | 144 | 127 | 666 | 222 | - | 60 | 13 | 99 | 50 |
| TaSOC1-3B | 931 | 138 | 127 | 666 | 222 | - | 60 | 10 | 102 | 50 |
| TaSVP-1A | 1282 | 183 | 415 | 684 | 228 | - | 60 | 9 | 103 | 56 |
| TaSVP-1B | 1297 | 198 | 415 | 684 | 228 | - | 60 | 9 | 103 | 56 |
| TaSVP-2A | 1041 | 86 | 277 | 678 | 226 | - | 60 | 18 | 95 | 53 |
| TaSVP-2B | 1043 | 89 | 276 | 678 | 226 | - | 60 | 18 | 95 | 53 |
| TaGGM13 | 1155 | 134 | 265 | 756 | 252 | - | 60 | 10 | 101 | 81 |
| TaAG-4A | 1073 | 49 | 259 | 765 | 255 | - | 60 | 12 | 102 | 81 |
| TaAG-4B | 1081 | 49 | 279 | 753 | 251 | - | 60 | 12 | 102 | 77 |
| TaSVP-3A | 1163 | 187 | 298 | 678 | 226 | - | 60 | 20 | 94 | 52 |
| TaSVP-3B | 1171 | 191 | 302 | 678 | 226 | - | 60 | 20 | 94 | 52 |
| TaAG-2A | 1141 | 107 | 215 | 819 | 273 | 37 | 60 | 13 | 101 | 62 |
| TaAG-2B | 1135 | 107 | 200 | 828 | 276 | 37 | 60 | 13 | 101 | 65 |
| TaAGL17-1 | 1119 | 123 | 276 | 720 | 240 | - | 60 | 12 | 100 | 68 |
| TaAGL17-2A | 1155 | 117 | 342 | 696 | 232 | - | 60 | 13 | 100 | 59 |
| TaAGL17-2B | 1164 | 126 | 342 | 696 | 232 | - | 60 | 13 | 100 | 59 |
| TaAGL17-2C | 1122 | 91 | 341 | 690 | 230 | - | 60 | 13 | 100 | 57 |
| TaAGL17-3A | 1012 | 40 | 249 | 723 | 241 | - | 60 | 12 | 99 | 70 |
| TaAGL17-3B | 1013 | 40 | 250 | 723 | 241 | - | 60 | 12 | 99 | 70 |

which belongs to the *SOC1/AGL20* subfamily, whose other members do not possess it; therefore its origin and functional meaning would require further studies. Multiple cDNAs cloned from independent amplifications with the same primer pair showed high identity (over 90%) at both nucleotide and amino acid levels, most probably because they derived from transcripts of MADS box genes located in homoeologous chromosomes. Most mutations detected in the coding region of the putative homoeologous cDNA sequences consisted of nucleotide substitutions, with prevalence of synonymous substitutions, and of short in-frame insertions/deletions (indels) of three, six and nine nucleotides, whereas the 5' and 3' untranslated regions showed higher rates of base substitutions and of indels of variable length (data not shown). The most remarkable differences observed among putative homoeologous sequences consisted of two frameshift mutations, in the C-terminal regions of the clone pairs TaSEP-2A/TaSEP-2B and TaAG-4A/TaAG-4B. Previous studies in other plant species have shown that frameshift mutations in the C-terminal domain of specific ancestral MADS-box MIKC-type genes may have contributed to their structural and functional divergence (8).

Southern analyses showed the presence in hexaploid wheat of three homoeologous copies for each of the 29 identified MIKC-type sequence (23 of them in Fig. 1), indicating that the genome of *T. aestivum* contains at least 87 (29x3) type II MIKC MADS-box genes. This assumption was further supported by aneuploid analysis of six sequences assigned by phylogenetic analysis to the *SEP/AGL2* subfamily. The three homoeologous sequences of both TaSEP-1 and TaSEP-2 were located in chromosome arms 4AS, 4BL and 4DL (Fig. 2a and 2d), those of TaSEP3 and TaSEP-5 in the long arms of group 5 chromosomes (Fig. 2b and 2e) and those of TaSEP-4 and TaSEP-6 in the short arms of group 7 chromosomes (Fig. 2c and 2f).

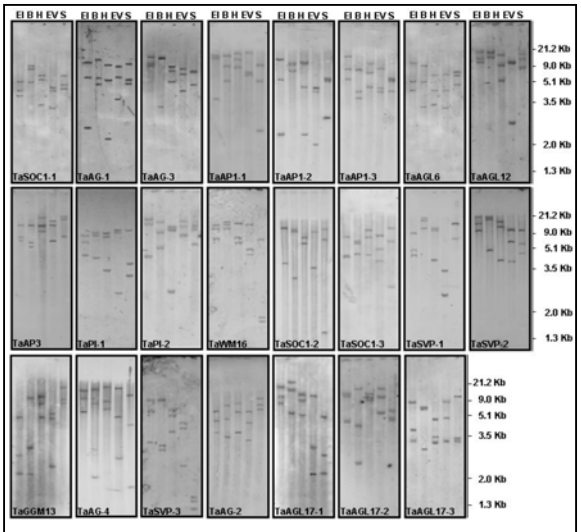


**Figure 1. Southern analysis of the hexaploid wheat cv Chinese Spring (CS) genomic DNA digested with *Eco*RI (EI), *Bam*HI (B), *Hin*dIII (H), *Eco*RV (EV) and *Sac*I (S) and hybridised with specific probes of 23 MIKC-type sequences of wheat cloned in this study.**

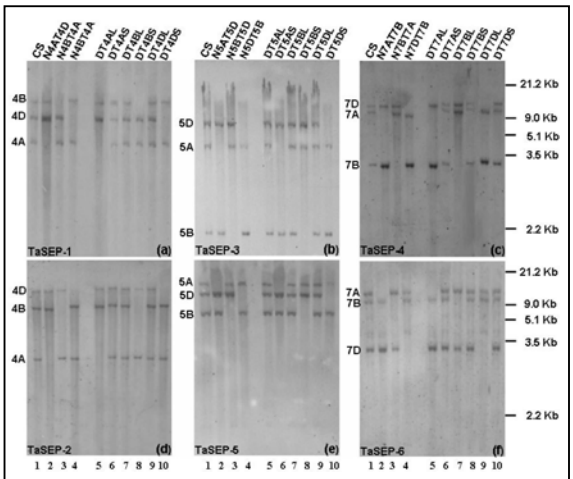


**Figure 2. Southern analysis of genomic DNA digested with *Bam*HI (a and c), *Hin*dIII (d and e), *Sac*I (b) and *Eco*RI (f) from CS and its nulli-tetrasomic and di-telosomic lines of homoeologous groups 4 (a and d), 5 (b and e) and 7 (c and f) and hybridised with specific probes for each of the six wheat *SEP*-like sequences.**

The chromosome arm locations of the six SEP-like genes of wheat were compatible with the map locations of their orthologous genes of rice and maize. In fact syntenic relationships, and sometimes high collinearities, have been shown between the chromosomal regions where the SEP-like genes of rice and maize are located and the 4, 5 and 7 chromosome homoeologous groups wherein the wheat genes were located (Table 2).

**Table 2. Chromosome arm locations and syntenic relationships between the SEP-like genes of wheat, rice and maize.**

| SEP-like wheat gene | Chromosome arm location | | |
|---|---|---|---|
| | Wheat | Rice orthologous* | Maize orthologous* |
| TaSEP-1 | 4AS, 4BL, 4DL | 3S (OsMADS1) | 1S (ZMM14), 9L (ZMM8) |
| TaSEP-2 | 4AS, 4BL, 4DL | 3S (OsMADS1) | 1S (ZMM14), 9L (ZMM8) |
| TaSEP-3 | 5AL, 5BL, 5DL | 9L (OsMADS8) | 7L (ZMM7) |
| TaSEP-4 | 7AS, 7BS, 7DS | 8L (OsMADS7) | 1L (ZMM6) |
| TaSEP-5 | 5AL, 5BL, 5DL | 3L (OsMADS34) | 1L (ZMM27) |
| TaSEP-6 | 7AS, 7BS, 7DS | 6S (OsMADS5) | 9S (ZMM3) |

*The chromosomal locations of *SEP*-like genes from rice and maize have been reported by Lee et al. (7) and Munster et al. (6).

## PHYLOGENETIC ANALYSIS

The evolutionary relationships between cloned wheat cDNAs and other known MIKC-type genes of plants were studied by the phylogeny reconstruction of 125 amino acid sequences deduced from 33 genes from *Arabidopsis*, 31 from rice, 32 from maize and 29 from wheat. The phylogenetic analysis revealed that 28 of the 29 non-homoeologous cDNA sequences of wheat were assigned to 11 of the 13 known plant subclasses of the MIKC-type gene family, whereas TA-WM16 and its putative ortholog *OsMADS32* of rice formed a separate clade. More extensive analyses would be needed to verify whether the *Arabidopsis* orthologs of these genes have been lost or are present only in monocot or grass genomes. The absence in wheat of sequences

homologous to *FLC*- and *AGL15*-like genes is consistent with previous studies, which found these genes only in *Brassicaceae* species, but not in other plants, even not in the sequenced genome of rice (3,7). These data indicate that at least 11 subfamilies of MIKC genes are common to both monocot and dicot species and that at least many different MIKC genes were carried by the common ancestor of monocots and dicots. Moreover, in agreement with previous phylogenetic studies, our results demonstrate that several gene duplications occurred in the lineage that originated grass species, preceding the separation of the lineages that led to extant maize, rice and wheat. Our phylogenetic analysis identified 27 minimal clades containing putative orthologs from both wheat and rice, 22 of them included also maize genes. The lack of maize orthologs in five minimal clades (belonging to the subfamilies *AGL12*, *AGL17*, *SOC1/AGL20* and *OsMADS32*) could be due either to incomplete MIKC gene sampling in that species or to their deletion after maize speciation. On the basis of these lines of evidence at least 22 (probably 27) MIKC genes were carried by the common ancestor of maize, rice and wheat, which diverged about 50-70 MYA. This represents a significant increase in comparison to the 11 MIKC-type genes that would have been present about 200 MYA in the common ancestor of monocots and eudicots. Most likely the increase of MIKC genes in the progenitor of grass species was caused by multiple duplications followed by gene diversification and could have been a key factor in the evolution of the complex inflorescences common to the species belonging to the Poaceae family.

## THE ABCDE MODEL OF FLOWER ORGAN IDENTITY IN WHEAT

The expression of the wheat cDNAs corresponding to the genes of the homeotic classes, A, B, Bs, C, D, and E was analysed by RT-PCR, quantitative real time RT-PCR and northern blot hybridisation in 18 different plant tissues and floral organs of Chinese Spring (data not shown). The expression patterns of these genes were compared with those of functionally characterised MADS-box genes of *Arabidopsis* and monocot species, such as rice and maize, in the same phylogenetic clades. Potential functions of the wheat genes corresponding to the cloned MADS-box sequences were assumed on the basis of sequence similarities and comparable expression patterns. The morphology of the grass flowers is quite different from that of dicots. However, molecular genetic studies of maize and rice suggest that lemma and palea are flower organs homologous to sepals of dicots and lodicules are modified petals, therefore the ABC model can effectively be extended to grass species (9). On the basis of these considerations and of the phylogenetic and expression analyses of the cloned cDNAs, we propose an adaptation of the "ABCDE model" of *Arabidopsis* to wheat. The wheat MIKC genes encoding the products, which most likely provide the A, B, C, D and E functions, are shown in Fig. 3. The additional function B sister (Bs) provided by TaGGM13 was introduced andmight be required together with the C and D functions for specifying the identities of carpels and ovules. The wheat class E (*SEP*-like) genes were split into the functional groups E1 and E2. The E1 activity would be provided by TaSEP-3 and TaSEP-4, whose redundant functions would be required for the development of lodicules, stamens and carpel. The E2 activity would be accomplished by TaSEP-1 and TaSEP-2, which are orthologs of the rice gene *LHS1*, and by TaSEP-5 and TaSEP-6.Theaction of these genes would be required specifically, together with that of the class A genes, for determining the lemma/palea identity. The expression patterns of TaSEP-1, TaSEP-2 and TaSEP-5 indicate that the corresponding genes would be involved also in the development of the reproductive organs of wheat flowers. Although the existence of functional A genes in grass species remains controversial, based on our expression analyses TaAP1-3 would likely be the putative wheat functional equivalent of *AP1* in *Arabidopsis*. The model presented here is mainly intended as a working hypothesis. Functional characterisation of the isolated wheat MIKC-type genes using loss- and gain-of-function mutants, analyses of transcript expression by *in situ* hybridisation and identification of protein-protein interaction patterns by yeast two-hybrid experiments will be required to confirm their role in flower organogenesis.

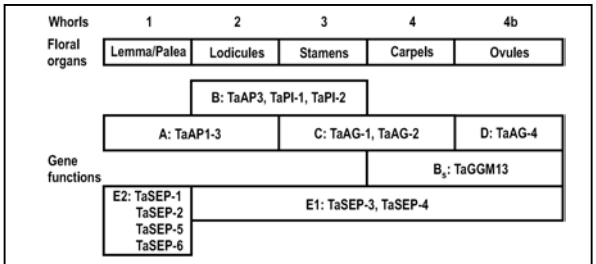


**Figure 3. The ABCDE model of flower organ identity in wheat.**

## REFERENCES


1. Messenguy F, Dubois E (2003). Gene 316:1-21.
2. Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, de Pouplana LR, Martinez-Castilla L, Yanofsky MF (2000). Proc Natl Acad Sci USA 97:5328-5333.
3. Becker A, Theissen G (2003). Molec Physiol Evol 29:464-489.
4. Theissen G (2001). Curr Opin Plant Biol 4:75-85.
5. Honma T, Goto K. (2001). Nature 409:525-529.
6. Munster T, Deleu W, Wingen LU, Ouzunova M, Cacharron J, Faigl W, Werth S, Kim JTT, Saedler H, Theissen G (2002). Maydica 47:287-301.
7. Lee S, Kim J, Son JS, Nam J, Jeong DH, Lee K, Jang S, Yoo J, Lee J, Lee D-Y, Kang HG, An G (2003). Plant Cell Physiol 44:1403-1411.
8. Vandenbussche M, Theissen G, Van de Peer Y, Gerats T (2003). Nucleic Acids Res 31:4401-4409.
9. Goto K, Kyozuka J, Bowman JL (2001). Curr Opin Genet Dev 11:449-456.