Mol Biol Rep (2013) 40:3901–3911
DOI 10.1007/s11033-012-2438-6

Genome-wide survey and expression analysis of the MADS-box gene family in soybean

Yongjun Shu · Diansi Yu · Dan Wang · Donglin Guo · Changhong Guo

Received: 29 August 2012 / Accepted: 18 December 2012 / Published online: 5 April 2013

Abstract MADS-box genes encode important transcription factors in plants that are involved in many processes during plant growth and development. An investigation of the soybean genome revealed 106 putative MADS-box genes. These genes were classified into two classes, type I and type II, based on phylogenetic analysis. The soybean type II group has 72 members, which is higher than that of *Arabidopsis*, indicating that soybean type II genes have undergone a higher rate of duplication and/or a lower rate of gene loss after duplication. Soybean MADS-box genes are present on all chromosomes. Like *Arabidopsis* and rice MADS-box genes, soybean MADS-box genes expanded through tandem gene duplication and segmental duplication events. There are many duplicate genes distributed across the soybean genome, with two genomic regions, i.e., MADS-box gene hotspots, where MADS-box genes with high degrees of similarity are clustered. Analysis of high-throughput sequencing data from soybean at different developmental stages and in different tissues revealed that MADS-box genes are expressed in embryos of various stages and in floral buds. This expression pattern suggests that soybean MADS-box genes play an important role in soybean growth and floral development.

Keywords MADS-box · Soybean · Phylogenetic analysis · Duplication · Expression patterns

Yongjun Shu and Diansi Yu contributed equally to this study.

Y. Shu · D. Yu · D. Wang · D. Guo · C. Guo (✉)
Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin 150025, Heilongjiang, People's Republic of China
e-mail: kaku_2008@163.com

Introduction

The MADS-box gene family is one of the most extensively studied transcription-factor gene families in eukaryotes (including animals, fungi, plants and other organisms) [1–3]. MADS-box proteins are characterized by the presence of a conserved domain of approximately 58–60 amino acids located in the N-terminal region known as the MADS-box domain. The MADS-box is a DNA binding domain that binds to CArG boxes (CC[A/T]6GG) [4, 5]. Based on phylogenetic analysis, the MADS-box gene family is divided into two categories, type I and type II, which originated by ancestral gene duplication. Type I MADS-box genes encode SRF-like domain proteins, while type II MADS-box genes encode MEF2-like proteins and plant-specific MIKC-type MADS-box proteins [6–8]. MIKC-type proteins generally contain four common domains. In addition to the MADS (M) domain, MIKC-type proteins contain the intervening (I), kertain-like (K) and C-terminal (C) domains [4, 6, 9, 10]. The I domain is the least conserved, which contributes to the DNA-binding specificity and dimerization of these proteins [3, 11]. The K domain is characterized by a coiled-coil structure, which mainly contributes to the dimerization of MADS-box proteins. The K domain, which is present in type II proteins but is absent in type I proteins, is more highly conserved [12–15]. The C domain is the least conserved domain, with high diversity, but it has been shown to play an important role in transcriptional activation and the formation of multimeric MADS-box protein complexes [16, 17]. Based on structural features, MADS-box genes can be divided into five groups. Type I genes, which have common ancestors, are subdivided into three groups, Mα, Mβ, and Mγ, based on the M domain of the encoded protein. Type II genes are subdivided into the MIKC^c and

MIKC* groups based on the diversity of the encoded I and K domains. The MIKC* group has longer I domains and less conserved K domains than the MIKC^c group [8, 18].

The MADS-box MIKC^c genes are the most well characterized group of MADS-box genes in plants. These genes play fundamental roles in many important processes during plant growth and development, including the following: determination of flowering time (SOC1 [SUPPRESSOR OF OVERESPRESSION OF CONSTANS1], FLC1 [FLOWERING LOCUS c], AGL24 [AGAMOUS-LIKE GENE 24], MAF1/FLM [MADS AFFECTING FLOWERING] and SVP (SHORT VEGETATIVE PHASE) [19–24]; floral meristem identity (AP1 [APETALA 1], FUL [FRUITFUL] and CAL [CAULIFLOWER]) [25–27]; the formation of floral organs (AP1, SEP1-3 [SEPALLATA 1-3], AP3 [APETALA 3], PI [PISTILLATA] and AG [AGAMOUS]) [5, 28]; fruit ripening (SHP1, SHP2 [SHATTERPROOF 1-2] and FUL) [27, 29] and seed pigmentation and embryo development (TT16 [TRANSPARENT TESTA16]) [30]. The most significant role of MADS-box genes is the determination of floral organ identity [31], which occurs via the ABC model. This model describes how the combined functions of three classes of MADS-box genes, along with AP2, specify floral organ identity.

Soybean is one of the most important sources of oil and vegetable protein for humans and animals. The soybean genome has been sequenced [32], which enables whole genome analysis of MADS-box genes in soybean. Therefore, in the present study, we determined the number of MADS-box genes in the soybean genome and characterized the gene structures, phylogenetic relationships, chromosomal locations, conserved motifs and expression patterns of MADS-box genes in soybean.

Materials and methods

Identification and classification of MADS-box genes

Soybean genomic and protein sequences were obtained from the Phytozome website (http://www.phytozome.net/soybean) [32]. The Hidden Markov Model (HMM) profile of the MADS-box domain (PF00319) was obtained from the Pfam website (pfam.sanger.ac.uk) [33], which was employed as a query to identify all possible MADS-box genes in soybean using HMMER (V3.0) software [34]. All of the candidate MADS-box proteins were aligned to *Arabidopsis* MADS-box proteins for classification into different groups. All of the annotation information concerning candidate MADS-box genes was obtained from the soybean genome website, and the numbers and distributions of introns in MADS-box genes were investigated using soybean genome annotation information.

Phylogenetic analysis of MADS-box genes

All candidate MADS-box protein sequences were aligned using ClustalW with default parameters [35]. The phylogenetic trees of all MADS-box proteins were generated using MEGA (V4.0) [36] with the neighbor-joining (NJ) method, with the following parameters: poisson correction, pairwise deletion and bootstrap (1,000 replicates).

Analysis of conserved motifs in MADS-box protein

The MADS-box protein sequences were analyzed using MEME software (Multiple EM for Motif Elicitation, V4.8.1) [37]. A MEME search was performed with the following parameters: (1) optimum motif width was set to $\geq$6 and $\leq$200; (2) the maximum number of motifs was set to identify 10 motifs and (3) occurrences of a single motif were distributed among the sequences with model: zero or one per sequence (-mod zoops). The MEME motifs were annotated using the Pfam database.

Chromosomal locations and gene duplication of MADS-box genes

Sequences of putative MADS-box genes (genomic sequences, CDS sequences) were obtained from the soybean genome database. The soybean MADS-box genes were blasted against each other to identify duplicate genes in which the similarity of the aligned regions was more than 85 % [38]. In addition, positional information about all of the MADS-box genes was investigated; the locations of the MADS-box genes on the soybean chromosomes were plotted using Matlab (R2008b) with custom script, and red lines were drawn linking duplicate genes on different chromosomes.

Expression analysis of soybean MADS-box genes *in silico*

The genome-wide transcriptome data of soybean seeds during several stages of seed development and throughout the soybean life cycle (obtained with high-throughput sequencing) were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov; accession numbers SRX062325–SRX062334). The transcript data were obtained from plant material including seeds at five stages of development (globular, heart, cotyledon, early-maturation, dry seeds), vegetative tissue (leaves, roots, stems, whole seedlings) and reproductive tissue (floral buds). All transcript data were analyzed with Matlab (R2008b) using Bioinformatics Toolbox.

Results

Identification and classification of MADS-box genes in soybean

To identify the full complement of MADS-box genes in soybean, the MADS domain (PF00319) was employed to search across the soybean genome. All 106 putative MADS-box genes were identified in soybean, which is similar to the number of MADS-box genes present in *Arabidopsis* (Table 1). More MADS members were identified with the MADS-domain screening method than were previously identified in the Plant Transcription Factor Database (PlantTFDB; 94 members), which employed protein homology searches [39]. By performing a homology search against *Arabidopsis* MADS-boxes, we were able to divide the soybean genes into two groups. A total of 72 genes were determined to be type II MADS-box genes (including MIKC^c and MIKC*), while 34 were confirmed to be type I MADS-box genes (including the Mα, Mβ and Mγ groups). However, while the total numbers of both type I and type II MADS-box genes in soybean (106) and *Arabidopsis* (107) are nearly identical, the number of type II genes is greater in soybean than in *Arabidopsis* or other plants, such as rice, maize, sorghum and cucumber. We also determined that soybean type II MADS-box genes usually contain multiple introns, with a maximum of 10 introns, while type I genes usually lack introns or have only a single intron. However, GmMADS73, 85, 91, 97, 103, 104, 105 and 106 have no more than three introns. These characteristics are consistent with those of MADS-box genes in *Arabidopsis*, rice, maize and other plants [14, 15, 40–42].

Phylogenetic analysis of MADS-box genes in soybean

To examine the phylogenetic relationships between MADS-box genes in detail, independent phylogenetic trees were constructed with type I and type II genes (Fig. 1). The type I soybean MADS-box genes were divided into 13 subfamilies. Subfamily CAL/FUL/AP1 has the most members, with 13 homologous genes. Many members are also present in subfamilies AG/SHP (12 members), AP3/PI (11 members), SEP (nine members) and SVP (seven members), while other subfamilies have fewer members, including SOC1 (four members), ANR1 (three members), FLC (two members), AGL6 (two members), AGL12 (two members), AGL15 (one member) and TT16 (one member). In addition, five genes, which are in the same group, were identified as MIKC*-type MADS-box genes. For type I soybean genes, the total numbers of genes in the Mα, Mβ and Mγ groups were similar to those of *Arabidopsis* and rice. However, the Mα and Mγ groups have more members, with 18 and 11 homologs, respectively (Fig. 2), while the Mβ group has just five members, which is significantly different from that of *Arabidopsis* and rice [14, 40].

Analysis of conserved motifs in MADS domain proteins

A total of 106 MADS-box genes from soybean were subjected to analysis to reveal conserved motifs shared among related proteins; ten conserved motifs, named motif 1–10, were identified (Fig. 3). Among these, the conserved motif encoding the MADS-box domain (motif 1) was found in all soybean MADS-box genes; this is the most conserved motif that was identified. Motif 5, which specifies the K domain, was found only in most MIKC^c group proteins (45 members). Motifs 2, 3, 4, 6, 7 and 9, representing the I region, are not highly conserved, except among closely related proteins, but these motifs were found in most MADS-box genes, from type I to type II genes. Finally, motifs 8 and 10, representing the C-terminal domain, are also weakly conserved in soybean MADS-box genes. These motifs are only present in some MIKC^c and Mα group proteins.

Chromosomal locations of MADS-box genes and their genomic duplication

The physical locations of the MADS-box genes on soybean chromosomes are shown in Fig. 4. Chromosomes 8 and 10 contain the most MADS-box genes (12 genes), while chromosome 17 contains the fewest (one gene). Other chromosomes contain between two and 11 MADS-box genes, suggesting that soybean MADS-box genes are distributed across these chromosomes. However, MADS-box genes are not randomly dispersed on each chromosome. Some chromosomes contain gene clusters or gene hotspots. For example, a short region of chromosome 10 contains ten MADS-box genes, and other chromosomes (chromosomes 5, 8, 11, 13 and 18) contain similar gene clusters. In addition, by performing gene duplication analysis, we found that many MADS-box genes (38.7 %, 41/106) are present in two or more copies. These gene duplications arose from tandem duplications and segment duplications. The soybean genome has undergone two duplication events. Therefore, many soybean genes are present in duplicate. Most MADS-box genes have undergone tandem duplication (45 duplications), while others have undergone segment duplication (17 duplications). Tandem duplications have produced MADS-box gene clusters or hotspots, while segment duplications have produced many homologs of MADS-box genes on different chromosomes, as indicated with red lines in Fig. 4.

Table 1 The MADS-box genes identified in soybean

Name	Gene locus	Group	Introns	Name	Gene locus	Group	Introns
GmMADS1	Glyma01g02880	MIKC	7	GmMADS54	Glyma02g45730	MIKC	4
GmMADS2	Glyma02g04710	MIKC	7	GmMADS55	Glyma14g03100	MIKC	7
GmMADS3	Glyma07g30040	MIKC	5	GmMADS56	Glyma08g42300	MIKC	7
GmMADS4	Glyma08g07260	MIKC	7	GmMADS57	Glyma18g12590	MIKC	7
GmMADS5	Glyma06g10020	MIKC	7	GmMADS58	Glyma13g29510	MIKC	7
GmMADS6	Glyma13g33030	MIKC	0	GmMADS59	Glyma15g09500	MIKC	6
GmMADS7	Glyma15g06300	MIKC	3	GmMADS60	Glyma08g12730	MIKC	7
GmMADS8	Glyma02g38090	MIKC	1	GmMADS61	Glyma04g43640	MIKC	7
GmMADS9	Glyma13g32810	MIKC	7	GmMADS62	Glyma06g48270	MIKC	7
GmMADS10	Glyma01g02530	MIKC	3	GmMADS63	Glyma05g29590	MIKC	4
GmMADS11	Glyma02g33040	MIKC	8	GmMADS64	Glyma09g40230	MIKC	6
GmMADS12	Glyma08g06980	MIKC	1	GmMADS65	Glyma18g45780	MIKC	6
GmMADS13	Glyma20g00400	MIKC	6	GmMADS66	Glyma05g03660	MIKC	5
GmMADS14	Glyma05g28130	MIKC	4	GmMADS67	Glyma20g29300	MIKC	6
GmMADS15	Glyma08g11110	MIKC	5	GmMADS68	Glyma13g02170	MIKC*	9
GmMADS16	Glyma12g17720	MIKC	1	GmMADS69	Glyma14g34160	MIKC*	10
GmMADS17	Glyma01g37470	MIKC	6	GmMADS70	Glyma07g35610	MIKC*	10
GmMADS18	Glyma11g07820	MIKC	6	GmMADS71	Glyma20g04500	MIKC*	10
GmMADS19	Glyma04g02980	MIKC	6	GmMADS72	Glyma05g00960	MIKC*	4
GmMADS20	Glyma06g02990	MIKC	6	GmMADS73	Glyma05g25930	Mα	3
GmMADS21	Glyma12g13560	MIKC	3	GmMADS74	Glyma05g35820	Mα	0
GmMADS22	Glyma18g33910	MIKC	3	GmMADS75	Glyma08g03790	Mα	0
GmMADS23	Glyma16g17450	MIKC	3	GmMADS76	Glyma18g20830	Mα	1
GmMADS24	Glyma04g42420	MIKC	5	GmMADS77	Glyma08g03820	Mα	1
GmMADS25	Glyma06g12380	MIKC	6	GmMADS78	Glyma10g10610	Mα	0
GmMADS26	Glyma13g09660	MIKC	6	GmMADS79	Glyma10g10930	Mα	0
GmMADS27	Glyma14g24590	MIKC	6	GmMADS80	Glyma10g10640	Mα	0
GmMADS28	Glyma05g28140	MIKC	7	GmMADS81	Glyma10g10770	Mα	0
GmMADS29	Glyma08g11120	MIKC	7	GmMADS82	Glyma10g10840	Mα	0
GmMADS30	Glyma11g36890	MIKC	6	GmMADS83	Glyma10g10860	Mα	0
GmMADS31	Glyma08g27670	MIKC	7	GmMADS84	Glyma10g10900	Mα	0
GmMADS32	Glyma18g50900	MIKC	7	GmMADS85	Glyma10g10690	Mα	2
GmMADS33	Glyma13g06730	MIKC	7	GmMADS86	Glyma10g11450	Mα	0
GmMADS34	Glyma19g04320	MIKC	7	GmMADS87	Glyma10g10920	Mα	0
GmMADS35	Glyma01g08130	MIKC	7	GmMADS88	Glyma20g27320	Mα	0
GmMADS36	Glyma18g00800	MIKC	2	GmMADS89	Glyma20g27330	Mα	0
GmMADS37	Glyma03g02210	MIKC	7	GmMADS90	Glyma10g40070	Mα	1
GmMADS38	Glyma07g08890	MIKC	7	GmMADS91	Glyma07g03400	Mβ	3
GmMADS39	Glyma01g08150	MIKC	7	GmMADS92	Glyma19g07170	Mβ	1
GmMADS40	Glyma02g13420	MIKC	7	GmMADS93	Glyma13g07720	Mβ	0
GmMADS41	Glyma08g36380	MIKC	7	GmMADS94	Glyma15g23350	Mβ	1
GmMADS42	Glyma16g13070	MIKC	7	GmMADS95	Glyma08g10080	Mβ	0
GmMADS43	Glyma05g07380	MIKC	7	GmMADS96	Glyma03g19880	Mγ	1
GmMADS44	Glyma17g08890	MIKC	7	GmMADS97	Glyma18g06010	Mγ	3
GmMADS45	Glyma08g27680	MIKC	6	GmMADS98	Glyma11g26260	Mγ	0
GmMADS46	Glyma18g50910	MIKC	7	GmMADS99	Glyma18g06040	Mγ	1
GmMADS47	Glyma04g31810	MIKC	2	GmMADS100	Glyma11g33460	Mγ	0
GmMADS48	Glyma09g27450	MIKC	2	GmMADS101	Glyma18g04760	Mγ	0
GmMADS49	Glyma10g38580	MIKC	7	GmMADS102	Glyma11g30640	Mγ	0
GmMADS50	Glyma20g29250	MIKC	7	GmMADS103	Glyma18g05980	Mγ	2
GmMADS51	Glyma16g32540	MIKC	7	GmMADS104	Glyma11g30500	Mγ	3
GmMADS52	Glyma09g36590	MIKC	6	GmMADS105	Glyma11g30630	Mγ	2
GmMADS53	Glyma12g00770	MIKC	6	GmMADS106	Glyma09g11550	Mγ	3

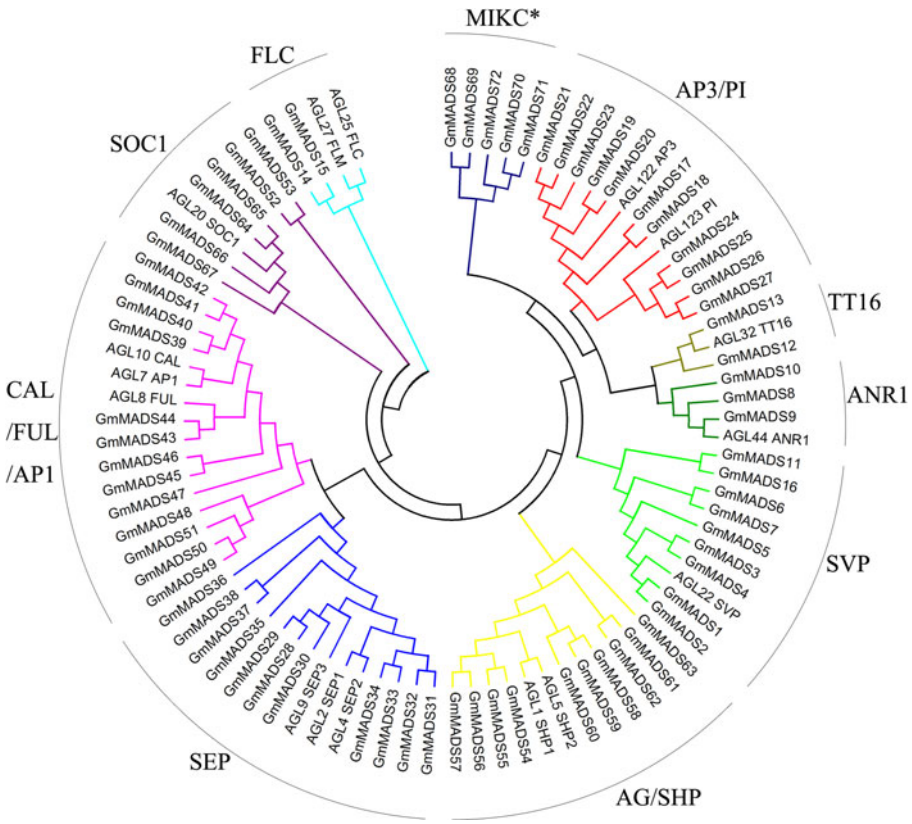


Fig. 1 Phylogenetic tree of soybean and *Arabidopsis* type II MADS-box genes

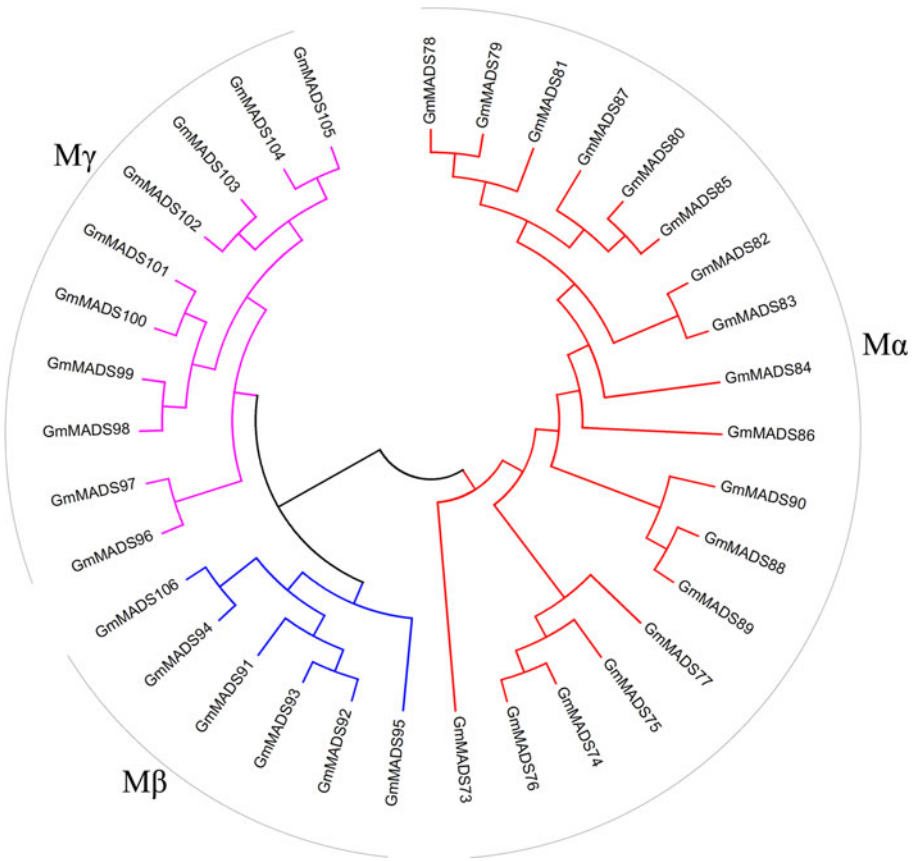


Fig. 2 Phylogenetic tree of soybean and *Arabidopsis* type I MADS-box genes

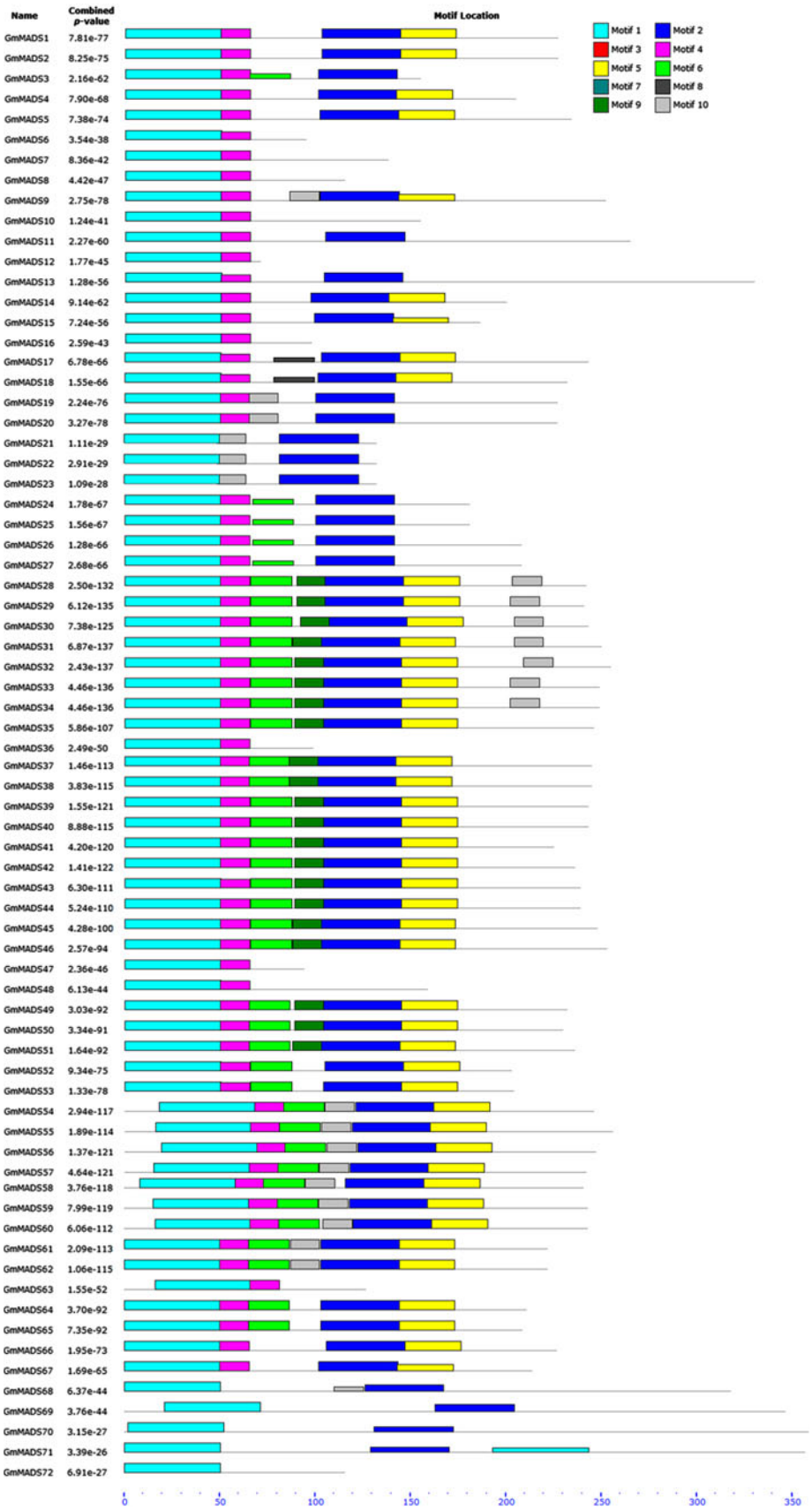
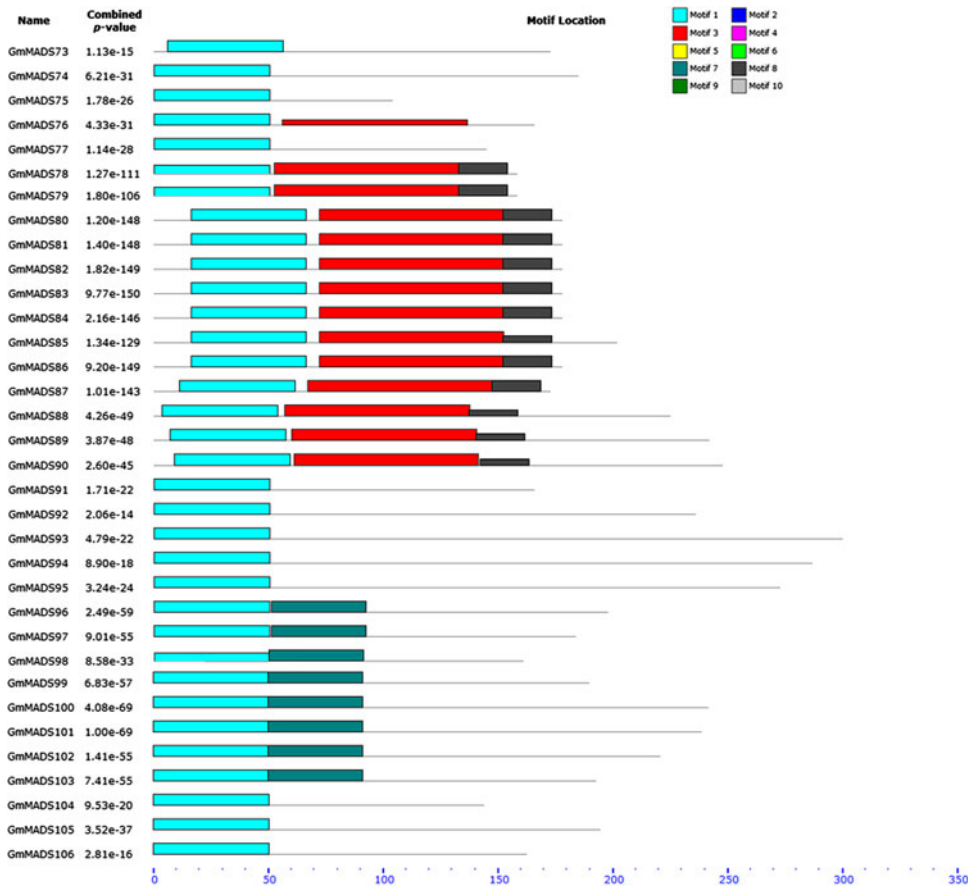


Fig. 3 Distribution of conserved motifs in soybean MADS-box proteins identified using the MEME search tool. Different motifs are indicated by *different colors*, and the names of all members and combined *p* values from different groups are shown on the *left side* of the figure. The order of the motifs corresponds to the position of the motifs in individual protein sequences. (Color figure online)

Fig. 3 continued



Expression pattern analysis of soybean MADS-box genes

Because high-throughput sequencing and gene expression analyses have been performed on many types of soybean tissues at various developmental stages, many soybean genetic sequences are available in the NCBI database. We therefore examined soybean transcriptome sequencing data from different tissues and developmental stages and collected all of the available MADS-box gene expression data. These genes were clustered into five groups based on their expression patterns. Most soybean MADS-box genes have a broad expression spectrum, except for group C, which has few transcripts across all tissues and developmental stages. There are 24 group C genes, most of which are type I MADS-box genes, mainly from groups Mβ and Mγ. Group A includes 28 members that are highly expressed during embryo and floral development. Most group A genes are in the MIKC^c group, which includes most SEP and AG/SHP subfamily genes, which is consist with previous reports [43, 44]. MADS-box genes from the Mα group, which are group B genes, have specific expression patterns at several stages of embryo development. Group D genes (21 genes), mainly comprising CAL/FUL/AP1 and AP3/PI subfamilies genes, are expressed in different tissues, including floral buds, leaves, roots and stems. However, these genes are mainly expresses in floral buds, as previously observed [45]. Finally, group E (15 genes), comprising the subfamilies SOC1, FLC, ANR1 and SVP, have broad expression patterns. These genes are expressed from embryo development to floral development and even in dry seeds. An SOC1 homolog gene was identified in soybean, and its expression pattern was similar to that of SOC1 in *Arabidopsis* [46].

Discussion

By performing comparative genomic and phylogenetic analysis, we determined that some characteristics of MADS-box genes that are present in species such as *Arabidopsis* and rice are present in soybean as well. However, there are more MADS-box type II genes in soybean than in other species. There are 45 and 43 type II MADS-box type genes in *Arabidopsis* and rice, respectively, whereas soybean has 72 MADS-box type II genes. This suggests that soybean type II MADS-box genes have undergone a higher rate of duplication and/or a lower rate of gene loss after

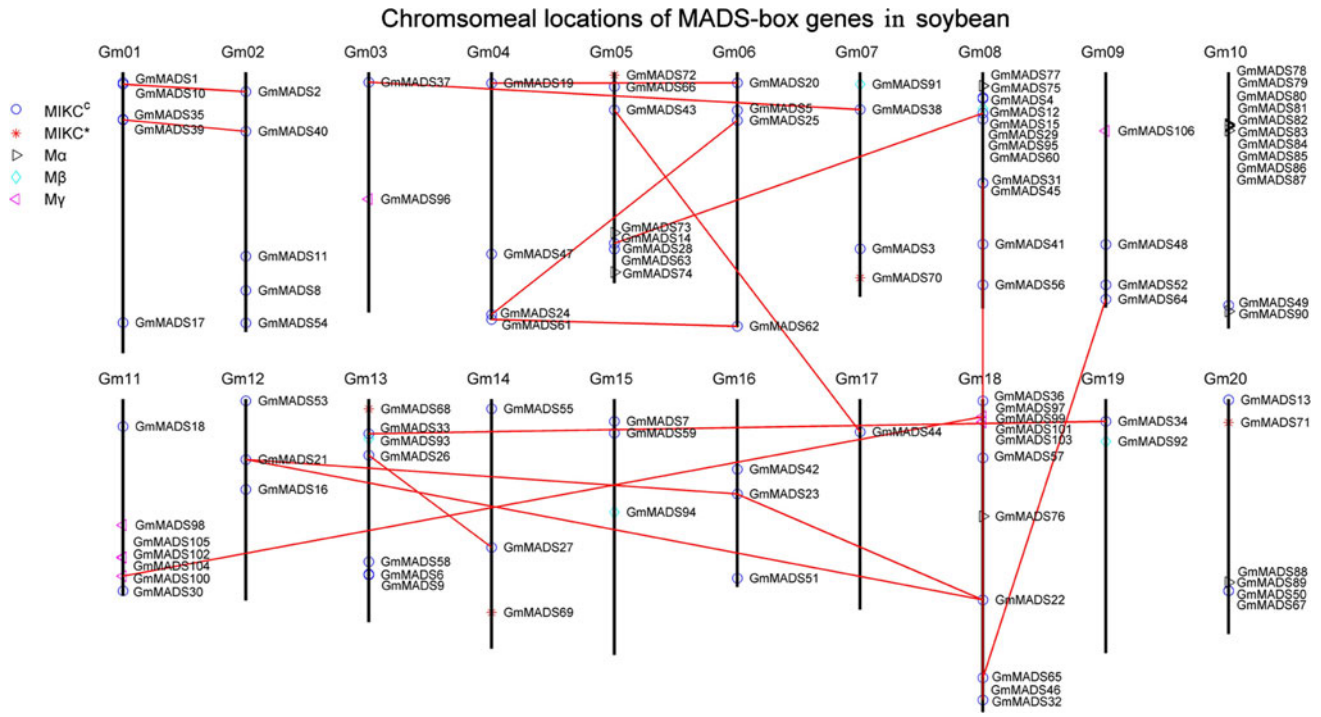


Fig. 4 Chromosomal locations of soybean MADS-box genes. The markers represent the group to which each MADS-box gene belongs, and the *red lines* connecting the MADS-box genes indicate segmental duplications on soybean chromosomes. (Color figure online)

duplication [47, 48]. Among soybean MADS-box genes, most are MIKC genes, while only five are MIKC* genes. Therefore, there are many MIKC genes in soybean, including complete subfamilies of MIKC genes. This indicates that the MIKC group genes are conserved among different species, while other groups are not. Therefore, MIKC group genes are enhanced in soybean, while other MADS-box genes have been lost or were never present. This notion is supported by the distribution pattern of introns; soybean MADS-box genes contain a bimodal distribution of introns. MIKC and MIKC* genes have many introns, while type I (Mα, Mβ and Mγ) genes lack introns or have single introns. Genes with many introns are considered to be conservative. In addition, different groups of soybean MADS-box genes have different distribution patterns in the genome. MIKC genes are distributed across all 20 soybean chromosomes, while other groups, including Mα, Mβ, Mγ, are mainly located on chromosomes 10 and 11. These two chromosomes account for approximately 9.5 % of the genome and contain 4.5 % of the MIKC genes, whereas they contain 47.1 % of type I genes, including Mα, Mβ and Mγ. In eukaryotic transcription factor families, genes expand by gene duplication, while type I and type II group MADS-box genes expand through different mechanisms [1, 49]. By performing homology analysis between genes, we determined that all type II gene duplications occurred between two different chromosomes via a process known as segmental duplication. However, the duplication of type I genes tends to occur within a single chromosome, which is known as internal chromosome duplication or tandem duplication. Both tandem and segmental duplication have played major roles in MADS-box gene expansion in the soybean genome. Because tandem duplication or internal chromosome duplication occur more frequently in soybean MADS-box type I genes, these genes originated and diverged more recently than type II genes, which indicates that type II genes are more conservative than type I genes in soybean.

Thus, different groups of MADS-box genes, with different duplication patterns, are under different evolutionary constraints. Therefore, type II genes are conservative and accumulate in the soybean genome by natural and artificial selection. Finally, the soybean genome contains 72 MADS-box genes, more than are present in *Arabidopsis*. Why are type II genes evolutionarily conserved in the soybean genome? Soybean MADS-box genes are selected or affected by many evolutionary constraints, which are still unknown. However, these genes are linked to soybean growth and development. As expected, soybean MADS-box type II genes are more highly expressed at different developmental stages and in different tissues than type I genes. There are 24 MADS-box genes clustered into group C (Fig. 5), which have very low expression levels or are not expressed. Among these genes, eight (11.1 %, 8/72) are type II, while 16 (47.1 %, 16/34) are type I genes. In addition, the 12 group B genes, which belong to group Mα

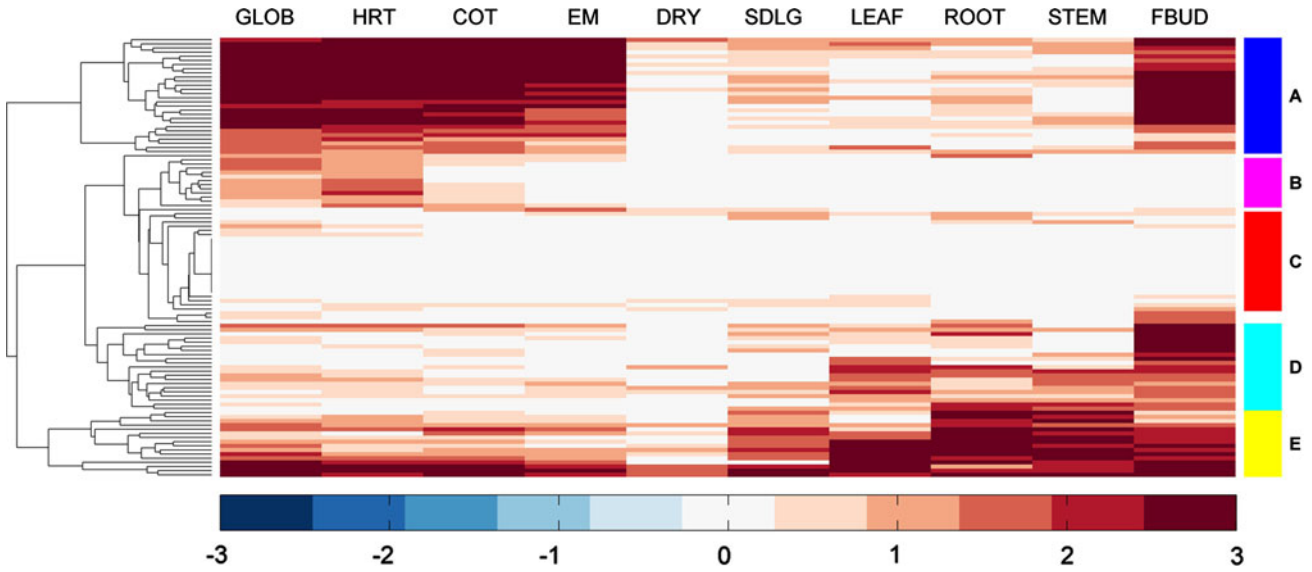


Fig. 5 Heat map of soybean MADS-box gene expression obtained from high-throughput sequencing data. Sources of the samples are as follows: GLOB (globular-stage embryos), HRT (heart-stage embryos), COT (cotyledon-stage embryos), EM (early maturation-stage embryos), DRY (dry soybean seeds), SDLG (whole seedlings 6 days after imbibition), LEAF (leaves), ROOT (roots), STEM (stems) and FBUD (floral buds)

of type I, are expressed at low levels during embryo development. These genes are mainly located on chromosome 10, and they have similar expression patterns. Groups A, D and E are type II genes and are highly expressed across all developmental stages and in all tissues. Among these genes, most are highly express in floral buds or during embryo development, while some of these genes are also present at other organs, such as leaves, roots and stems. The genes that are expressed in floral buds and during embryo development are mainly from large subfamilies. For example, group A genes, which are expressed during four developmental stages and in floral bud tissue, mainly comprise subfamilies AG/SHP (12) and SEP (9). Group D, which comprises subfamilies CAL/FUL/AP1 (13) and AP3/PI (11), are expressed in different organs, with the highest level of expression in floral buds. However, group E genes are expressed across every developmental stage and in every tissue, even in dry seeds and whole seedlings. Group E comprises small subfamilies such as AGL12 (2), SOC1 (4) and ANR1 (3), which is similar to other species [41, 50]. This indicates that soybean MADS-box gene redundancy and functional diversity also exists in subfamilies of soybean MADS-box type II genes. Subfamilies CAL/FUL/AP1, AP3/PI, SEP and AG/SHP exhibit high degrees of gene redundancy, while FLC, AGL6, AGL12, AGL15, SOC1, TT16 and ANR1 contain few members, without redundancy. Therefore, large subfamilies have conservative functions that are likely present in all species, while small subfamilies exhibit little conservation and may have been lost in some species. For example, the AGL12 and FLC genes are absent in cucumber, but their functions are not conservative and are therefore performed by other genes.

Conclusions

In summary, we have identified 106 MADS-box genes in soybean, which are clustered into the type I and type II groups. These results are consistent with previous studies. However, the soybean genome has more MADS-box type II genes than *Arabidopsis* or rice. By phylogenetic analysis, genes structure analysis and chromosome location analysis, we found that type II genes are more conservative than type I genes, which has played a major role in soybean development, growth and flower formation. In addition, soybean MADS-box gene expression patterns were investigated using high-throughput sequencing data. Type II MADS-box genes are highly expressed during seed development and in floral buds, while type I genes are not expressed or are expressed at low levels during seed development or in floral buds. While we were able to classify soybean MADS-box genes and elucidate the evolution and expression patterns of these genes, further functional analysis of these genes will be required to advance our understanding of their biological roles in soybean.

Acknowledgments This work was supported by Grants from the National Major Project for Cultivation of Transgenic Crops (2011ZX08004-002-003), the Science and Technology Research Plan from Education Department of Heilongjiang Province (12521149), the Natural and Science Foundation of Heilongjiang Province (QC2011C108), the Science and Technology Innovative Research Team in Higher Educational Intuitions of Heilongjiang Province

(2010TD10), and the Innovation Research Group of Harbin Normal University (KJTD201102).

References

1. Becker A, Winter K-U, Meyer B, Saedler H, Theißen G (2000) MADS-box gene diversity in seed plants 300 million years ago. Mol Biol Evol 17(10):1425–1434
2. Messenguy F, Dubois E (2003) Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. Gene 316(0):1–21
3. Riechmann JL, Krizek BA, Meyerowitz EM (1996) Dimerization specificity of *Arabidopsis* MADS domain homeotic proteins A-PETALA1, APETALA3, PISTILLATA, and AGAMOUS. Proc Natl Acad Sci USA 93(10):4793–4798
4. Norman C, Runswick M, Pollock R, Treisman R (1988) Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell 55(6):989–1003
5. Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM (1990) The protein encoded by the *Arabidopsis* homeotic gene agamous resembles transcription factors. Nature 346(6279):35–39
6. Theißen G, Kim J, Saedler H (1996) Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. J Mol Evol 43(5):484–516
7. Alvarez-Buylla ER, Liljegren SJ, Pelaz S, Gold SE, Burgeff C, Ditta GS, Vergara-Silva F, Yanofsky MF (2000) MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. Plant J 24(4):457–466
8. De Bodt S, Raes J, Van de Peer Y, Theißen G (2003) And then there were many: MADS goes genomic. Trends Plant Sci 8(10): 475–483
9. Cho S, Jang S, Chae S, Chung K, Moon Y-H, An G, Jang S (1999) Analysis of the C-terminal region of *Arabidopsis thaliana* APETALA1 as a transcription activation domain. Plant Mol Biol 40(3):419–429
10. Yang Y, Fanning L, Jack T (2003) The K domain mediates heterodimerization of the *Arabidopsis* floral organ identity proteins, APETALA3 and PISTILLATA. Plant J 33(1):47–59
11. Henschel K, Kofuji R, Hasebe M, Saedler H, Münster T, Theißen G (2002) Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. Mol Biol Evol 19(6):801–814
12. Ma H, Yanofsky MF, Meyerowitz EM (1991) AGL1-AGL6, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. Genes Dev 5(3):484–495
13. Davies B, Egea-Cortines M, de Andrade Silva E, Saedler H, Sommer H (1996) Multiple interactions amongst floral homeotic MADS box proteins. EMBO J 15(16):4330–4343
14. Parenicová L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B et al (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*. Plant Cell 15(7):1538–1551
15. Diaz-Riquelme J, Lijavetzky D, Martinez-Zapater JM, Carmona MJ (2009) Genome-wide analysis of MIKCC-type MADS box genes in grapevine. Plant Physiol 149(1):354–369
16. Kramer EM, Irish VF (1999) Evolution of genetic mechanisms controlling petal development. Nature 399(6732):144–148
17. Honma T, Goto K (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature 409(6819): 525–529
18. Kofuji R, Sumikawa N, Yamasaki M, Kondo K, Ueda K, Ito M, Hasebe M (2003) Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. Mol Biol Evol 20(12):1963–1977
19. Michaels SD, Amasino RM (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11(5):949–956
20. Hartmann U, Höhmann S, Nettesheim K, Wisman E, Saedler H, Huijser P (2000) Molecular cloning of SVP: a negative regulator of the floral transition in *Arabidopsis*. Plant J 21(4):351–360
21. Samach A, Onouchi H, Gold SE, Ditta GS, Schwarz-Sommer Z, Yanofsky MF, Coupland G (2000) Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. Science 288(5471):1613–1616
22. Scortecci KC, Michaels SD, Amasino RM (2001) Identification of a MADS-box gene, FLOWERING LOCUS M, that represses flowering. Plant J 26(2):229–236
23. Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, Amasino RM (2003) AGL24 acts as a promoter of flowering in *Arabidopsis* and is positively regulated by vernalization. Plant J 33(5):867–874
24. Kaufmann K, Melzer R, Theißen G (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene 347(2):183–198
25. Alejandra Mandel M, Gustafson-Brown C, Savidge B, Yanofsky MF (1992) Molecular characterization of the *Arabidopsis* floral homeotic gene APETALA1. Nature 360(6401):273–277
26. Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR (1993) Control of flower development in *Arabidopsis thaliana* by APETALA1 and interacting genes. Development 119(3):721–743
27. Gu Q, Ferrandiz C, Yanofsky MF, Martienssen R (1998) The FRUITFULL MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. Development 125(8):1509–1517
28. 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(6783):200–203
29. Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL, Yanofsky MF (2000) SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. Nature 404(6779):766–770
30. Nesi N, Debeaujon I, Jond C, Stewart AJ, Jenkins GI, Caboche M, Lepiniec L (2002) The TRANSPARENT TESTA16 locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. Plant Cell 14(10):2463–2479
31. 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(9):1435–1447
32. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278): 178–183
33. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R et al (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34(suppl 1):D247–D251
34. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39(Suppl 2):W29–W37
35. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680
36. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software wersion 4.0. Mol Biol Evol 24(8):1596–1599

37. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34(Suppl 2):W369–W373
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
39. Zhang H, Jin J, Tang L, Zhao Y, Gu X, Gao G, Luo J (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. Nucleic Acids Res 39(suppl 1):D1114–D1117
40. Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M (2005) Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. Plant Mol Biol 59(1):191–203
41. Zhao Y, Li X, Chen W, Peng X, Cheng X, Zhu S, Cheng B (2011) Whole-genome survey and characterization of MADS-box gene family in maize and sorghum. Plant Cell Tissue Organ Cult 105(2):159–173
42. Hu L, Liu S (2012) Genome-wide analysis of the MADS-box gene family in cucumber. Genome 55(3):245–256
43. Thakare D, Tang W, Hill K, Perry SE (2008) The MADS-domain transcriptional regulator AGAMOUS-like 15 promotes somatic embryo development in *Arabidopsis* and soybean. Plant Physiol 146(4):1663–1672
44. Wang Y, Zhang X, Liu Z, Zhang D, Wang J, Liu D, Li F, Lu H (2012) Isolation and characterization of an AGAMOUS -like gene from *Hosta plantaginea*. Mol Biol Rep 39(3):2875–2881
45. Chi Y, Huang F, Liu H, Yang S, Yu D (2011) An APETALA1-like gene of soybean regulates flowering time and specifies floral organs. J Plant Physiol 168(18):2251–2259
46. Zhong X, Dai X, Xv J, Wu H, Liu B, Li H (2012) Cloning and expression analysis of GmGAL1, SOC1 homolog gene in soybean. Mol Biol Rep 39(6):6967–6974
47. Leseberg CH, Li A, Kang H, Duvall M, Mao L (2006) Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. Gene 378:84–94
48. Nam J, Kim J, Lee S, An G, Ma H, Nei M (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci USA 101(7):1910–1915
49. Theissen G, Becker A, Di Rosa A, Kanno A, Kim JT, Münster T, Winter K-U, Saedler H (2000) A short history of MADS-box genes in plants. Plant Mol Biol 42(1):115–149
50. Li H-L, Wang Y, Guo D, Tian W-M, Peng S-Q (2011) Three MADS-box genes of *Hevea brasiliensis* expressed during somatic embryogenesis and in the laticifer cells. Mol Biol Rep 38(6): 4045–4052