1 Identification and characterization of MADS-box gene family in pigeonpea, their wild relatives and 2 comparative phylogenetic analysis with grain legumes

3 Kuldeep Kumar¹, Harsha Srivastava¹, Antara Das¹, Kishor U. Tribhuvan¹, Kumar Durgesh², Rekha Joshi²,
4 Amitha Mithra Sevanthi¹, Pradeep Kumar Jain¹, Nagendra Kumar Singh¹, Kishor Gaikwad^{1*}

5 ¹ICAR-National Institute for Plant Biotechnology, New Delhi, India.

6 ²Division of Genetics, ICAR-IARI, New Delhi, India

7 *Corresponding author's email address: kish2012@nrcpb.org

8 Abstract

9 MADS-box genes are classes of transcription factors involved in various physiological and developmental
10 processes in plants. Here, genome wide identification of MADS-box genes was done in *Cajanus cajan*,
11 identifying 102 members, classified into two different groups based on their gene structure. The gene
12 based phylogeny of *C. cajan* MADS-box genes, and some grain legumes was developed to detect their
13 gene homologs in *C. cajan*. The status of all these genes was analyzed in three wild relatives i.e. *C.*
14 *scarabaeoides*, *C. platycarpus* and *C. cajanifolius*. A total of 41 MADS-box genes were found to be
15 missing in wild type cultivars hinting towards their role in domestication and evolution. Single copy of
16 Flowering locus C (*FLC*) and Short vegetative phase (*SVP*), while three copies of Suppressor of
17 activation of Constans 1 (*SOCI*) was found to be present. One *SOCI* gene i.e. CcMADS1.5 was found to
18 be missing in all wild relatives, also forming separate clade in phylogeny, revealing its origin through
19 duplication followed by divergence, and role in domestication. Expression profiling of major MADS-box
20 genes involved in flowering was done in different tissues viz vegetative meristem vegetative leaf,
21 reproductive meristem and reproductive bud. Gene based time tree of *FLC* and *SOCI* gene dictates their
22 divergence from *Arabidopsis* before 71 and 23 million year ago (mya) respectively. This study provides
23 valuable insights into the functions, characteristics and evolution of MADS-box proteins in grain legumes
24 with emphasis on *C. cajan*, which may help in further characterizing these genes.

25 **Keywords:** MADS-box, flower induction, *FLC*, *SOCI*, *SVP*

26 1. Introduction

27 Pigeonpea (*Cajanus cajan* (L.) Millsp.), is a high protein (20–23%) food legume, important for
28 providing nutritional security of the Indian subcontinent. Globally, it ranks sixth in the grain
29 legume production and is grown across 7.02 million hectares (mha) with yearly production of
30 6.80 million tons and productivity of 969.1 kg/ha. In India alone, *C. cajan* is cultivated in about
31 5.38 mha area (75% of global area) with an average production of 4.87 million tonnes,

accounting for 71% of the global production [1] It is diploid legume species with $2n=22$ chromosome with an approximate 850 Mb genome size.

Flowering is an important event in the life cycle of plants. It is regulated by many endogenous and exogenous factors like age, hormonal balance within the plant, carbon to nitrogen ratio, photoperiod, temperature etc. [2-3]. The endogenous factors responsible for flowering are regulated at both translational and post translational levels, while environmental factors regulate flowering through external stimulus [4]. MADS-box gene family is one of the most important gene family involved in flowering and homeotic organ development in plants. The loss of function mutants for MADS genes are responsible for abnormal development of flowering organs [5]. MADS family includes a set of highly conserved genes [6] where the N-terminal (55 to 60 amino acid) are highly conserved across the family and “MADS” represent four gene families; Mini chromosomal maintenance-1 of *S. cerevisiae*; Agamous of *A. thaliana*; Deficiens of *A. majus* and Serum response factor of *Homo sapiens* [7]. Along with MADS domain some of the genes also code for a protein with an additional ‘K’ domain at the C-terminal end of protein [8]. The coding sequences of MADS-box family genes other than the MADS-box domain are variable and play a major role in various morphological and developmental process such as flowering, root development, fruits development etc [9].

In previous reports several methodologies have been used to classify MADS genes in plants. On the basis of presence of additional domains, MADS family genes were classified in two major groups; type I and type II MADS. The type I class of MADS contains only a SRF (SERUM RESPONSE FACTOR/ MADS) domain, while type II class contains an Intervening (I-box) domain, a Keratin like coiled coil (K-box) at the C terminal end of the protein and a MADS domain [10] (Figure 1). Type II MADS-box genes are also known as MIKC type genes as they have a MADS-box at the N-terminal and an I-box and K box towards the C terminal of the protein. On the basis of length of the K domain, Type II MADS-box genes have been further classified into two lineages $MIKC^C$ and $MIKC^*$ [11]. $MIKC^*$ has an extended K domain, while $MIKC^C$ has a normal length K domain. $MIKC^*$ genes are further divided into P and S monophyletic lineages [12]. Parenicová et al. [13], followed a different methodology for classification of MADS genes of *Arabidopsis thaliana*, type I MADS genes were grouped into three sub-groups $M\alpha$, $M\beta$ and $M\gamma$, and type II MADS were sub-grouped into $MIKC^C$ and $MIKC^*$ on the basis of their phylogenetic clustering. MADS-box genes were also divided into 13 different gene subfamilies, termed *AG*, *AGL2*, *AGL6*, *AGL12*, *AGL15*, *AGL17*, *DEF*, *FLC*, *GGM13* (Bsister), *GLO*, *SQUA*, *SVP/STMADS11*, and *SOC1/TM3* on the basis of phylogenetic clustering [14]. Characterization of MADS genes have been carried out in many plants species including *A. thaliana* [15], Apple [16], Maize [17],

tomato [18] and rice [19] which has provided information regarding the evolution, conservation and function in diverse physiological and morphological traits.

Pigeonpea displays large diversity in terms of days to flowering, ranging from 55 days to 160 days. It is well known that flower development is controlled by various pathways that are regulated by hormones, day length duration and various environmental factors. Also, in case of pigeonpea, late duration crops are preferred due to their high yield potential. In Pigeonpea, genome wide study, classification and gene based phylogeny of MADS-box genes is lacking. Keeping this in mind, a comprehensive study of MADS-box family genes of *C. cajan* was designed. Three wild type cultivars, *C. scarabaeoides* (ICP 15747), *C. platycarpus* (ICP 15665) and *C. cajanifolius* (ICP 15629) were used to search for copy number variation of MADS-box genes in comparison with a cultivated variety. A comparative analysis has been done to study the evolution of MADS-box genes in several grain legumes including *Arachis hypogea*, *Glycine max*, *Cicer arietinum*, *Phaseolus vulgaris* and pigeonpea. A phylogenetic analysis among different grain legumes was conducted to study the evolution pattern and search for orthologs and paralogs on the similarity basis. As all grain legumes have evolved from a common ancestor, the gene family present in them would have diverged during the time of evolution [20]. Flowering is regulated by multiple genes having both positive and negative effects. Two of the key genes involved in flowering are Flowering locus C (*FLC*) and Suppressor of activation of Constans (*SOC*) which are MADS-box proteins that regulate the process of flowering in negative and positive manner respectively. Hence, analysis of their point of divergence may provide insight into their evolution as pigeonpea has considerable diversity in terms of flowering. As MADS-box genes are one of the key players of flower initiation and development, comprehensive analysis of their number, expression pattern and phylogenetic relationship will aid our understanding of this complex pathway.

2. Material and methods

2.1 Identification, domain prediction and classification of MADS-box proteins

For genome-wide identification of MADS-box containing genes, protein sequences of *C. cajan* was downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). The conserved MADS-box domain sequence (pfam identifier PF00319) representing SRF-type transcription factor (DNA-binding and dimerization domain) consisting of a core domain of 90 amino acids was downloaded from pfam database (<https://pfam.xfam.org/>) in Stockholm format. MADS-box domain containing proteins present in the *C. cajan* protein database were identified using HMMER (v 3.2.1). Simultaneously, all the MADS-box genes reported in Legume Information System (LIS) database (<https://legumeinfo.org/>) were obtained using gene search through description option. Total MADS-box protein identified from HMMER and LIS

database were compared and redundant genes were removed using the CD-HIT webserver (<http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi>). For the study of evolutionary relationship of *C. cajan* MADS protein with other legume species, MADS protein sequences were obtained for other legume species including *A. hypogaea*, *C. arietinum*, *G. max* and *P. vulgaris* from NCBI using similar methodology.

All the isolated protein sequences were analyzed for presence of MADS-box and other domains present in the protein sequences using NCBI Batch CD-search (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>) with e-value value 0.01 [21]. The classification of MADS-box containing proteins into type I and type II MADS were based on presence of K domain, besides the MADS box domains. Type I MADS box genes contain only MADS box domain, while type II contains K box domain in addition to MADS box. Type I MADS-box genes were further grouped as M α and M β on the basis of their phylogenetic clustering.

2.2 Identification of chromosomal position, nomenclature and schematic representation of MADS genes

The chromosomal position of all the identified MADS proteins were obtained by BLAST search tool available at legume information system (LIS) database. All the genes were renamed based on their position on a specific chromosome and in an ascending order. While renaming the MADS genes of *C. cajan*, all the information such as genus name, gene family name, chromosome number and position were taken into account, for example in CcMADS1.1, Cc represent *C. cajan*, MADS represent presence of MADS domain, 1 depicts its presence on chromosome and .1 depict its first position on chromosome1. Subsequent MADS genes on chromosome 1 were renamed as CcMADS1.2, CcMADS1.3 and so on. Genes present on scaffolds were also named similarly way but instead of CcMADS, CcMADSSca nomenclature was used. The theoretical molecular weight and pI of the proteins was calculated using Compute pI/Mw tool (https://web.expasy.org/compute_pi/), with the resolution set to 'average'. The chromosomal position of all the MADS genes were represented schematically using mapchart v 2.0 tools [22]. Plant genome duplication database was used to check the duplication of each 102 genes [23].

2.3 Phylogenetic analysis, gene structure prediction and copy number variations in *C. cajan* wild relatives

Multiple sequence alignments and preparation of un-rooted tree through neighbor joining method (NJ) of full length proteins were performed using MEGA-X [24] software with 1000 bootstrap replicates. The LIS coordinates of each gene were used to obtain complete gene sequences using bedtools. The GSDS2.0

web server (<http://gsds.cbi.pku.edu.cn/>) was used for gene structure prediction [25]. Raw sequence reads of three wild relatives *C. scarabaeoides* (ICP15747), *C. platycarpus* (ICP15665) and *C. cajanifolius* (ICP15629) were downloaded from NCBI. The reference-based assemblies of these three genotypes were performed using GATK pipeline (<https://software.broadinstitute.org/gatk/>). *C. cajan* MADS genes orthologs present in these three wild relatives were searched using local BLAST search with default parameters, followed by their translational analysis through AUGUSTUS gene prediction tool web server (<http://bioinf.uni-greifswald.de/augustus/submission.php>). Blast2go analysis [26] was performed for annotation function of MADS genes identified from *C. cajan* and wild relatives.

2.4 Gene expression profiling of MADS genes

Expression profile of selected MADS genes across the four tissues, vegetative leaf, vegetative meristem, reproductive leaf and reproductive meristem of *C. cajan* were analyzed using ‘Gene expression atlas data’ [27]. The FPKM values of each gene corresponding to different tissues were retrieved from gene expression atlas data and used for construction of heatmap using gplots package of ‘R’ software. For experimental validation of gene expression profiling of important MADS-box gene through qRT-PCR, total RNA was isolated using Spectrum Plant Total RNA Kit (MERCK or SIGMA, USA) from Vegetative Shoot Apical Meristem (SAM), reproductive SAM, reproductive bud and reproductive leaf tissues sampled from ICP 20338 genotype of *C. cajan*. DNase (Thermo Fisher Scientific, Waltham, USA) treatment was performed to remove DNA contamination from the isolated RNA. Quantification of total RNA was carried out using Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, USA). Integrity and quality of RNA were checked by electrophoresis on 1.2% denaturing agarose gel. cDNA synthesis was performed using Revert Aid first strand cDNA synthesis kit (Thermo Fisher Scientific, Waltham, USA) as per manufacturer's protocol.

qRT-PCR analysis was performed using CFX96 touch real time PCR detection system (Bio-rad) using Brilliant III Ultra-Fast SYBR Green qPCR Master Mix kit (Agilent Technologies, USA). qRT-PCR primers were designed using primer quest software of Integrated DNA Technologies (IDT). Each qRT-PCR reaction contained 2 µl of cDNA, 0.5 µl of each primer, 10 µl SYBR green master mix and 7.5 µl DEPC water. *Tubulin-3α* gene specific primers were used as internal control. Reaction conditions were set at 94°C for 3 min, followed by 40 cycles of 94°C for 30 s, 60°C for 15 s and 72°C for 20 s. Similar reaction condition were followed for melting curve analysis (56°C to 95°C) with increasing annealing temperature at 0.5°C in every 10s. All primer information used for real time PCR is given in Table S1. Relative expression pattern of each sample were analyzed by the ‘comparative Ct method’[28]. The

internal control for normalize Ct values U6 snRNA was used. All reactions were conducted in three technical replicates.

2.5 Molecular clock for estimating divergence time and STRING analysis for FLC AND SOC1 gene

As each gene member of a family has a different time of divergence, we have only focused on two genes, *FLC* and *SOC1* for the study of divergence time due to their importance in flower induction and development. Protein sequences for these two genes from *Arabidopsis thaliana*, *Glycine max* and *C. cajan* were taken to develop their molecular clocks to know the divergence time for them separately. MEGA-X software with RelTime method was used to develop the molecular clock using default parameters [24]. RelTime method has better analysis potential and commonly used method for molecular clock analysis because it performs better than other methods when applied to very large computer simulated datasets where evolutionary rates vary extensively among lineages [29]. FLC genes of *A. thaliana* were used as an outgroup in the analysis. Function protein association network was developed for important regulators of flowering time MADS-box genes including *FLC*, *SOC1*, and *SVP* using STRING database v10 [30].

3. Result and Discussion

3.1 Genome wide identification of *C. cajan* MADS genes

A total of 82 and 102 MADS-box domain containing proteins were found in pigeonpea using HMMER analysis and LIS database respectively. CD hit analysis confirmed that all 102 members were unique and non-redundant; hence, all 102 MADS genes from LIS were taken for further analysis. The complete detail of each gene including gene id, LIS identifier, protein identifier, gene coordinates, theoretical *pI*, theoretical molecular weight and functions are given in Table S2. All the identified 102 MADS genes were renamed using the nomenclature system discussed method section. Identification of MADS genes in other legumes including *A. hypogea*, *C. arietinum*, *G. max*, *P. vulgare* and *A. thaliana* revealed the presence of 89, 83, 129, 77 and 120 MADS-box genes respectively. Kumar et al. [15] identified dormancy associated loci and flowering locus C gene in apple and studied their role in flowering as well dormancy associated pathway. Likewise, *HUA ENHANCER 4* that delays flowering by upregulating the MADS-box repressor genes *FLC* and *MAF4* has also been reported from these studies [14].

3.2 Identification of protein domains and classification of MADS genes

The results of batch CD search revealed that all of the identified proteins possessed the conserved MADS-box domain with few exceptions which are noted below (Figure 2 and 3). A total of 63 type I and 39 type II MADS-box genes were found in *C. cajan* in our analysis. The same methodology was used by [9] for the classification of MADS-box gene in *Arabidopsis*.

Among type II members, 12 genes including CcMADS1.6, CcMADS3.1, CcMADS3.6, CcMADS6.3, CcMADS8.1, CcMADS11.3, CcMADSsca10.1, CcMADSsca11.1, CcMADSsca13.2, CcMADSsca17.2, CcMADSsca18.1 and CcMADSsca21.1 possessed only K box and lacked the trademark SRF domain. Although these proteins lacked the MADS domain, they were annotated and reported as MADS-box family members by Blast2GO analysis (Table S3) as well as in many previous studies [31, 32, 33]. CcMADS1.6, a *SOCI* gene positively regulates floral transition of meristematic stem [31]. Other genes lacking MADS domain including *AGL8*, *AGL24* and *JOINTLESS* are also vital MADS-box orthologs. *AGL 8* is a transcription factor that regulates meristem identity redundantly with *CAULIFLOWER* and *APETALA 1* [34]. *AGL24*, *SVP* and *JOINTLESS* shares extensive amino acid sequence similarity and belong to same clade [35]. Usually *SVP* delays flowering by working in coordination with *FLC*, but its paralog *AGL 24* contradicts its function and promotes flowering [36].

3.3 Gene structure prediction

Gene structure prediction of type I genes shows that they are comparatively small in size and have few introns. In type I MADS-box genes only, CcMADS1.3, CcMADS11.5, CcMADS11.9, CcMADSsca16.1, CcMADSsca17.1, CcMADSsca19.1, CcMADSsca21.2, CcMADSsca23.1, CcMADSsca24.1, CcMADSsca26.1, CcMADSsca4.2 and CcMADSsca9.6 were found to possess introns (Figure 4). Similar analysis of type II gene indicates that presence of introns is more common (Figure 5) with the exception of CcMADS11.7 (MADS-box protein *CMB1*-like). This type of bimodal distribution of introns has also been reported in *Brassica napus* [37] and rice [18]. Most of the members belonging to the common clade in phylogeny show a similar gene structure.

3.4 Chromosome wise distribution of MADS-box gene

The pictorial representation of all genes based on their chromosomal position is given in Figure S1. Chromosome 2 contains the highest number of MADS-box genes (13 genes), while chromosome 5 and 10 do not contain any MADS-box genes. All members are widely distributed throughout the genome and several genes were found to be clustered as tandem duplications (Figure S1).

3.5 Gene based phylogeny of *C. cajan* MADS-BOX gene

A gene based phylogeny was developed for all *C. cajan* MADS-box genes to examine their evolution pattern and presence of paralogous genes (Figure 6). The phylogenetic tree was divided into three major clades, clade I contains 12 MADS-box gene, clade II contains 22 MADS genes, and clade III contains the most MADS-box genes (68 MADS-box genes). Clade II is divided into two subgroups, subgroup 1 contains mixture of AGL proteins and PHERES proteins while in subgroup II all members were exclusively AGL80 proteins or their isoforms except CcMADS9.1 (a *PHERES* 2 like protein). Clade I was solely comprised of type II while clade II has only type I MADS-box genes. Clade III on the other hand has a mixture of both type I and II MADS-box genes. Clade III has diverse set of proteins; which can be divided into 10 different subgroups. Subgroup 1 exclusively has type I MADS-box proteins while rest of the subgroups have both type I and type II MADS-box proteins. Phylogenetic analysis of Type I and II MADS-box genes resulted in monophyletic cluster in animals and yeast but not plants [11], our results are also in accordance with it.

A single copy of Flowering locus C (CcMADSsca1.3) is present in *C. cajan* as well as its wild relatives and fall separately into subgroup 5 of clade III, sharing homology with CcMADS2.5 (AGL 15 protein). *SOC1* is an important positive regulator of flowering and belongs to type II MADS domain (i.e. *SOC1* protein). Through annotation analysis, *SOC1* gene was found to have three paralogs, CcMADS1.5, CcMADS1.6 and CcMADS7.1. The CcMADS1.6 is grouped in clade I while other two paralogs are in clade III. Another repressor of flowering namely Short Vegetative Phase gene (*SVP*) is present in single copy on chromosome 6. Interestingly it falls in the same subgroup III of clade III as *JOINTLESS* protein (CcMADS3.6, CcMADS6.3, CcMADS8.1 and CcMADS9.1). It has been previously reported that *JOINTLESS* is a homolog of *SVP* in some plants such as [38].

3.6 Identification of *C. cajan* MADS gene orthologs from its wild relatives

Identification of *C. cajan* MADS gene orthologs present in its three wild relatives viz *C. scarabaeoides* (ICP15747), *C. platycarpus* (ICP15665) and *C. cajanifolius* (ICP15629) revealed the presence of 71, 69 and 68 MADS-box genes in each respectively. It was found that 31 to 34 MADS-box genes were missing in the three wild relatives compared to cultivated variety. Status of all these genes in term of functions, presence/absence is given in Table S3.

In all wild type genotypes CcMADS1.6 (a *SOC1* protein), CcMADS6.6 (floral homeotic agamous like gene), CcMADS9.3 (floral homeotic agamous like gene), CcMADSsca1.2 (AGL 9), CcMADSsca9.1 (floral homeotic *PHERES* 2 like gene), CcMADSsca11.2 (AGL 21), CcMADSsca15.2 (AGL 80) and other

proteins as mentioned in Table S4 were absent. These genes may have risen by duplication from related counterpart gene in the cultivated variety. Selection pressure for duplication and divergence of these genes may be due to their vital role in flowering. Similar results were obtained during the analysis of SPL genes in *Arabidopsis* [37]. For example, *AGL 80* plays an important role in female gametophyte development and pollen interaction. In our analysis we found 15 *Agamous 80* or their isoforms, suggesting their importance in plant survival. Similarly, *CcMADS6.6* and *CcMADS9.3* (floral homeotic agamous genes) have 6 paralogs, which are probably involved in the control of organ identity during the early development of flowers. They are also required for development of stamens and carpels. Interestingly *CcMADSsca11.2* (*AGL 21*) is a single copy gene, which is reported to express in the quiescent centre of root cap and phloem cell and involved in salt and drought stresses [38-39]. *CcMADSsca9.1* (floral homeotic *PHERES 2* like gene) also has a role in gametophyte and seed development [40]. As the seed bearing capacity is significantly more in cultivated pigeonpea genotypes, this may be the reason for their duplication in cultivated varieties. Similarly *AGL9* is involved in early flower development [41]. A similar pattern has been observed [42] in *Arabidopsis* where loss of introns, domains, pre mature stop codon, alternative splice sites, poly A tails etc were found after duplication.

Combined phylogenetic analysis of all MADS-box genes in both cultivated and wild type genotypes reveal information about origin and similarity of unique genes present only in cultivated genotypes. A total of 41 MADS-box genes were missing in atleast one wild species analyzed in this study (Table S4). Most of the genes clustered with similar paralogs in *C. cajan* or orthologs in wild species (Figure S2). Only a few genes showed divergence in structure including *CcMADS1.6* which codes for a *SOC1* protein and was found only in *C. cajan*. It is in a separate out group in the phylogenetic tree that shows less similarity with other MADS-box protein. This gene may have evolved from duplication of an *SOC1* gene followed by mutation accumulation and divergence, which could account for its lack off similarity to other *SOC1* genes (*CcMADS1.5* and *CcMADS7.1*). All the MADS-box genes absent in the three wild relatives are highlighted in Figure S2 and Figure S3. Divergence in the sequence of all three paralogs of *SOC1* gene in pigeonpea was also confirmed by alignment of their protein sequences (Figure S3) and was found in accordance with phylogenetic analysis. A similar report on an ortholog of *SOC1* named as UNSHAVEN (*UNS*) sharing a sequence as well expression similarity was observed in *Petunia hybrida* [43]. To check the evolution pattern in grain legumes, MADS-box genes of *C. areitinum*, *A. hypogea*, *C.cajan*, *G. max* and *P.vulgaris* were used together to develop a common phylogenetic tree (Figure S4). This phylogenetic tree provides us interesting information about the evolution pattern of MADS-box genes and orthologs in other grain legumes. The phylogeny was divided into 15 clades (A-O) as highlighted by different colours. All orthologs were clustered in the same clades, for example FLC

gene of *C. cajan* clustered together with *FLC* of *G. max* (XP_014631160) in clade M. The members of each MADS-box protein sub-family ($M\alpha$, $M\beta$, $M\gamma$ MIKC^C and MIKC^{*}) of different plant species were clustered in the same clade. Similar results were also obtained while studying MADS-box genes in apple [15].

3.7 String analysis and selection of flowering related candidate genes

All three SOC1 proteins, CcMADS1.5, CcMADS1.6 and CcMADS7.1 were most similar to GLYMA18G45780.1, and were found to interact with two *FRIGIDA* like proteins (Figure 7A). These two *FRIGIDA* like proteins (Identifier: GLYMA04G38071.1 and GLYMA06G17005.1) are required for regulation of flowering time in late-flowering phenotypes [44]. They are involved in the enrichment of a WDR5A-containing COMPASS-like complex at the 'FLOWERING LOCUS C' that trimethylates histone H3 'Lys-4', leading to *FLC* up-regulation and an increase in RNA level [45].

Similarly, STRING analysis of CcMADS6.2 (SVP protein) showed it to be an ortholog of the *G. max* *SVP* gene. It interacts with another *SVP* isoform GLYMA06G10020.3 (Figure 7B). The interaction of *SVP* and *SOC1* is regulated by few amino acids that are highly conserved [46]. But no such interaction between *SVP* and *SOC* was detected in our analysis.

String analysis of CcMADSsca1.3 (*FLC*) revealed its interactions with histone modification complexes (GLYMA08G02490.4) and serine protein kinase (GLYMA12G08210.3) (Figure 7C). It has been reported that *FLC* is regulated at both transcription and post transcription level [47]. *FLC* inhibits *SOC1* by directly binding to its promoter, As STRING analysis detects protein-protein interaction [48], no interaction between *FLC* and *SOC1* was observed.

3.8 Expression analysis of selected genes using qRT-PCR

The comparative gene expression analysis of all selected MADS-box genes is provided in Figure 8. The gene expression analysis of all three SOC1 genes (CcMADS1.5, 1.6 and 7.1) reveals a similar profile, as all the three paralogs were found to be upregulated in reproductive SAM as compared to the vegetative SAM. Overexpression of *SOC1* suppressed the late flowering phenotype caused by the high expression of *FLC* in winter annual plants, indicating that *SOC1* is a downstream target of *FLC*. Also, loss of function and gain of function mutants of *SOC* exhibits late flowering and early flowering. Our expression analyses of *SOC1* are in accordance with [49]. Similarly, the repressor of *SOC1* gene i.e. *FLC* and *FRI* (activator of *FLC*) were upregulated in vegetative leaves as compared to reproductive leaves. *SVP*, another repressor of floral transition was showing similar expression pattern as *FLC*. *FRI* protein was involved in the activation of *FLC*, hence it also delays flowering [50]. *SOC1* gene acts as floral

integrator, it is activated by CO (a positive regulator of flowering) gene via FT while repressed by the repressor complex formed by the integration of FLC and SVP genes (both are negative regulator of flowering) [48].

Tissue specific local gene expression data developed [27] was used to develop heat map of only 79 MADS-box genes which were found in the expression atlas for vegetative SAM, vegetative leaf, reproductive SAM, reproductive leaf and reproductive bud to search for MADS-box genes involved in flowering. On the basis of expression pattern as shown in Heatmap (Figure S5), all genes can be divided into two major groups. Group I contain the genes which were repressed in reproductive bud and having negligible expression in reproductive leaf and reproductive SAM. Group I genes mainly includes *FLC*, *SVP* and *JOINTLESS* genes and their paralogs. *FLC* (CcMADSsca1.3) and *SVP* (CcMADS6.2) regulates flowering time by negatively regulating Flowering locus T (FT) gene [51], hence there suppressed expression in reproductive bud and high expression in vegetative leaf is justifiable. Group II contains SEPALLATA genes (CcMADSsca20.1), which is involved in floral identity development [52]; agamous, apetala 1 and Cauliflower genes which acts as a floral homoeotic gene [34, 53]. As all genes of group II are mostly involved in flower homeotic development, hence they are expressing in reproductive tissues like reproductive bud, reproductive leaf but not in vegetative tissues and vegetative leaf.

3.9 Time tree

Time tree reveals that the FLC gene of legumes used to construct the tree, diverged from *Arabidopsis* around 71 mya to *C. arietinum* lineage (Figure 9a), followed by their divergence into two different lineages around 49 mya, one with the *G. max* FLC gene and another with the *C. cajan* FLC gene. This result is also in accordance with the fact that *G. max* is believed to be the closest progenitor [54]. Similarly, all SOC1 genes diverged from *Arabidopsis* before 23 mya. Interestingly CcMADS7.1 appears to be similar to the SOC1 gene of *Glycine max* while CcMADS1.6 was close to the SOC1 gene of *Arachis hypogea*. CcMADS1.5 paralog of SOC1, which was absent in all three wild relatives, diverged separately from the other two paralogs of *C. cajan* SOC1 at the same time they diverged from *Arabidopsis* i.e. before 23 mya (Figure 9b).

4. Conclusion

This study determined 102 MADS-box transcription factors family genes in the *C. cajan* genome. The genome organization of *C. cajan* MADS-box genes revealed that genes are scattered throughout the genome. These genes were classified into two groups on the basis of presence of extra domains such as K-box domain. Type I MADS-box genes were more abundant than type II genes. Structural analysis

focusing on conserved MADS-box domains reveals that few MADS-box genes lacked the conserved domain, which may be due to duplication followed by divergence. Gene based phylogeny revealed clustering of paralogs genes. FLC and SVP were found in single copies while the SOC1 gene was found to have three functional copies. Of the three SOC1 genes paralogs, CcMADS1.5 formed a separate clade from two other counterparts in phylogeny as well as the time tree. Also, CcMADS1.5 was absent in all three wild relatives, suggesting the role of SOC genes in domestication. Comparative analysis revealed the absence of many MADS-box genes in wild species suggesting the important role of MADS-box genes in domestication and evolution. Gene based molecular clock for SOC1 and FLC genes shows their divergence from *Arabidopsis* around 23 and 71 mya, respectively. Overall our findings are largely consistent with previous studies in *Arabidopsis* and other crops and will provide the basal information for understanding the role of these genes in flowering in pigeonpea.

5. Competing of interest

The author(s) declares no competing interests.

6. Acknowledgment

We acknowledge the support provided by Director, ICAR-NIPB, for this work.

7. Author contribution

KG and KK conceived and designed the study. KK analysed performed all the bioinformatics and expression analysis, HS contributed in all bioinformatics work, KD and RJ provides the samples for real-time analysis, AD, KT and PK contributed in qRT-PCR and data analysis, KK, AD, HS, KT, KD, RJ, AM, NK and KG compiled and interpreted the data and wrote the manuscript. All authors have read and approved the final manuscript

8. Reference

- [1] FAOSTAT. <http://www.fao.org> Accessed 2 Feb 2019
- [2] G. Bernier, C. Périlleux, A physiological overview of the genetics of flowering time control, Plant Biotechnol. J. (2005). <https://doi.org/10.1111/j.1467-7652.2004.00114.x>.
- [3] L.H. Cho, J. Yoon, G. An, The control of flowering time by environmental factors, Plant J. (2017). <https://doi.org/10.1111/tpj.13461>.
- [4] W. Deng, H. Ying, C.A. Helliwell, J.M. Taylor, W.J. Peacock, E.S. Dennis, FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*, Proc. Natl. Acad. Sci. (2011). <https://doi.org/10.1073/pnas.1103175108>.

- [5] F. Messenguy, E. Dubois, Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development., *Gene*. 316 (2003) 1–21. <http://www.ncbi.nlm.nih.gov/pubmed/14563547> (accessed July 4, 2019).
- [6] G. Theissen, J.T. Kim, H. Saedler, 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 (1996) 484–516. <http://www.ncbi.nlm.nih.gov/pubmed/8875863> (accessed July 4, 2019).
- [7] Z. Schwarz-Sommer, P. Huijser, W. Nacken, H. Saedler, H. Sommer, Genetic control of flower development by homeotic genes in *Antirrhinum majus*, *Science* (80). (1990). <https://doi.org/10.1126/science.250.4983.931>.
- [8] W. Tröbner, L. Ramirez, P. Motte, I. Hue, P. Huijser, W.E. Lönig, H. Saedler, H. Sommer, Z. Schwarz-Sommer, GLOBOSA: a homeotic gene which interacts with DEFICIENS in the control of *Antirrhinum* floral organogenesis., *EMBO J.* 11 (1992) 4693–704. <http://www.ncbi.nlm.nih.gov/pubmed/1361166> (accessed November 5, 2019).
- [9] G. Theissen, R. Melzer, F. Rümpler, MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution, *Development*. (2016). <https://doi.org/10.1242/dev.134080>.
- [10] A. Becker, G. Theissen, The major clades of MADS-box genes and their role in the development and evolution of flowering plants., *Mol. Phylogenet. Evol.* 29 (2003) 464–89. <http://www.ncbi.nlm.nih.gov/pubmed/14615187> (accessed July 4, 2019).
- [11] K. Henschel, R. Kofuji, M. Hasebe, H. Saedler, T. Münster, G. Theissen, Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*, *Mol. Biol. Evol.* (2002). <https://doi.org/10.1093/oxfordjournals.molbev.a004137>.
- [12] J. Nam, J. Kim, S. Lee, G. An, H. Ma, M. Nei, Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms., *Proc. Natl. Acad. Sci. U. S. A.* 101 (2004) 1910–5. <https://doi.org/10.1073/pnas.0308430100>.
- [13] L. Parenicová, S. de Folter, M. Kieffer, D.S. Horner, C. Favalli, J. Busscher, H.E. Cook, R.M. Ingram, M.M. Kater, B. Davies, G.C. Angenent, L. Colombo, Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world., *Plant Cell*. (2003).
- [14] G. Theissen, A. Becker, A. Di Rosa, A. Kanno, J.T. Kim, T. Münster, K.U. Winter, H. Saedler, A short history of MADS-box genes in plants., *Plant Mol. Biol.* 42 (2000) 115–49. <http://www.ncbi.nlm.nih.gov/pubmed/10688133> (accessed July 8, 2019).
- [15] S. Ortuño-Miquel, E. Rodríguez-Cazorla, E.A. Zavala-Gonzalez, A. Martínez-Laborda, A. Vera, *Arabidopsis* HUA ENHANCER 4 delays flowering by upregulating the MADS-box repressor genes FLC and MAF4, *Sci. Rep.* (2019). <https://doi.org/10.1038/s41598-018-38327-3>.

- 1 [16] G. Kumar, P. Arya, K. Gupta, V. Randhawa, V. Acharya, A.K. Singh, Comparative phylogenetic
2 analysis and transcriptional profiling of MADS-box gene family identified DAM and FLC-like
3 genes in apple (*Malus domestica*), Sci. Rep. 6 (2016) 1–13. <https://doi.org/10.1038/srep20695>.
- 4 [17] Q. Dong, F. Wang, J. Kong, Q. Xu, T. Li, L. Chen, H. Chen, H. Jiang, C. Li, B. Cheng, Functional
5 analysis of ZmMADS1a reveals its role in regulating starch biosynthesis in maize endosperm, Sci.
6 Rep. (2019). <https://doi.org/10.1038/s41598-019-39612-5>.
- 7 [18] Q. Xie, Z. Hu, Z. Zhu, T. Dong, Z. Zhao, B. Cui, G. Chen, Overexpression of a novel MADS-box
8 gene SIFYFL delays senescence, fruit ripening and abscission in tomato., Sci. Rep. 4 (2014) 4367.
9 <https://doi.org/10.1038/srep04367>.
- 10 [19] R. Arora, P. Agarwal, S. Ray, A.K. Singh, V.P. Singh, A.K. Tyagi, S. Kapoor, MADS-box gene
11 family in rice: Genome-wide identification, organization and expression profiling during
12 reproductive development and stress, BMC Genomics. (2007). [https://doi.org/10.1186/1471-2164-8-](https://doi.org/10.1186/1471-2164-8-242)
13 [242](https://doi.org/10.1186/1471-2164-8-242).
- 14 [20] X. Qiao, Q. Li, H. Yin, K. Qi, L. Li, R. Wang, S. Zhang, A.H. Paterson, Gene duplication and
15 evolution in recurring polyploidization-diploidization cycles in plants 06 Biological Sciences 0604
16 Genetics, Genome Biol. 20 (2019). <https://doi.org/10.1186/s13059-019-1650-2>.
- 17 [21] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C.J. Lanczycki, S. Lu, F. Chitsaz, M.K. Derbyshire, R.C.
18 Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z.
19 Wang, R.A. Yamashita, D. Zhang, C. Zheng, L.Y. Geer, S.H. Bryant, CDD/SPARCLE: Functional
20 classification of proteins via subfamily domain architectures, Nucleic Acids Res. (2017).
21 <https://doi.org/10.1093/nar/gkw1129>.
- 22 [22] R.E. Voorrips, MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs, J.
23 Hered. 93 (2002) 77–78. <https://doi.org/10.1093/jhered/93.1.77>.
- 24 [23] T. H. Lee, H. Tang, X. Wang, A. H. Paterson, PGDD: a database of gene and genome duplication in
25 plant. Nucleic Acid Research, Doi: 10.1039/nar/gks1104 (2012) doi:
26 <https://doi.org/10.1093/nar/gks1104>
- 27 [24] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics
28 analysis across computing platforms, Mol. Biol. Evol. (2018).
29 <https://doi.org/10.1093/molbev/msy096>.
- 30 [25] B. Hu, J. Jin, A.Y. Guo, H. Zhang, J. Luo, G. Gao, GSDS 2.0: An upgraded gene feature
31 visualization server, Bioinformatics. (2015). <https://doi.org/10.1093/bioinformatics/btu817>.
- 32 [26] A. Conesa, S. Götz, Blast2GO: A comprehensive suite for functional analysis in plant genomics, Int.
33 J. Plant Genomics. 2008 (2008). <https://doi.org/10.1155/2008/619832>.
- 34 [27] L.T. Pazhamala, S. Purohit, R.K. Saxena, V. Garg, L. Krishnamurthy, J. Verdier, R.K. Varshney,
35 Gene expression atlas of pigeonpea and its application to gain insights into genes associated with
36 pollen fertility implicated in seed formation, J. Exp. Bot. (2017). <https://doi.org/10.1093/jxb/erx010>.

- [28] T.D. Schmittgen, K.J. Livak, Analyzing real-time PCR data by the comparative CT method, *Nat. Protoc.* 3 (2008) 1101–1108. <https://doi.org/10.1038/nprot.2008.73>.
- [29] K. Tamura, F.U. Battistuzzi, P. Billing-Ross, O. Murillo, A. Filipinski, S. Kumar, Estimating divergence times in large molecular phylogenies, *Proc. Natl. Acad. Sci. U. S. A.* 109 (2012) 19333–19338. <https://doi.org/10.1073/pnas.1213199109>.
- [30] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, M. Kuhn, P. Bork, L.J. Jensen, C. Von Mering, STRING v10: Protein-protein interaction networks, integrated over the tree of life, *Nucleic Acids Res.* (2015). <https://doi.org/10.1093/nar/gku1003>.
- [31] C. Dorca-Fornell, V. Gregis, V. Grandi, G. Coupland, L. Colombo, M.M. Kater, The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems, *Plant J.* 67 (2011) 1006–1017. <https://doi.org/10.1111/j.1365-313X.2011.04653.x>.
- [32] H. Zhang, W. Teng, J. Liang, X. Liu, H. Zhang, Z. Zhang, X. Zheng, MADS1, a novel MADS-box protein, is involved in the response of *Nicotiana benthamiana* to bacterial harpinXoofile:///Users/ziegler/Documents/Mantelin 2011-Plant J. aphid Rce.pdf, *J. Exp. Bot.* 67 (2016) 131–141. <https://doi.org/10.1093/jxb/erv448>.
- [33] G.B. Seymour, C.D. Ryder, V. Cevik, J.P. Hammond, A. Popovich, G.J. King, J. Vrebalov, J.J. Giovannoni, K. Manning, A SEPALLATA gene is involved in the development and ripening of strawberry (*Fragaria×ananassa* Duch.) fruit, a non-climacteric tissue, *J. Exp. Bot.* (2011). <https://doi.org/10.1093/jxb/erq360>.
- [34] C. Ferrándiz, Q. Gu, R. Martienssen, M.F. Yanofsky, Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER., *Development.* (2000).
- [35] D. Liu, D. Wang, Z. Qin, D. Zhang, L. Yin, L. Wu, J. Colasanti, A. Li, L. Mao, The SEPALLATA MADS-box protein SLMBP21 forms protein complexes with JOINTLESS and MACROCALYX as a transcription activator for development of the tomato flower abscission zone. *Plant J.* 77(2):284-96. (2014) doi: <https://doi.org/10.1111/tpj.12387>. Epub 2013 Dec 20.
- [36] X. Liu, Z. Sun, W. Dong, Z. Wang, L. Zhang, Expansion and functional divergence of the SHORT VEGETATIVE phase (SVP) genes in eudicots, *Genome Biol. Evol.* (2018). <https://doi.org/10.1093/gbe/evy235>.
- [37] Y. Wu, Y. Ke, J. Wen, P. Guo, F. Ran, M. Wang, M. Liu, P. Li, J. Li, H. Du, Evolution and expression analyses of the MADS-box gene family in *Brassica napus*, *PLoS One.* (2018). <https://doi.org/10.1371/journal.pone.0200762>.
- [38] Y. Gan, S. Filleur, A. Rahman, S. Gotensparre, B.G. Forde, Nutritional regulation of ANR1 and other root-expressed MADS-box genes in *Arabidopsis thaliana*., *Planta.* 222 (2005) 730–42. <https://doi.org/10.1007/s00425-005-0020-3>.

- [39] C. Yu, S. Su, Y. Xu, Y. Zhao, A. Yan, L. Huang, I. Ali, Y. Gan, The effects of fluctuations in the nutrient supply on the expression of five members of the AGL17 clade of MADS-box genes in rice., *PLoS One*. 9 (2014) e105597. <https://doi.org/10.1371/journal.pone.0105597>.
- [40] C. Kohler, L. Hennig, C. Spillane, S. Pien, W. Gruissen, U. Grossniklaus, The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. *Genes Dev*. 15;17(12):1540-53 (2003). <https://doi.org/10.1101/gad.257403>
- [41] M.A. Mandel, M.F. Yanofsky, The Arabidopsis AGL9 MADS box gene is expressed in young flower primordia, *Sex. Plant Reprod*. 11 (1998) 22–28. <https://doi.org/10.1007/s004970050116>.
- [42] D. Benovoy, G. Drouin, Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome. *Journal of Molecular Evolution*, 62(5, J. Mol. Evol. (2006). <https://doi.org/10.1007/s00239-005-0045-z>.
- [43] S. Ferrario, J. Busscher, J. Franken, T. Gerats, M. Vandenbussche, G.C. Angenent, R.G.H. Immink, Ectopic expression of the petunia MADS box gene UNSHAVEN accelerates flowering and confers leaf-like characteristics to floral organs in a dominant-negative manner, *Plant Cell*. 16 (2004) 1490–1505. <https://doi.org/10.1105/tpc.019679>.
- [44] U. Johanson, J. West, C. Lister, S. Michaels, R. M. Amsino, C. Dean, Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time." *Science* 290:344-347(2000). DOI: <https://doi.org/10.1126/science.290.5490.344>
- [45] D. Jiang, X. Gu, Y. He, Establishment of the winter-annual growth habit via FRIGIDA-mediated histone methylation at FLOWERING LOCUS C in Arabidopsis., *Plant Cell*. 21 (2009) 1733–46. <https://doi.org/10.1105/tpc.109.067967>.
- [46] C. Li, H. Gu, W. Jiang, C. Zou, D. Wei, Z. Wang, Q. Tang, Protein interactions of SOC1 with SVP are regulated by a few crucial amino acids in flowering pathways of Brassica juncea, *Acta Physiol. Plant*. (2019). <https://doi.org/10.1007/s11738-019-2838-5>.
- [47] A. Mouradov, F. Cremer, G. Coupland, Control of flowering time: interacting pathways as a basis for diversity., *Plant Cell*. (2002). <https://doi.org/10.1105/tpc.001362>.
- [48] J. Lee, I. Lee, Regulation and function of SOC1, a flowering pathway integrator, *J. Exp. Bot*. 61 (2010) 2247–2254. <https://doi.org/10.1093/jxb/erq098>.
- [49] A. Samach, H. Onouchi, S.E. Gold, G.S. Ditta, Z. Schwarz-Sommer, M.F. Yanofsky, G. Coupland, Distinct roles of constans target genes in reproductive development of Arabidopsis, *Science* (80). 288 (2000) 1613–1616. <https://doi.org/10.1126/science.288.5471.1613>.
- [50] S. Kim, K. Choi, C. Park, H.J. Hwang, I. Lee, SUPPRESSOR of FRIGIDA4, encoding a C2H2-type zinc finger protein, represses flowering by transcriptional activation of Arabidopsis FLOWERING LOCUS C, *Plant Cell*. 18 (2006) 2985–2998. <https://doi.org/10.1105/tpc.106.045179>.

- [51] H.L. Jeong, J.Y. Seong, H.P. Soo, I. Hwang, S.L. Jong, H.A. Ji, Role of SVP in the control of flowering time by ambient temperature in Arabidopsis, *Genes Dev.* (2007). <https://doi.org/10.1101/gad.1518407>.
- [52] S. Pelaz, G.S. Ditta, E. Baumann, E. Wisman, M.F. Yanofsky, B and C floral organ identity functions require SEPALLATA MADS-box genes., *Nature.* 405 (2000) 200–3. <https://doi.org/10.1038/35012103>.
- [53] M.F. Yanofsky, H. Ma, J.L. Bowman, G.N. Drews, K.A. Feldmann, E.M. Meyerowitz, The protein encoded by the Arabidopsis homeotic gene *agamous* resembles transcription factors, *Nature.* (1990). <https://doi.org/10.1038/346035a0>.
- [54] P. Smýkal, C.J. Coyne, M.J. Ambrose, N. Maxted, H. Schaefer, M.W. Blair, J. Berger, S.L. Greene, M.N. Nelson, N. Besharat, T. Vymyslický, C. Toker, R.K. Saxena, M. Roorkiwal, M.K. Pandey, J. Hu, Y.H. Li, L.X. Wang, Y. Guo, L.J. Qiu, R.J. Redden, R.K. Varshney, Legume Crops Phylogeny and Genetic Diversity for Science and Breeding, CRC. *Crit. Rev. Plant Sci.* 34 (2015) 43–104. <https://doi.org/10.1080/07352689.2014.897904>.

Figure 1: Basic structure of type I and II MADS-box genes. In type II MADS-box genes any of the extra domains i.e. I, K and C is found, K box is the most frequent of which.

Figure 2: Conserved domain structure reveals the presence of only MADS-box domains in type I MADS-box protein

Figure 3: Conserved domain structure reveals the presence of K box domain besides MADS-box domains in type II MADS-box protein in *C. cajan*

Figure 4: Gene structure of type I reveals that presence of intron is rare phenomenon in them

Figure 5: Gene structure of type I reveals that presence of intron is common phenomenon in them

Figure 6: Gene based phylogeny of all MADS-box genes identified in *C. cajan*

Figure 7: (A) STRING analysis showing interaction network of SOC1 (1.5, 1.6 and 7.1) genes; (B) SVP gene and (C) FLC gene

Figure 8: Expression pattern of some important floral regulator in different tissues as revealed by quantitative rt-PCR

Figure 9: (A) Time tree showing molecular divergence (in million years ago) of FLC genes in *C. cajan*, *G. max* and *Arabidopsis* and (B) SOC1 genes in *C. cajan*, *G. max*, *C. arietinum*, *Arachis hypogaea* and *Arabidopsis*

1

2 **Figure S1:** Schematic representation of MADS distribution along all chromosomes

Figure S2: Gene based phylogeny of all MADS-box genes in cultivated as well wild genotypes; blue, black, yellow and red squares represents *C. scarabaeoides*, *C. cajanifolius*, *C. platycarpus* and *C. cajan* representatives respectively

```

CcMADS1.6      -----MNQRISPYLT--LDYQQVRQGDVK-----
CcMADS7.1      MVRGKTQMRRIENATSRQVTFSKRRNGLLKKAFELSVLCDAEVALIIFSPRGKLYEFAST
CcMADS1.5      MVRGKTQMKRIENATSRQVTFSKRRNGLLKKAFELSVLCDAEVALIIFSPRGKLYEFA--
               .**      :      : : : * : *

CcMADS1.6      -----FLGEGLGSCS
CcMADS7.1      SMQETIERYYRRHNSAQTVNRSDEQNMQHLKQETANLMKKIELLEASKRKLLGEGLGSCS
CcMADS1.5      -----

CcMADS1.6      IEELHRIEQQLQRSISNVRARKTQVFMEQIEELKEKEKALLEEKARLSEKHGTDQP----
CcMADS7.1      LEELQQIEQQLENSVNNVRARKNQVYKEQIEQLKVKEKALYAENARLCEQLGIQPQAEAE
CcMADS1.5      -----NSRYC-----
               : : * .

CcMADS1.6      PTTKDQTENQPQNYAESSLSDEVETELFIGLPHSRTRGVISCWKRNGPFRKQNGPFCLKT
CcMADS7.1      AATRDPKENQP--YAESSPSSEVETELFIGLPRSS-----
CcMADS1.5      -----SP--FSS-----FSNFPFSCS-----
               . *      : : .      *      : : * *

CcMADS1.6      QSDISESETSLERTVPKAKRVVSLSEMSL
CcMADS7.1      -----
CcMADS1.5      -----

```

Figure S3: Multiple sequence alignment for all 3 paralogs of SOC1 gene using clustal omega

Figure S4: Combined phylogenetic tree developed by MADS-box isolated from selected grain legumes

Figure S5: Heat map showing expression of MADS-box gene in different vegetative and reproductive tissues

1 **Table S1:** Details of primers used for qRT-PCR

Gene ID	Primer Name	Forward Primer	Length	Tm	GC Percent	Reverse Primer	Length	Tm	GC Percent	Amplicon
cajca.C.cajan_20579	CeMADS1.5	GCGTATAGAGAACGCCACAA	20	62	50	CAACTTCAGCATCACACAGAAC	22	62	45.5	104
cajca.C.cajan_20582	CcMADS1.6	GATCCACAGCCAACAACAAAG	21	62	47.6	CGGTCTCCACATCTGAACCTTAG	22	62	50	85
cajca.C.cajan_18407	CeMADS7.1	CTGTGTGAGCAGTTAGGTATCC	22	62	50	TGGACTGCTTTCTGCATAGG	20	62	50	93
cajca.C.cajan_12250	CcMADS6.2	GCGAGGGAGAAGATTCAGATTA	22	61	45.5	GCATCACACAGAACCGAAAG	20	61	50	119
cajca.C.cajan_24259	FLC	GGGAAGAAGAAGGTGGAGATAAA	23	62	43.5	TCGCATAGAACGGACAACCTC	20	62	50	116
XM_020358256.2	FRI	CCTGTTGGGCAGAGTTATGT	20	62	50	GATGCTCTGGAACACCTGAA	20	62	50	100

2

3 **Table S2:** List of 102 MADS-box genes identified in pigeonpea and their characteristics

S. no.	Name	Chromosome and scaffold	LIS identifier	Protein identifier	Start position	End position	Protein length	Function	Theoretical pI and MW
1	CeMADS1.1	Chr. 1	cajca.C.cajan_19616	XP_020211538.1	4408594	4409320	271	Agamous-like MADS-box protein AGL80	9.76 / 27606.94
2	CeMADS1.2	Chr. 1	cajca.C.cajan_19733	XP_020211435.1	5734709	5735434	191	Agamous-like MADS-box protein AGL80	9.41 / 22469.05
3	CeMADS1.3	Chr. 1	cajca.C.cajan_20239	XP_020220789.1	10888762	10892695	358	Agamous-like MADS-box protein AGL30	8.53 / 40441.60
4	CeMADS1.4	Chr. 1	cajca.C.cajan_20351	XP_020220122.1	12295393	12299915	214	Floral homeotic protein pmads 2-like	9.04 / 25196.88
5	CeMADS1.5	Chr. 1	cajca.C.cajan_20579	XP_020220768.1	14829984	14830218	77	MADS-box protein SOC1	10.04 / 8794.21
6	CeMADS1.6	Chr. 1	cajca.C.cajan_20582	KYP76934	14861427	14866911	177	MADS-box protein SOC1	6.85 / 20253.80
7	CeMADS1.7	Chr. 1	cajca.C.cajan_20583	XP_020214835.1	14874502	14880529	243	MADS-box transcription factor 17	8.96 / 28119.12
8	CeMADS2.1	Chr. 2	cajca.C.cajan_04971	XP_020228054.1	5283756	5288408	331	MADS-box transcription factor 1	5.42 / 37729.29
9	CeMADS2.2	Chr. 2	cajca.C.cajan_04979	XP_020228064.1	5318852	5324582	331	MADS-box transcription factor 2	5.42 / 37810.37
10	CeMADS2.3	Chr. 2	cajca.C.cajan_04992	KYP72524	5436633	5436786	50	Agamous-like MADS-box protein AGL11	11.53 / 5868.04
11	CeMADS2.4	Chr. 2	cajca.C.cajan_05842	XP_020239322.1	14248177	14252726	245	Agamous-like MADS-box protein AGL5	8.98 / 28043.57
12	CeMADS2.5	Chr. 2	cajca.C.cajan_06129	XP_020207051.1	30597529	30597781	255	Agamous-like MADS-box protein AGL15	9.85 / 29694.78
13	CeMADS2.6	Chr. 2	cajca.C.cajan_07398	XP_020232325.1	17252506	17255778	83	MADS-box protein AGL42-like	9.76 / 9547.21
14	CeMADS2.7	Chr. 2	cajca.C.cajan_07422	XP_020232396.1	30831579	30831831	83	MADS-box protein AGL42-like	9.76 / 9547.21
15	CeMADS2.8	Chr. 2	cajca.C.cajan_07427	XP_020237639.1	30914480	30934112	227	MADS-box transcription factor 17 isoform X2	9.12 / 26437.46
16	CeMADS2.9	Chr. 2	cajca.C.cajan_07729	XP_020233147.1	34375886	34376417	176	Agamous-like MADS-box protein AGL61	9.30 / 20490.00
17	CeMADS2.10	Chr. 2	cajca.C.cajan_07730	XP_020233156.1	34380504	34380831	108	Agamous-like MADS-box protein AGL61	6.74 / 12422.10
18	CeMADS2.11	Chr. 2	cajca.C.cajan_07731	XP_020233169.1	34384059	34384752	230	Agamous-like MADS-box protein AGL62	8.78 / 25961.68
19	CeMADS2.12	Chr. 2	cajca.C.cajan_07734	XP_020209627.1	34414011	34414701	239	Agamous-like MADS-box protein AGL62	9.36 / 25653.20
20	CeMADS2.13	Chr. 2	cajca.C.cajan_07735	XP_020233190.1	34428973	34429570	198	Agamous-like MADS-box protein AGL62	9.19 / 22283.75
21	CeMADS3.1	Chr. 3	cajca.C.cajan_08738	XP_020212024.1	7336087	7340320	179	Agamous-like MADS-box protein AGL8	6.23 / 20883.58
22	CeMADS3.2	Chr. 3	cajca.C.cajan_08739	RDX62358	7340131	7340320	62	Floral homeotic protein APETALA 1	10.21 / 7144.38
23	CeMADS3.3	Chr. 3	cajca.C.cajan_08743	XP_020213192.1	7377547	7389463	62	Agamous-like MADS-box protein AGL3	10.11 / 7129.43

24	CcMADS3.4	Chr. 3	cajca.C.cajan_08946	XP_020213894.1	9335735	9338037	143	Floral homeotic protein GLOBOSA	9.14 / 16299.71
25	CcMADS3.5	Chr. 3	cajca.C.cajan_09006	XP_020213859.1	10037816	10040216	76	MADS-box protein JOINTLESS	10.15 / 8812.26
26	CcMADS3.6	Chr. 3	cajca.C.cajan_09005	KYP70056	10047221	10047452	172	MADS-box protein JOINTLESS	4.66 / 19259.12
27	CcMADS3.7	Chr. 3	cajca.C.cajan_10816	KYP71854	27804378	27804633	84	Developmental protein SEPALLATA 3	9.78 / 9972.57
28	CcMADS4.1	Chr. 4	cajca.C.cajan_20893	XP_020215268.1	317855	324180	201	Agamous-like MADS-box protein AGL12	9.16 / 23002.90
29	CcMADS6.1	Chr. 6	cajca.C.cajan_11329	XP_020217667.1	3432998	3435782	223	MADS-box protein FBP24	6.34 / 26125.94
30	CcMADS6.2	Chr. 6	cajca.C.cajan_12250	KYP66335	12873463	12873734	86	MADS-box protein SVP	9.91 / 9851.49
31	CcMADS6.3	Chr. 6	cajca.C.cajan_12251	XP_020218071.1	12877201	12880879	168	MADS-box protein JOINTLESS	5.20 / 18912.15
32	CcMADS6.4	Chr. 6	cajca.C.cajan_12291	XP_020216806.1	13381678	13390090	224	MADS-box transcription factor 23-like	8.78 / 26006.74
33	CcMADS6.5	Chr. 6	cajca.C.cajan_12821	XP_020216957.1	18826605	18828551	236	Floral homeotic protein DEFICIENS	9.39 / 27439.75
34	CcMADS6.6	Chr. 6	cajca.C.cajan_12868	XP_020216886.1	19382581	19383589	335	floral homeotic protein AGAMOUS-like	9.07 / 38294.71
35	CcMADS6.7	Chr. 6	cajca.C.cajan_13317	XP_020216947.1	23629895	23630567	223	Agamous-like MADS-box protein AGL61	9.22 / 25132.90
36	CcMADS7.1	Chr. 7	cajca.C.cajan_18407	XP_020219513.1	11874840	11886028	213	MADS-box protein SOC1-like isoform X1	8.92 / 24402.66
37	CcMADS8.1	Chr. 8	cajca.C.cajan_15708	XP_020220161.1	3529423	3529936	170	MADS-box protein JOINTLESS	9.58 / 19429.58
38	CcMADS8.2	Chr. 8	cajca.C.cajan_15709	XP_020222045.1	3539127	3539661	177	Agamous-like MADS-box protein AGL29	9.67 / 19974.99
39	CcMADS8.3	Chr. 8	cajca.C.cajan_15710	XP_020221917.1	3560511	3561048	178	Agamous-like MADS-box protein AGL29	8.26 / 20314.17
40	CcMADS8.4	Chr. 8	cajca.C.cajan_15712	XP_020221887.1	3569130	3569673	180	Agamous-like MADS-box protein AGL29	9.05 / 20094.13
41	CcMADS8.5	Chr. 8	ajca.C.cajan_17116	XP_020220153.1	18752885	18753161	91	agamous-like MADS-box protein AGL29	10.06 / 10505.55
42	CcMADS8.6	Chr. 8	cajca.C.cajan_17127	KYP63093	18789986	18792276	149	MADS-box protein JOINTLESS	6.21 / 17237.59
43	CcMADS8.7	Chr. 8	cajca.C.cajan_17128	XP_020221760.1	18796678	18796888	60	MADS-box protein JOINTLESS	10.03 / 8005.57
44	CcMADS8.8	Chr. 8	cajca.C.cajan_17149	KYP63093	18999921	19000206	94	MADS-box protein JOINTLESS	10.16 / 10907.03
45	CcMADS9.1	Chr. 9	cajca.C.cajan_22935	XP_020223003.1	8970848	8985849	236	MADS-box protein JOINTLESS	9.38 / 27463.75
46	CcMADS9.2	Chr. 9	cajca.C.cajan_22967	KYP61215	9274654	9274840	60	MADS-box transcription factor 57	10.27 / 6863.92
47	CcMADS9.3	Chr. 9	cajca.C.cajan_22975	XP_020222130.1	9362269	9363277	335	floral homeotic protein AGAMOUS-like	9.07 / 38294.71
48	CcMADS11.1	Chr. 11	cajca.C.cajan_00220	XP_020227817.1	2053324	2054460	378	MADS-box transcription factor PHERES 2-like	5.73 / 43094.25
49	CcMADS11.2	Chr. 11	cajca.C.cajan_00703	XP_020225031.1	7289209	7289395	61	MADS-box transcription factor 57	10.15 / 6982.18
50	CcMADS11.3	Chr. 11	cajca.C.cajan_00709	KYP54538	7316123	7317637	139	MADS-box transcription factor 27	6.71 / 16021.38
51	CcMADS11.4	Chr. 11	cajca.C.cajan_01433	XP_020226323.1	15464035	15469653	240	Agamous-like MADS-box protein AGL15	8.82 / 27285.20
52	CcMADS11.5	Chr. 11	cajca.C.cajan_01724	XP_020225648.1	18838734	18843202	360	MADS-box protein GGM13	8.54 / 41947.78
53	CcMADS11.6	Chr. 11	cajca.C.cajan_02642	XP_020226034.1	29023845	29034291	229	Agamous-like MADS-box protein AGL14	9.42 / 26983.91
54	CcMADS11.7	Chr. 11	cajca.C.cajan_02899	XP_020227937.1	32032982	32033482	89	MADS-box protein CMB1-like	9.91 / 9971.43
55	CcMADS11.8	Chr. 11	cajca.C.cajan_02900	XP_020227938.1	32041349	32042729	152	Agamous-like MADS-box protein AGL8 isogeny	9.65 / 17933.55
56	CcMADS11.9	Chr. 11	cajca.C.cajan_03143	XP_020225349.1	34675538	34676708	372	agamous-like MADS-box protein AGL36	5.29 / 40932.43
57	CcMADSsca1.1	NW_017984067.1	cajca.C.cajan_24257	XP_020229782.1	193155	199428	241	Agamous-like MADS-box protein AGL9-like protein	8.82 / 27507.31
58	CcMADSsca1.2	NW_017984067.1	cajca.C.cajan_24258	XP_020229791.1	205552	215326	241	Agamous-like MADS-box protein AGL9-like protein	8.22 / 27599.34
59	CcMADSsca1.3	NW_017984067.1	cajca.C.cajan_24259	XP_020229765.1	217113	217326	70	MADS-box protein FLOWERING LOCUS C	10.24 / 8081.56
60	CcMADSsca2.1	NW_017984069.1	cajca.C.cajan_25279	XP_020229986.1	641978	646607	149	floral homeotic protein apetala 1-like	9.63 / 17594.23
61	CcMADSsca3.1	NW_017984121.1	cajca.C.cajan_26683	XP_020232939.1	429390	429969	192	Agamous-like MADS-box protein AGL62	8.85 / 22409.57
62	CcMADSsca4.1	NW_017984161.1	cajca.C.cajan_34567	XP_020234510.1	362953	363142	62	Agamous-like MADS-box protein AGL8	10.21 / 7136.36
63	CcMADSsca4.2	NW_017984161.1	cajca.C.cajan_34569	XP_020234496.1	408084	412378	197	truncated transcription factor CAULIFLOWER A-like	9.16 / 22942.57
64	CcMADSsca5.1	NW_017984170.1	cajca.C.cajan_31945	XP_020234727.1	316808	325078	240	agamous-like MADS-box protein AGL1 isoform X1	9.51 / 27664.25
65	CcMADSsca6.1	NW_017984224.1	cajca.C.cajan_35672	XP_020236340.1	130427	130910	160	agamous-like MADS-box protein AGL62	6.37 / 18189.48
66	CcMADSsca6.2	NW_017984224.1	cajca.C.cajan_35673	XP_020236341.1	158190	158856	221	agamous-like MADS-box protein AGL62	9.51 / 25123.49
67	CcMADSsca7.1	NW_017984260.1	cajca.C.cajan_31406	XP_020237330.1	248834	257123	226	floral homeotic protein DEFICIENS	9.33 / 26248.97
68	CcMADSsca8.1	NW_017984271.1	cajca.C.cajan_34534	XP_020237602.1	78195	79784	192	agamous-like MADS-box protein AGL62	9.53 / 21473.85

69	CcMADSsca9.1	NW_017984272.1	cajca.C.cajan_32024	XP_020237635.1	80623	81496	238	MADS-box transcription factor PHERES 2-like	5.53 / 27092.80
70	CcMADSsca9.2	NW_017984272.1	cajca.C.cajan_32025	XP_020237625.1	99462	99984	173	agamous-like MADS-box protein AGL80	9.32 / 19795.04
71	CcMADSsca9.3	NW_017984272.1	cajca.C.cajan_32026	XP_020237626.1	102000	103017	338	agamous-like MADS-box protein AGL80	6.47 / 39490.02
72	CcMADSsca9.4	NW_017984272.1	cajca.C.cajan_32028	XP_020237628.1	116234	116978	247	agamous-like MADS-box protein AGL80	8.57 / 28494.93
73	CcMADSsca9.5	NW_017984272.1	cajca.C.cajan_32030	XP_020237629.1	127468	128173	234	agamous-like MADS-box protein AGL80	6.64 / 27323.36
74	CcMADSsca9.6	NW_017984272.1	cajca.C.cajan_32032	XP_020237617.1	152042	152822	259	agamous-like MADS-box protein AGL80	6.03 / 29716.18
75	CcMADSsca9.7	NW_017984272.1	cajca.C.cajan_32033	XP_020237614.1	162135	162849	237	agamous-like MADS-box protein AGL80	9.06 / 26974.22
76	CcMADSsca10.1	NW_017984431.1	cajca.C.cajan_36667	KYP41312	190514	192747	166	Agamous-like MADS-box protein AGL11	9.01 / 19233.90
77	CcMADSsca10.2	NW_017984431.1	cajca.C.cajan_36668	XP_020240315.1	199379	199565	61	Floral homeotic protein AGAMOUS	9.93 / 7082.24
78	CcMADSsca11.1	NW_017984433.1	cajca.C.cajan_38112	XP_020240365.1	180878	181543	129	agamous-like MADS-box protein AGL16 isoform X7	5.73 / 14740.83
79	CcMADSsca11.2	NW_017984433.1	cajca.C.cajan_38114	XP_020240364.1	204325	204739	137	Agamous-like MADS-box protein AGL21	8.93 / 15650.26
80	CcMADSsca12.1	NW_017984466.1	cajca.C.cajan_34456	KYP40749	201471	202239	255	MADS-box transcription factor PHERES 2	5.63 / 29232.74
81	CcMADSsca13.1	NW_017984487.1	cajca.C.cajan_40507	XP_020202289.1	56662	56896	77	Agamous-like MADS-box protein AGL9 isogeny	9.93 / 8953.49
82	CcMADSsca13.2	NW_017984487.1	cajca.C.cajan_40508	KYP40409	60053	62486	185	Developmental protein SEPALLATA 2	5.80 / 21416.06
83	CcMADSsca13.3	NW_017984487.1	cajca.C.cajan_40512	XP_020202293.1	154039	165342	168	Agamous-like MADS-box protein AGL8	9.73 / 19499.39
84	CcMADSsca14.1	NW_017984494.1	cajca.C.cajan_38133	XP_020202375.1	81261	81969	235	agamous-like MADS-box protein AGL62	9.71 / 26604.50
85	CcMADSsca15.1	NW_017984515.1	cajca.C.cajan_36447	XP_020202584.1	16162	16612	149	agamous-like MADS-box protein AGL80	9.62 / 17417.44
86	CcMADSsca15.2	NW_017984515.1	cajca.C.cajan_36448	XP_020202577.1	29548	30031	160	agamous-like MADS-box protein AGL80	9.27 / 18338.31
87	CcMADSsca16.1	NW_017984544.1	cajca.C.cajan_37010	RDY10356	49221	49624	52	agamous-like MADS-box protein AGL80	10.01 / 6134.38
88	CcMADSsca17.1	NW_017984584.1	cajca.C.cajan_39656	XP_020203367.1	93238	93556	80	Floral homeotic protein AGAMOUS	9.95 / 9128.60
89	CcMADSsca17.2	NW_017984584.1	cajca.C.cajan_39657	KYP39131	98036	101503	181	Floral homeotic protein AGAMOUS	7.82 / 20686.12
90	CcMADSsca18.1	NW_017984687.1	cajca.C.cajan_36782	KYP38093	102670	105534	167	Floral homeotic protein AGAMOUS	7.91 / 18676.05
91	CcMADSsca18.2	NW_017984687.1	cajca.C.cajan_36783	XP_020204254.1	109510	109765	84	Floral homeotic protein AGAMOUS	10.17 / 9729.30
92	CcMADSsca19.1	NW_017984726.1	cajca.C.cajan_40661	KYP35461	75907	76582	173	Agamous-like MADS-box protein AGL80	9.49 / 20027.32
93	CcMADSsca20.1	NW_017984783.1	cajca.C.cajan_38665	XP_020205008.1	74032	80497	243	developmental protein SEPALLATA 1-like isoform X2	9.20 / 27790.55
94	CcMADSsca20.2	NW_017984783.1	cajca.C.cajan_38667	XP_020205018.1	92347	101111	172	Agamous-like MADS-box protein AGL8	9.54 / 20104.13
95	CcMADSsca21.1	NW_017984803.1	cajca.C.cajan_43547	XP_020205191.1	46632	55709	192	MADS-box transcription factor 27	7.02 / 22548.68
96	CcMADSsca21.2	NW_017984803.1	cajca.C.cajan_43548	KYP36959	63709	66387	77	MADS-box transcription factor 57	10.02 / 8878.23
97	CcMADSsca22.1	NW_017985009.1	cajca.C.cajan_42409	XP_020206215.1	3042	4230	395	agamous-like MADS-box protein AGL103	5.71 / 44231.82
98	CcMADSsca23.1	NW_017985028.1	cajca.C.cajan_46899	KYP35461	49061	49808	197	Agamous-like MADS-box protein AGL80	9.68 / 22606.21
99	CcMADSsca24.1	NW_017985176.1	cajca.C.cajan_42711	XP_020207324	29323	29830	119	Agamous-like MADS-box protein AGL80	9.80 / 13527.86
100	CcMADSsca25.1	NW_017985198.1	cajca.C.cajan_44203	XP_020206928.1	4031	16909	244	Cauliflower D isoform	8.98 / 28327.47
101	CcMADSsca26.1	NW_017985343.1	cajca.C.cajan_48060	XP_020207324.1	31172	31926	201	Agamous-like MADS-box protein AGL80	9.51 / 22809.40
102	CcMADSsca27.1	NW_017985461.1	cajca.C.cajan_43653	XP_020207574.1	26311	26644	110	agamous-like MADS-box protein AGL62	4.89 / 12659.32

1

2

3

4

5

Table S3: Putative function and status of genes lacking MADS box domain as revealed by B2GO annotation in cultivated as well as wild type analysis

S. no	Gene name	Functional annotation using Blast2GO analysis	<i>C. cajan</i>	<i>C. Scarabaeoides</i>	<i>C. Platycarpus</i>	<i>C. Cajanifolius</i>
1	CcMADS1.6	MADS box protein SOC	P	A	A	A
2	CcMADS3.1	MADS box protein AGL8	P	P	P	P
3	CcMADS3.6	MADS box protein JOINTLESS	P	A	P	A
4	CcMADS6.3	MADS box protein JOINTLESS	P	P	P	P
5	CcMADS8.1	MADS box protein JOINTLESS	P	A	A	A
6	CcMADS11.3	MADS box transcription factor 27	P	A	A	A
7	CcMADSsca10.1	Agamous like MADS box protein	P	P	P	P
8	CcMADSsca11.1	Agamous like MADS box protein	P	P	A	P
9	CcMADSsca13.2	SEPALLATA 2	P	P	P	P
10	CcMADSsca17.2	Floral homeotic protein AGAMOUS	P	P	P	P
11	CcMADSsca18.1	floral homeotic protein AGAMOUS	P	P	P	P
12	CcMADSsca21.1	MADS box transcription factor 27	P	A	A	A

Table S4: Status of different MADS box genes of *C. cajan* in its wild relatives

S.no	Gene name	function	<i>C cajan</i>	<i>C. cajanifolius</i>	<i>C. scarabaeoides</i>	<i>C. platycarpus</i>
1	CcMADS1.1	AGL80	P	A	P	P
2	CcMADS1.2	AGL80	P	A	A	A
3	CcMADS1.6	SOC1	P	A	A	A
4	CcMADS1.7	TF 17	P	P	A	P
5	CcMADS2.3	AGL11	P	A	A	A
6	CcMADS2.7	AGL 42	P	A	P	A
7	CcMADS3.3	AGL 3	P	A	P	P
8	CcMADS3.6	Jointless	P	A	A	P
9	CcMADS6.1	FBP 24	P	A	P	A
10	CcMADS6.6	Agamous like	P	A	A	A
11	CcMADS8.1	Jointless	P	A	A	A
12	CcMADS8.4	AGL 29	P	A	A	A
13	CcMADS8.5	AGL 29	P	A	A	P
14	CcMADS9.1	Jointless	P	A	A	A
15	CcMADS9.3	Agamous like	P	A	A	A
16	CcMADS11.1	PHERES like	P	A	A	A
17	CcMADS11.2	Transcription factor 57	P	A	A	A
18	CcMADS11.3	Transcription factor 57	P	A	A	A
19	CcMADS11.4	AGL 15	P	A	A	A
20	CcMADS11.8	AGL 8	P	A	A	A
21	CcMADSsca1.2	AGL 9	P	A	P	A
22	CcMADSsca3.1	AGL 62	P	A	A	A
23	CcMADSsca6.2	AGL 62	P	A	A	A
24	CcMADSsca9.1	PHERES like	P	A	A	A

25	CcMADSsca9.2	AGL 80	P	A	A	A
26	CcMADSsca9.4	AGL 80	P	A	A	A
27	CcMADSsca9.5	AGL 80	P	A	A	A
28	CcMADSsca10.2	Agamous like	P	P	P	A
29	CcMADSsca11.1	AGL 16	P	P	P	A
30	CcMADSsca11.2	AGL 21	P	P	A	P
31	CcMADSsca15.1	AGL 80	P	A	P	P
32	CcMADSsca15.2	AGL 80	P	P	A	P
33	CcMADSsca16.1	AGL 80	P	P	P	A
34	CcMADSsca19.1	AGL 80	P	A	A	A
35	CcMADSsca20.2	AGL 80	P	P	P	A
36	CcMADSsca21.1	Transcription factor 27	P	A	A	A
37	CcMADSsca22.1	AGL 103	P	A	A	A
38	CcMADSsca23.1	AGL 80	P	A	A	A
39	CcMADSsca24.1	AGL 80	P	A	A	A
40	CcMADSsca25.1	Cauliflower D isoform	P	A	A	A
41	CcMADSsca26.1	AGL 80	P	A	A	A

Note: Other genes are present in both wild type and cultivated genotypes as mentioned in table S1.