BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(3)

Computational identification and analysis of MADS box genes in *Camellia sinensis*

Madhurjya Gogoi*, Sangeeta Borchetia & Tanoy Bandyopadhyay

Department of Biotechnology, Tea Research Association, Tocklai Tea Research Institute, Jorhat-785008, Assam, India; Madhurjya Gogoi – Email: mdhjyagogoi@gmail.com; *Corresponding author

Received January 23, 2015; Accepted February 18, 2015; Published March 31, 2015

Abstract:
MADS (Minichromosome Maintenance1 Agamous Deficiens Serum response factor) box genes encode transcription factors and they play a key role in growth and development of flowering plants. There are two types of MADS box genes- Type I (serum response factor (SRF)-like) and Type II (myocyte enhancer factor 2 (MEF2)-like). Type II MADS box genes have a conserved MIKC domain (MADS DNA-binding domain, intervening domain, keratin-like domain, and c-terminal domain) and these were extensively studied in plants. Compared to other plants very little is known about MADS box genes in *Camellia sinensis.* The present study aims at identifying and analyzing the MADS-box genes present in *Camellia sinensis*. A comparative bioinformatics and phylogenetic analysis of the *Camellia sinensis* sequences along with *Arabidopsis thaliana* MADS box sequences available in the public domain databases led to the identification of 16 genes which were orthologous to Type II MADS box gene family members. The protein sequences were classified into distinct clades which are associated with the conserved function of flower and seed development. The identified genes may be used for gene expression and gene manipulation studies to elucidate their role in the development and flowering of tea which may pave the way to improve the crop productivity.

Keywords: Bioinformatics, Crop productivity, Flowering, MADS box genes, Tea, Transcription factor.

Background:

MADS (Minichromosome Maintenance1 Agamous Deficiens Serum response factor) box genes are one of the best studied transcription factor family that are key regulators of development in almost all groups of eukaryotes and plays important role in the growth and development of flowering plants **[1].** The MADS box proteins possess a highly conserved DNA-binding MADS domain having a length of around 60 amino acids. MADS box genes are divided into two types–type I (serum response factor (SRF)-like) and type II (myocyte enhancer factor 2 (MEF2)-like). Type II has a conserved MIKC domain (MADS DNA-binding domain, intervening domain, keratin-like domain, and c-terminal domain). Only a few MADS-box genes of type I have been functionally characterized, whereas type II MADS box genes have been extensively studied in plants **[2]**. A lot of progress has been made in deciphering the molecular mechanism involved in the floral transition **[3]**. Further complete genome sequence of *Arabidopsis* provided more clear picture of the complexity and diversity of MADS box genes **[4]**. Many MADS box genes have been identified which are involved in various steps of transition from vegetative to reproductive growth. Most of the flowering genes encode transcription factors of MADS-box domain. Compared to other plants very little is known about the MADS box genes in *Camellia sinensis.* The floral buds of *C. sinensis* were found to be a major sink of assimilates produced by the maintenance foliage and are considered to be a limiting factor in proper partitioning of assimilates. It has been found that root starch was enriched when flower buds were controlled and it induced more vegetative growth. Considering the role of MADS box genes in the flowering of plants and its possible implication in improving tea productivity by controlling flowering with gene manipulation, the present study aimed at identifying and analyzing the MADS-box genes present in *Camellia sinensis*. Comparative bioinformatics and phylogenetic analysis identified the probable orthologous of *Arabidopsis* MADS box genes in tea. The protein sequences of the identified genes were classified into distinct clades and were found to be associated with the conserved function of flower and seed development. Biotechnological interventions on the identified MADS box genes will elucidate their role in flowering of tea and may also lead to increase in tea crop productivity.

Methodology:

Database search of MADS box sequences

NCBI NR, NCBI dbEST and NCBI TSA databases were used for the search of *Camellia sinensis* MADS Box sequences. Search for *Camellia sinensis* sequences was conducted using the tblastn module of NCBI blast. The query sequence for blast used was the band consensus of MADS region generated by the COBBLER program (Consensus Biasing by Locally Embedding Residues) **[5]** of the published MADS-box sequences of *Arabidopsis thaliana*. The *Camellia sinensis* blast hits having significant similarity (E-value cutoff 1e-15) and score greater than 100 were selected.

The reads obtained from blast hit were combined together, clustered and assembled using CAP3 program to form the contigs and singletons. The names of the contigs were prefixed as CsC and the singleton names were prefixed as CsS, followed by the number. To define putative coding frame of the transcripts, the NCBI ORF Finder tool was used. The transcripts open reading frame was determined and corresponding protein sequences were retrieved. Those sequences which were partial or had incomplete ORF were discarded from further analysis.

Conserved domain identification

For the verification of the MADS box conserved domain, the protein sequences were inspected by the NCBI Batch CD-Search program **[6]** and sequences without MADS box domain were discarded.

Phylogenetic analysis

For the Phylogenetic analysis, *Arabidopsis thaliana* MADS box gene sequences and corresponding protein sequences were retrieved from TAIR database (The Arabidopsis Information Resource) based on keyword search and published gene sequences. Amino acid sequences were used for phylogenetic analysis as they are more conserved compared to high variability of nucleotide sequences. The dataset for phylogenetic analysis contained the *Camellia sinensis* predicted MADS box protein sequences and the *Arabidopsis thaliana* MADS box protein sequences. MEGA 5 software **[7]** was used for the phylogenetic analysis. Sequence alignment was performed in MEGA using ClustalW and the phylogenetic tree was obtained using Neighbor-joining method **[8]** with Poisson distances and the pair-wise deletion option. For the reliability of the tree 1000 bootstrap replication were performed.

Motif identification and analysis

MEME program **[9]** was used for the identification of the motif cluster present in the MADS box sequences. The order of sequences in the phylogenetic tree was maintained in the input file for MEME program to facilitate the observation of common motif between the closely related sequences

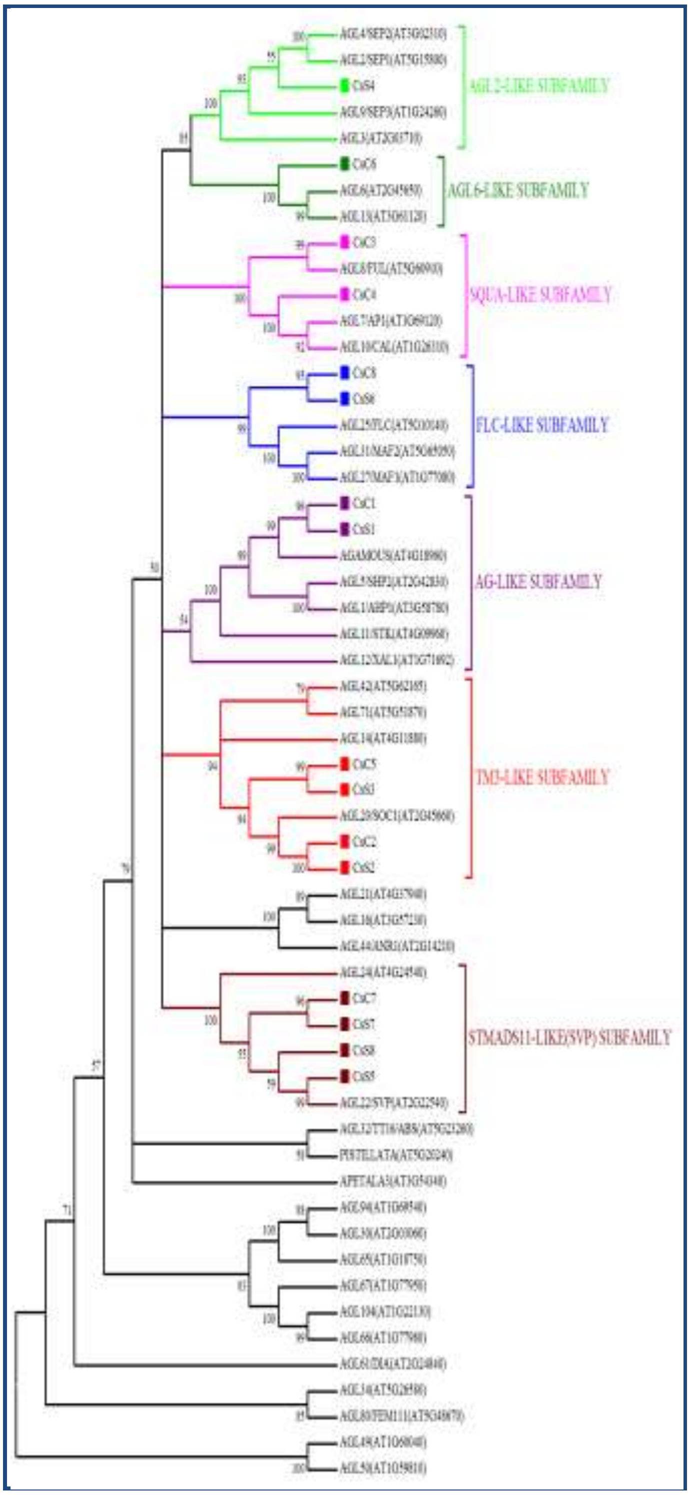

Figure 1: Phylogenetic Tree: Phylogenetic tree constructed based on MADS box protein sequences of *Camellia sinensis* and published *Arabidopsis thailiana* MADS box protein sequences. Neighbor-joining comparison model was used with poisson distances and Pairwise deletion option for the construction of the phylogenetic tree. Bootstrap values smaller than 50% were omitted and corresponding branches were merged. The *Camellia sinensis* protein sequences in the phylogenetic tree together with the *Arabidopsis thailiana* protein sequences were grouped mainly into seven subfamily (square bracket covering the subfamily members). The colours in the phylogenetic tree are used to graphically distinguish the subfamilies**.**

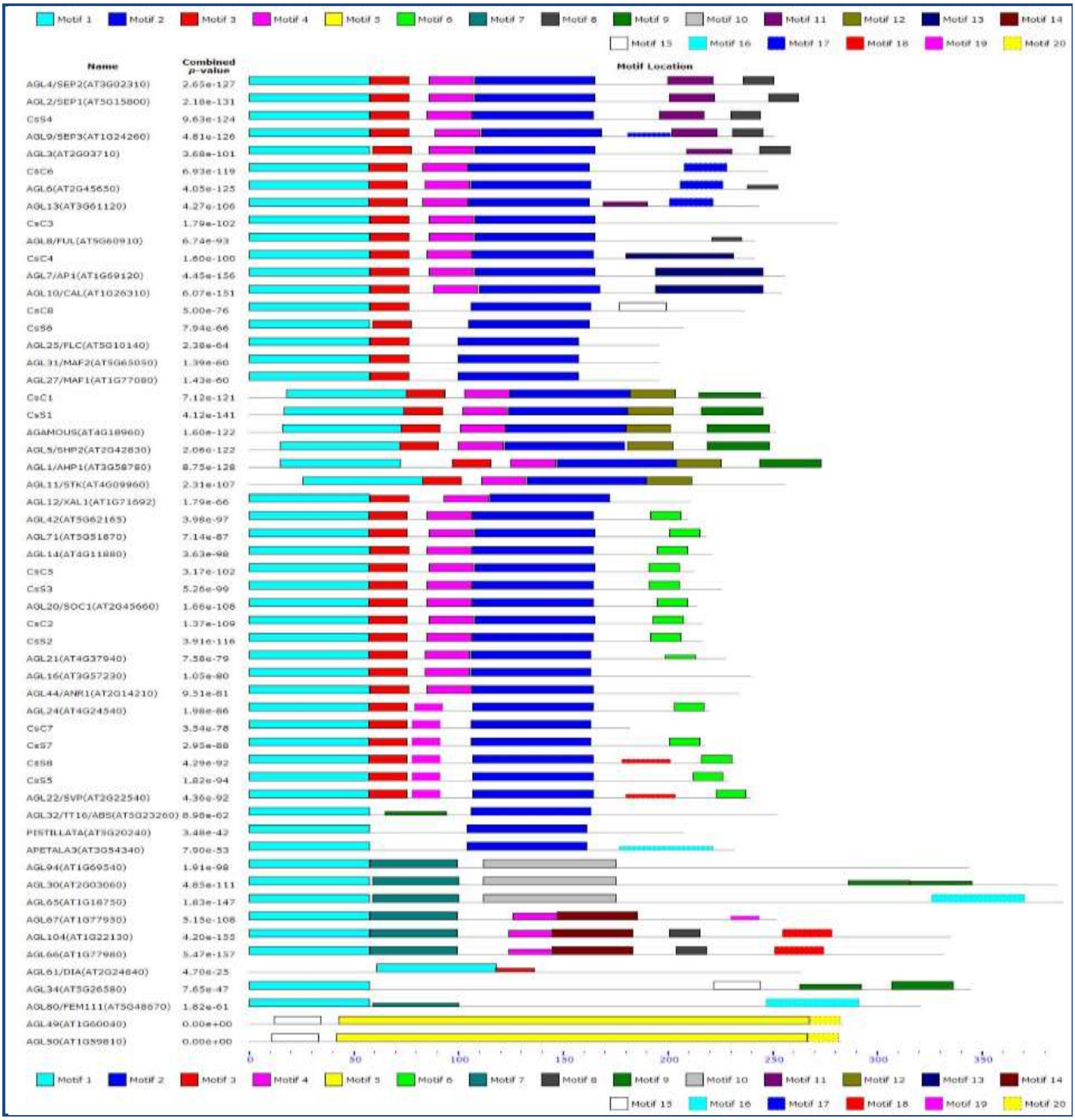

Figure 2: Graphic representation showing the complete grouping motifs of *Camellia sinensis* and *Arabidopsis thailiana* MADS box sequences obtained using MEME program (Multiple Expectation Minimization for Motif Elicitation, http://meme.sdsc.edu/ meme/ meme.html). The parameters used were: Distribution of motif occurrences- Zero or one per sequence, maximum number of motifs-20, Maximum motif width- 300 and Minimum motif width-6.

Result & Discussion:

The assembly of the sequences using CAP3 resulted in 13 contigs and 13 singletons. Based on the results of NCBI Batch CD-Search program, sequences without MADS box domain were discarded. Finally only 8 contigs and 8 singletons were further used for the analysis. Except two contigs (namely-CsC7, CsC8) and three singletons (namely-CsS5, CsS6, CsS7), all the others sequences were found to be complete in both N and C terminals **Table 1 (see supplementary material).** However all the sequences represented perfect MADS box domain including the two contigs and three singletons mentioned above. Accession number of contigs and singleton are provided in **Table 2 (see supplementary material).** Phylogenetic tree **(Figure 1)** comprising MADS box protein sequences of *Arabidopsis* and *C. sinensis*, showed that the sequences were clustered into different groups. All the transcripts from *C. sinensis* grouped with different subfamilies of type II MADS-box protein **(Figure 1).** The sequences could be further visualized by the analysis of the motif grouping results **(Figure 2)** of MEME program.

AGL2 subfamily

AGL2 subfamily is sister to the AGL6 subfamily with only one transcript from *C. sinensis* clustering with it **(Figure 1).** The CsS4 transcript appears to be homologous to AGL2 and AGL4 along with *AGL3*, *AGL9* forming one clade (**Figure 1).** For this subfamily motif grouping among sequences reflected the conserved feature **(Figure 2).** Studies pointed out that AGL2 gene may play a fundamental role in the floral organ identity and development along with seeds and embryo development **[3]**. By suppressing the expression of native AGL2 gene and other regulatory element linked to this gene by biotechnological approaches like antisense, co-suppression, gene replacement etc. delay in flower may be achieved which may led to increase in the length of vegetative phase and thus increase in vegetative tissue yield particularly in case of foliage crops.

AGL6 subfamily

The contig CsC6 clustered with the AGL6 subfamily and is seen to be very closely related to *AGL6* and *AGL13* **(Figure 1).** The overlying role of the AGL2 and AGL6 subfamily genes is also evident from the phylogenetic tree grouping of the gene subfamilies under one superclade **(Figure 1).** Sequences within the subfamily showed high similarity in motif grouping **(Figure 2**). In *Arabidopsis*, *AGL6* and *AGL13* belong to the AGL6 subfamily. Studies showed that AGL6 subfamily plays a key role in regulating floral organ identity **[10]** and floral meristem determinacy in rice, maize and Petunia hybrid. It also control circadian clock and is involved in the negative regulation of the FLC/MAF subfamily genes and positive regulation of FT genes **[11]**.

SQUA-Like subfamily

Two contigs namely CsC3 and CsC4 grouped together with the SQUA-Like subfamily (**Figure 1).** CsC3 appears to be homologous to Fruitfull *(FUL)* and CsC4 is homologous to Cauliflower (*CAL)* and Apetala 1 (*AP1)*. The functions of MADS box genes of the SQUA-Like subfamily includes controlling of transition from vegetative to reproductive growth, determining identity of the floral organ and regulating fruit maturation **[12]** . Apetala 1 (*AP1*) is one of the members of this subfamily. Together with the *FUL* and *CAL* genes, *AP1* act redundantly to control inflorescence architecture and meristem identity. Constitutive over expression of AP1 gene led to early flowering in transgenic *Chrysanthemum* plant **[13]**. Studies can be undertaken to down regulate or knockdown AP1 and related genes to see if this would extend the vegetative phase duration of foliage crop like tea.

FLOWERING LOCUS C (FLC) subfamily

The two transcripts from *C. sinensis* namely CsC8 and CsS6 formed another small group which is homologous to *FLC* (*AGL 25).* These two transcripts along with *FLC* (*AGL 25*), *MAF1* (*AGL27*) and *MAF2* (*AGL31*) gene of *Arabidopsis* formed one clade to represent the FLC subfamily in the phylogenetic tree **(Figure 1).** The *FLC* is a flowering transition repressor and also other members of the FLC like subfamily are directly involved in the flowering process to seasonal environmental factors. It also controls major life-history transition-seed germination and its expression is associated with natural variation in temperature-dependent germination **[14]**. Transgenic Chinese cabbage overexpressing *Brassica rapa* FLC showed delay in flowering and remained in vegetative phase for longer time **[15]**. Thus *FLC* gene appears to be one of the important target for genetic manipulation to increase biomass and get high yield in vegetative tissues.

AG subfamily

The two transcripts namely CsC1 and CsS1, clustered with AG subfamily were homologous to *AGAMOUS* (*AT4G18960*) of *Arabidopsis* **(Figure 1).** The motif grouping pattern **(Figure 2)** is highly similar among the two *C. sinensis* transcripts and *Arabidopsis AGAMOUS* (*AT4G18960*). The floral homeotic gene *AGAMOUS* (*AG*), a class C gene of the MADS-box transcription factor family is necessary for specification and development of stamen and carpels along with floral meristem determinacy **[16]**. In *Arabidopsis*, it interacts with other MADS box proteins and plays an important role for the induction of reproductive organ development **[17]**. Besides AGAMOUS, few more genes namely AGL1, AGL5, AGL11 and AGL12 **(Figure 1)** constitute the AG subfamily. In situ hybridization studies have shown that AGAMOUS gene is transcribed with strong expression only in the third and fourth floral whorl, after the floral bud formation just before the primordia of stamens and carpels **[3]**. It doesn't interfere with the normal vegetative growth of the plant and thus it may acts as a suitable target in genetic modification for crop improvement.

Tomato MADS box transcription factor3 (TM3) like genes subfamily

A total of four transcripts namely CsC2, CsC5, CsS2 and CsS3 were grouped with the TM3 like subfamily **(Figure 1)** with common grouping motifs **(Figure 2).** All these four transcripts showed being homologue to *AGL20/SOC1.* The TM3 like subfamily clade also contains *AGL14*, *AGL42* and *AGL71* genes of Arabidopsis. Many member of this subfamily showed expression both in vegetative and reproductive organs of angiosperms. **[18]**. The *SOC1* gene of TM3 subfamily is regulated by several pathways and it co-ordinate the responses to environmental signals. As SOC1 acts as a major hub in the regulatory network of floral timing and development **[19]**, it may acts as an important target for biotechnological intervention for crop improvement by means of over expression or under expression of the SOC1 polypeptide. Over expression of SOC1 may result in early flowering, increase in flower production and increase in fruit production. Whereas under expression of the same may benefit foliage crops by extending vegetative phase duration.

STMADS11-Like (SVP - SHORT VEGETATIVE PHASE)

SVP is another major subfamily where four transcripts from *C. sinensis* grouped with it **(Figure 1).** CsC7 and CsS7 appear to be homologue of *AGL24* gene, whereas CsS5 and CsS8 were homologous to the *AGL22* or *SVP* gene. *SVP* gene play important role in two developmental phases of plants. During the vegetative phase it acts as a repressor of the floral transition and later it plays a vital role in the specification of floral meristem **[20]**. Alteration in gene expression of the SVP gene in *C. sinensis* via genetic modification may lead to late flowering and extension of vegetative growth phase.

Conclusion:

In this study, presence of Type II MADS domain protein members and absence of member from type I MADS domain proteins were observed in *Camellia sinensis*. In future, studies can be undertaken to understand the molecular mechanisms of the MADS-box proteins in growth, development and stress conditions of *C. sinensis* and also steps can be undertaken to develop trans or cisgenic plants without seeds. MADS box gene family appears to be promising target to obtain such plants considering its role in reproductive development of plants. Seedless genetically modified tea plant with desired characters if obtained might be highly welcomed by tea community, provided the route to ethical issue relating to transgenics is clear. Seedless transgenic plants with desired trait will also eliminate the danger of uncontrolled dispersal of transgenic plants in environment. Tea plant is mainly propagated through vegetative cloning so absence of seed should not be a problem for its cultivation. Issue relating to narrowing down of genetic diversity of tea through such seedless plantation can be overcome by taking steps to preserve the genetic pool of the tea. As tea is a foliage crop vegetative phase is more important than the reproductive phase. Control of flowering may increase the vegetative growth. There is a possibility to enhance the vegetative growth of tea plants by manipulating the genes involved in flower development. Thus, understanding the role of MADS box gene family in tea plant may pave the way for improving the crop productivity and benefit the tea industry.

Conflict of interest: None

Acknowledgment:

The authors are grateful to Dr. N. Muraleedharan, Director, Tea Research Association for his constant encouragement and support. This work was supported by Department of Electronics and Information Technology (DeitY), Govt. of India.

References:

[1] Kapazoglou A *et al. BMC Plant Biol.* 2012 **12:** 166 [PMID: 22985436]

[2] Hanschel K *et al. Mol Biol Evol.* 2002 **19**: 801 [PMID: 12032236]

[3] Goto KJ *Biosci.* 1996 **21**: 369

[4] Oliveira RR *et al. Plant Mol Biol Rep.* 2010 **28**: 460

[5] http://blocks.fhcrc.org/blocks/cobbler.html

[6] Marchler-Bauer A *et al. Nucleic Acids Res.* 2011 **39**: D225 [PMID: 21109532]

[7] Tamura K *et al. Mol Biol Evol.* 2011 **28**: 2731 [PMID: 21546353]

[8] Saitou N & Nei M, *Mol Biol Evol.* 1987 **4**: 406 [PMID: 3447015]

[9] Bailey TL *et al. Nucleic Acids Res.* 2006 **34**: W369 [PMID: 16845028]

[10] Koo SC *et al. Plant J.* 2010 **62**: 807 [PMID: 20230491]

[11] Yoo SK *et al. Plant J.* 2011 **65**: 62 [PMID: 21175890]

[12] Fornara F *et al. Plant Physiol.* 2004 **135**: 2207 [PMID: 15299121]

[13] Shulga OA *et al. In Vitro Cellular & Developmental Biology - Plant.* 2011 **47**: 553

[14] Chiang GCK *et al. Proc Natl Acad Sci USA.* 2009 **106**: 11661 [PMID: 19564609]

[15] Kim SY *et al. Plant Cell Rep.* 2007 **26**: 327 [PMID: 17024448]

[16] Yellina AL *et al. Evodevo.* 2010 **1**: 13 [PMID: 21122096]

[17] Honma T & Goto K, *Nature* 2001 **409**: 525 [PMID: 11206550]

[18] Theissen G *et al. Plant Mol Biol.* 2000 **42**: 115 [PMID: 10688133]

[19] Immink RG *et al. Plant Physiol.* 2012 **160**: 433 [PMID: 22791302]

[20] Liu C *et al. Dev Cell.* 2009 **16**: 711 [PMID: 19460347]

Edited by P Kangueane

Citation: Gogoi *et al.* Bioinformation 11(3): 115-121 (2015)

Supplementary material:

Table 1: NCBI Batch CD-Search results.

Query: (Q#N) in the table below indicates the ordinal number (N) of the query sequence from the original input file; **Hit type:** Represents various confidence levels (specific hits, non-specific hits) and domain model scope (superfamilies, multi-domains); **PSSM ID:** It is the unique identifier for a domain model's position-specific scoring matrix (PSSM); **From and To:** Shows the range of amino acids in the query protein sequence to which the domain model aligns; **E-value:** Indicates the statistical significance of the hit as the likelihood the hit was found by chance; **Bitscore:** The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account; **Accession:** The accession column indicate the accession number of the hit, which can either be a domain model or a superfamily cluster; **Short name :** Shows the short name of a conserved domain, which concisely defines the domain; **Incomplete :** If the hit to a conserved domain is partial (i.e., if the alignment found by RPS-BLAST omitted more than 20% of the CD's extent at either the n- or c-terminus or both), this column will be populated with one of the following values: N: incomplete at the N-terminus C: incomplete at the C-terminus NC: incomplete at both the N-terminus and C-terminus (the alignment found by RPS-BLAST omitted more than 40% of the CD's total extent) If the hit to a conserved domain is complete, then this column will be populated with a dash (-); **Superfamily:** This column is populated only for domain models that are specific or non-specific hits, and it lists the accession number of the superfamily to which the domain model belongs. (If the hit is to a superfamily itself, then this column is simply populated with a dash because the superfamily accession is already listed in the preceding "Accession" column.) (Source: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml).

Query	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accession	Short name	Incomplete	Superfamily
Q#1 - >CsC1	specific	238165	20	93	1.06E-43	144.232	cd00265	MADS_MEF2_like	-	cl00109
Q#1 - >CsC1	superfamily	241616	20	93	1.06E-43	144.232	cl00109	MADS superfamily	-	-
Q#1 - >CsC1	specific	250655	95	191	1.46E-33	118.398	pfam01486	K-box	-	cl03234
Q#1 - >CsC1	superfamily	250655	95	191	1.46E-33	118.398	cl03234	K-box superfamily	-	-
Q#2 - >CsC2	specific	238165	3	75	6.55E-43	141.15	cd00265	MADS_MEF2_like	-	cl00109
Q#2 - >CsC2	superfamily	241616	3	75	6.55E-43	141.15	cl00109	MADS superfamily	-	-
Q#2 - >CsC2	specific	250655	83	172	1.50E-20	82.9598	pfam01486	K-box	-	cl03234
Q#2 - >CsC2	superfamily	250655	83	172	1.50E-20	82.9598	cl03234	K-box superfamily	-	-
Q#3 - >CsC3	specific	238165	2	79	4.30E-42	141.15	cd00265	MADS_MEF2_like	-	cl00109
Q#3 - >CsC3	superfamily	241616	2	79	4.30E-42	141.15	cl00109	MADS superfamily	-	-
Q#3 - >CsC3	specific	250655	75	168	2.36E-31	113.391	pfam01486	K-box	-	cl03234
Q#3 - >CsC3	superfamily	250655	75	168	2.36E-31	113.391	cl03234	K-box superfamily	-	-
Q#4 - >CsC4	specific	238165	2	79	5.75E-42	139.61	cd00265	MADS_MEF2_like	-	cl00109
Q#4 - >CsC4	superfamily	241616	2	79	5.75E-42	139.61	cl00109	MADS superfamily	-	-
Q#4 - >CsC4	specific	250655	89	167	1.83E-32	115.317	pfam01486	K-box	-	cl03234
Q#4 - >CsC4	superfamily	250655	89	167	1.83E-32	115.317	cl03234	K-box superfamily	-	-
Q#5 - >CsC5	specific	238165	3	76	4.83E-40	133.446	cd00265	MADS_MEF2_like	-	cl00109
Q#5 - >CsC5	superfamily	241616	3	76	4.83E-40	133.446	cl00109	MADS superfamily	-	-
Q#5 - >CsC5	specific	250655	76	171	2.64E-21	84.8858	pfam01486	K-box	-	cl03234
Q#5 - >CsC5	superfamily	250655	76	171	2.64E-21	84.8858	cl03234	K-box superfamily	-	-
Q#6 - >CsC6	specific	238165	2	73	7.52E-46	149.625	cd00265	MADS_MEF2_like	-	cl00109
Q#6 - >CsC6	superfamily	241616	2	73	7.52E-46	149.625	cl00109	MADS superfamily	-	-
Q#6 - >CsC6	specific	250655	82	169	1.42E-33	118.398	pfam01486	K-box	-	cl03234
Q#6 - >CsC6	superfamily	250655	82	169	1.42E-33	118.398	cl03234	K-box superfamily	-	-
Q#7 - >CsC7	specific	238165	3	75	1.25E-38	128.824	cd00265	MADS_MEF2_like	-	cl00109
Q#7 - >CsC7	superfamily	241616	3	75	1.25E-38	128.824	cl00109	MADS superfamily	-	-
Q#7 - >CsC7	superfamily	250655	106	170	3.30E-09	50.603	cl03234	K-box superfamily	N	-
Q#8 - >CsC8	specific	238165	2	76	1.96E-38	130.365	cd00265	MADS_MEF2_like	-	cl00109

Q#8 - >CsC8	superfamily	241616	2	76	1.96E-38	130.365	cl00109	MADS superfamily	-	-
Q#8 - >CsC8	superfamily	250655	102	161	0.001036	36.3507	cl03234	K-box superfamily	N	-
Q#9 - >CsS1	specific	238165	19	95	1.28E-44	146.543	cd00265	MADS_MEF2_like	-	cl00109
Q#9 - >CsS1	superfamily	241616	19	95	1.28E-44	146.543	cl00109	MADS superfamily	-	-
Q#9 - >CsS1	specific	250655	91	190	2.47E-36	125.717	pfam01486	K-box	-	cl03234
Q#9 - >CsS1	superfamily	250655	91	190	2.47E-36	125.717	cl03234	K-box superfamily	-	-
Q#10 - >CsS2	specific	238165	3	75	6.14E-43	141.15	cd00265	MADS_MEF2_like	-	cl00109
Q#10 - >CsS2	superfamily	241616	3	75	6.14E-43	141.15	cl00109	MADS superfamily	-	-
Q#10 - >CsS2	specific	250655	75	171	2.41E-24	93.3602	pfam01486	K-box	-	cl03234
Q#10 - >CsS2	superfamily	250655	75	171	2.41E-24	93.3602	cl03234	K-box superfamily	-	-
Q#11 - >CsS3	specific	238165	3	78	1.14E-41	138.069	cd00265	MADS_MEF2_like	-	cl00109
Q#11 - >CsS3	superfamily	241616	3	78	1.14E-41	138.069	cl00109	MADS superfamily	-	-
Q#11 - >CsS3	specific	250655	78	172	4.02E-22	87.5822	pfam01486	K-box	-	cl03234
Q#11 - >CsS3	superfamily	250655	78	172	4.02E-22	87.5822	cl03234	K-box superfamily	-	-
Q#12 - >CsS4	specific	238165	2	78	3.42E-45	148.084	cd00265	MADS_MEF2_like	-	cl00109
Q#12 - >CsS4	superfamily	241616	2	78	3.42E-45	148.084	cl00109	MADS superfamily	-	-
Q#12 - >CsS4	specific	250655	75	173	3.20E-33	117.628	pfam01486	K-box	-	cl03234
Q#12 - >CsS4	superfamily	250655	75	173	3.20E-33	117.628	cl03234	K-box superfamily	-	-
Q#13 - >CsS5	specific	238165	2	75	5.09E-38	128.824	cd00265	MADS_MEF2_like	-	cl00109
Q#13 - >CsS5	superfamily	241616	2	75	5.09E-38	128.824	cl00109	MADS superfamily	-	-
Q#13 - >CsS5	superfamily	250655	107	173	8.90E-16	69.863	cl03234	K-box superfamily	N	-
Q#14 - >CsS6	specific	238165	2	77	1.84E-36	124.202	cd00265	MADS_MEF2_like	-	cl00109
Q#14 - >CsS6	superfamily	241616	2	77	1.84E-36	124.202	cl00109	MADS superfamily	-	-
Q#14 - >CsS6	superfamily	250655	100	171	1.12E-12	61.0034	cl03234	K-box superfamily	N	-
Q#15 - >CsS7	specific	238165	3	69	8.07E-39	130.75	cd00265	MADS_MEF2_like	-	cl00109
Q#15 - >CsS7	superfamily	241616	3	69	8.07E-39	130.75	cl00109	MADS superfamily	-	-
Q#15 - >CsS7	superfamily	250655	106	171	4.89E-08	48.2918	cl03234	K-box superfamily	N	-
Q#16 - >CsS8	specific	238165	12	77	4.86E-32	113.416	cd00265	MADS_MEF2_like	-	cl00109
Q#16 - >CsS8	superfamily	241616	12	77	4.86E-32	113.416	cl00109	MADS superfamily	-	-
Q#16 - >CsS8	superfamily	250655	70	170	1.31E-15	69.4778	cl03234	K-box superfamily	-	-

Table 2: Contigs and Singletons sequences of *C sinensis* used in the analysis of MADS box genes after assembly.

Contigs	NCBI Accession	Singletons	NCBI Accession
CsC1	HP740334, GAAC01041694	CsS1	HP751613
CsC2	KA286264, GAAC01012023, GAAC01000659	CsS2	KA285338
CsC3	HP768082, KA302732, GAAC01010609, KA281908	CsS3	HP753371
CsC4	GAAC01050564, HP740164	CsS4	GAAC01009021
CsC5	HP755617, GAAC01006965	CsS5	HP765055
CsC6	GAAC01024468, HP753883	CsS6	KA286666
CsC7	GAAC01014198, KA287022	CsS7	KA281481
CsC8	HP753445, KA294793	CsS8	KA281696