GENOME-WIDE CHARACTERIZATION AND PHYLOGENETIC AND EXPRESSION ANALYSES OF THE CALEOSIN GENE FAMILY IN SOYBEAN, COMMON BEAN AND BARREL MEDIC

Yue Shen[1,2], Qing-Li Jia[1], Ming-Zhe Liu[1], Zhuo-Wei Li[1], Li-Li Wang[1], Cui-Zhu Zhao[1], Zhi-Xi Li[2,*] and Meng Zhang[1,*]

[1] *College of Agronomy, Northwest A&F University*, Yangling 712100, Shaanxi, People's Republic of China

[2] *College of Food Science and Engineering, Northwest A&F University*, Yangling 712100, Shaanxi, People's Republic of China

*__Corresponding authors__: zhangm@nwsuaf.edu.cn; lizhxi@nwsuaf.edu.cn

Received: September 16, 2015; **Revised**: December 29, 2015; **Accepted**: December 29, 2015; **Published online**: April 27, 2016

Abstract: Caleosin are a class of calcium-binding proteins embedded in the phospholipid monolayer of lipid droplets. In addition to maintaining the structure of lipid droplets, caleosin proteins are involved in dormancy and lipid signaling, and are associated with the stress response via their histidine-dependent peroxygenase activity. To date, caleosins have been studied in *Arabidopsis thaliana*. However, little is known about these genes in legumes, including the most cultivated oilseed crop, soybean. In this paper, 20 caleosin genes in soybean, common bean and barrel medic were studied. Among these, 13 caleosin genes, including 3 in *Glycine max*, 5 in *Phaseolus vulgaris* and 5 in *Medicago truncatula*, are identified for the first time. The structures, characteristics and evolution of the 20 caleosin proteins are analyzed. Expansion patterns show that tandem duplication was the main reason for the caleosin family expansion in the legume. Expression profiles indicate that L-caleosin in soybean and common bean are more important than H-caleosin, which is just the opposite in *Arabidopsis thaliana*. *GmaCLO2*, *PvuCLO1*, *PvuCLO3* and *MtrCLO3* may play important roles, while *GmaCLO6*, *GmaCLO10* and *MtrCLO4* may lose their function in the examined tissues. In addition, according to the results of *cis*-element analyses, we propose potential functions for the more important caleosin genes in leguminous plants. Our work provides helpful information for further evolution and function analyses of the caleosin gene family in soybean, common bean and barrel medic.

Key words: Caleosin; evolution; expression; function; legume

INTRODUCTION

Soybean (*Glycine max*) is not only the largest source of animal protein feed, but also the most cultivated oilseed crop, and the second largest vegetable oil source in the world. Soybean seed oil is used to supply essential nutrition for humans as well as renewable materials for bioenergy production [1].

Triacylglycerols (TAGs), which are packed in organelles called lipid droplets (LDs), are the most abundant ingredient in seed oil. TAGs in LDs are enveloped by a monolayer of phospholipids embedded with some unique proteins including oleosins, caleosins and steroleosins [2,3].

Caleosins are a class of proteins that are only found in plants and fungi [4]. Caleosins are structurally similar to the oleosins, consisting of three domains: an *N*-terminal hydrophilic domain, a central hydrophobic domain containing a proline knot for anchoring LDs, and a *C*-terminal hydrophilic domain. Unlike the protein structure of oleosins, the *N*-terminal of caleosins include an EF-hand calcium-binding motif and the *C*-terminal of caleosins contains several phosphorylation sites [3].

Studies of caleosins, which mainly focus on the model plant *Arabidopsis thaliana*, suggest that they may have a structural role and other unique functions. For instance, AtCLO1 is associated with lipid breakdown; AtCLO2 is involved in dormancy; AtCLO1, AtCLO2, AtCLO3 and AtCLO4 have haem-dependent peroxygenase activity, which may be associated with plant stress responses [2,5-9]. However, limited information is available on leguminous caleosin proteins. Recently, a 27 KDa and a 29 KDa caleosin protein were identified in extracted soybean LDs [10]. Seven caleosin genes were identified in the soybean (*Glycine max*) genome [11]. However, some questions remain unanswered. Are there any additional caleosin genes in soybean? How many caleosin genes are there in common bean and barrel medic? How do they express and evolve? Which of these caleosin genes are the most important and what are the functions of these important genes?

The soybean genome provides an opportunity to analyze the caleosin gene family at a genomic level. The paleopolyploid soybean went through genome duplication 58 million years ago (MYA) and 13 MYA [12]. After the duplication, many genes underwent loss or local duplication. The common bean (*Phaseolus vulgaris*) and soybean split 19.2 MYA [13], and barrel medic (*Medicago truncatula*) and soybean diverged 54 MYA [14]. The available genome information about common bean [13] and barrel medic [14] facilitates our study of the evolution patterns and other analyses of the caleosin gene family in soybean.

In this study, 10, 5 and 5 caleosin genes were identified in the latest genomes of soybean (*Glycine maxWm82.a2.v1*), common bean (*Phaseolus vulgaris v1.0*) and barrel medic (*Medicago truncatulaMt4.0v1*), respectively. Moreover, the classification, characteristics, evolution analyses of these caleosin proteins and expression profiles of the genes were analyzed. In addition, potential functions of important leguminous caleosin proteins are proposed.

MATERIALS AND METHODS

Identification of caleosin genes and characterization of caleosin proteins in three sequenced legume genomes

In order to obtain putative caleosin genes in three sequenced legume genomes, BLAST and Keyword searches were executed. The sequences of 8 Arabidopsis caleosin genes were obtained from the Arabidopsis Information Resource (TAIR10.0, http://www.arabi-dopsis.org/) [15], and used as queries to search each of the latest legume protein databases using “BLASTP” in phytozome 10 (http://phytozome.jgi.doe.gov/pz/portal.html) [16], with E-value<10^{-10}. For soybean, the latest database was *Glycine max-Wm82.a2.v1*; for common bean it was *Phaseolus vulgaris v1.0*, and for barrel medic it was *Medicago truncatulaMt4.0v1*. Sequences of candidate legume caleosin proteins were gathered to run BLASTP in TAIR to get the most similar corresponding genes in Arabidopsis. If the best hit gene in Arabidopsis was not a “caleosin”, the candidate legume caleosin was excluded. The Keyword search was also performed using “caleosin” as a query in each legume genome of Phytozome 10 [16]. Then, the redundant sequences were deleted manually. The deduced sequences of candidate caleosin proteins were checked for caleosin domains (IPR007736) using InterProScan 5 (http://www.ebi.ac.uk/interpro/search/sequence-search) [17].

The molecular weights [13] and isoelectric points (pI) of caleosin proteins were calculated on ExPASy (http://web.expasy.org/compute_pi/) [18]. Ser, Thr and Tyr phosphorylation sites of caleosin proteins in the three legumes were predicted by NetPhos 2.0 (http://www.cbs.dtu.dk/services/NetPhos/) [19]. To figure out the motif’s distribution in leguminous caleosin proteins, MEME (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) [20] was operated with 10 different motifs. The motifs were annotated by InterProScan5 (http://www.ebi.ac.uk/interpro/search/sequence-search) [17] and multiple alignment.

Multiple alignment and phylogenetic analysis

Multiple sequence alignment of leguminous caleosin proteins was performed by ClustalX 2.1 [21]. Shading and output of the results were executed using Boxshade 3.21 in ExPASy (http://www.ch.embnet.org/software/BOX_form.html) [18]. To evaluate the evolution relationship of caleosin proteins in Arabidopsis and the legumes, multiple sequence alignment was performed again with the leguminous caleosin proteins and 8 Arabidopsis caleosin proteins. Then, a neighbor-joining (NJ) tree was constructed by MEGA 6.06 [22] with the parameters: Poisson model, pairwise deletion of gaps and 1000 bootstrap; a maximum-likelihood (ML) tree was constructed by PhyML 3.0 (http://www.atgc-montpellier.fr/phyml/) [23] with the parameters: LG model, substitution rate 6, SPR improvement and 100 bootstrap. Both trees were visualized using MEGA 6.06 [22].

Chromosome distribution and duplication of caleosin genes

The location and flanking genes of caleosin genes were obtained in JBrowse in phytozome 10 (http://phytozome.jgi.doe.gov/pz/portal.html) [16] using loci names. The caleosin genes separated by 0 to 5 genes were regarded as tandem duplication [24]. Segmental duplication was searched according to a previous study [25]. Ks values were calculated using the codeml program of PAML in PAL2NAL (http://www.bork.embl.de/pal2nal/) [26]. The ages of duplication were calculated using the formula T=Ks/2λ; for soybean, λ is $5.17*10^{-9}$ [12].

Expression analysis of caleosin genes

Expression patterns of caleosin genes were examined by analyzing data in public databases. We used RNA-seq data and ESTs (expressed sequence tags) for soybean and common bean, and microarray data and ESTs for barrel medic. ESTs and cDNAs were collected from the EST database in the NCBI (http://www.ncbi.nlm.nih.gov/nucest/). Leguminous caleosin genes were used as queries to search the corresponding EST database, using BLASTN, with the following parameters: E-value<10^{-10}, identity>95% and length>200bp. To investigate gene expression in different tissues/organs, all libraries were assigned into 12, 8, 10 synthetic libraries in soybean, common bean and barrel medic, respectively (Table S2). Then, the corresponding ESTs of each gene in each synthetic library were tallied and normalized to transcripts per million (TPM).

RNA-seq data of soybean and common bean were obtained from PhytoMine in Phytozome10.0, using gene loci as queries.

Microarray data of barrel medic were searched in the database *Medicago truncatula* gene expression atlas (http://mtgea.noble.org/v3/) [27]. All microarray data were normalized using mean values of all replicates of the same experiment. As the genomes from version 4 were used, expression data could not found by gene ID search which includes version 3.5. Instead, BLASTN was employed to find the corresponding microarrays, using barrel medic caleosin genes as queries with identity>95%, E-value<10^{-10} and length>200bp.

A heat map was drawn for visualizing expression by R script [28] with log2 transformed normalized expression values.

Cis-regulatory elements analysis of important leguminous caleosin genes

The 2 kb upstream sequences of important leguminous caleosin genes were obtained from corresponding genomes. Putative *cis*-regulatory elements were identified by New PLACE (https://sogo.dna.affrc.go.jp/cgi-bin/sogo.cgi?sid=&lang=en&pj=640&action=page&page=newplace) [29].

RESULTS

Identification of caleosin genes and classification of caleosin proteins in soybean, common bean and barrel medic

Based on Keyword and BLAST searches in Phytozome, 11, 5 and 5 caleosin genes were identified in soybean, common bean and barrel medic, respectively. After BLASTP in TAIR, no caleosin genes were deleted. Among these 21 leguminous caleosin genes, two adjacent genes in soybean were found. BLASTP in TAIR for these two genes showed that the upstream one was homologous to the 5' end of At5g55240 while the downstream one was homologous to the 3' end of At4g26740. Then, the genomic regions spanning these two genes were obtained, and were re-predicted using the software FGENESH. The obtained result showed that there was a single gene, renamed Glyma.20G200900*, suggesting wrong annotation in the soybean genome. Then, 10, 5 and 5 caleosin genes were finally identified in soybean, common bean and barrel medic, respectively.

Caleosin genes were named according to their orders on chromosomes: *GmaCLO1-GmaCLO10* in soybean, *PvuCLO1-PvuCLO5* in common bean, and *MtrCLO1-MtrCLO5* in barrel medic (Table 1).

Multiple alignments were performed using all 20 complete protein sequences. The results of some alignments were poor, which might be due to the severely truncated sequences. Subsequently, these truncated proteins, GmaCLO6 and MtrCLO4, were deleted. Multiple alignments were performed again using the remaining 18 complete protein sequences (Fig. S1). According to the results of multiple alignment,

Table 1. Classification and characteristics of caleosin proteins in three legumes

name	Phytozome gene locus ID	Alias	pI	Mw	location	type
GmaCLO1	Glyma.03G249900	Glyma03g41030	7.10	27167.95	Chr03:44595527..44597546	H
GmaCLO2	Glyma.09G123800	Glyma09g22306	8.92	22592.73	Chr09:30003271..30006728	L
GmaCLO3	Glyma.09G124200	Glyma09g22330	9.24	22515.79	Chr09:30055462..30058048	L
GmaCLO4	Glyma.09G124400	Glyma09g22455	7.07	22568.75	Chr09:30189230..30195627	L
GmaCLO5	Glyma.09G124600	Glyma09g22580	7.86	22828.84	Chr09:30310712..30316610	L
GmaCLO6	Glyma.09G124700		NA[a]	NA	Chr09:30320809..30321917	L
GmaCLO7	Glyma.09G137700	Glyma09g25350	6.13	18357.12	Chr09:34132657..34133577	H
GmaCLO8	Glyma.10G189900	Glyma10g33350	6.04	22178.12	Chr10:42289892..42291691	H
GmaCLO9	Glyma.19G247500	Glyma19g43680	8.95	35959.36	Chr19:49371851..49373718	H
GmaCLO10	Glyma.20G200900*[b]		5.69	14822.02	Chr20:43806781..43809024	H
PvuCLO1	Phvul.003G020500		7.13	23082.10	Chr03:1804467..1810304	L
PvuCLO2	Phvul.003G020700		9.12	22475.66	Chr03:1815434..1820386	L
PvuCLO3	Phvul.003G020800		9.28	22753.86	Chr03:1825785..1829000	L
PvuCLO4	Phvul.006G104600		6.71	27176.96	Chr06:22086070..22087862	H
PvuCLO5	Phvul.007G130900		5.90	27579.41	Chr07:31818481..31820678	H
MtrCLO1	Medtr1g032160		8.60	23533.80	chr1:11392576..11395858	L
MtrCLO2	Medtr1g032190		9.34	26456.41	chr1:11401162..11404856	L
MtrCLO3	Medtr1g073640		6.08	27289.98	chr1:32673521..32675028	H
MtrCLO4	Medtr6g082510		NA	NA	chr6:30835040..30835673	NA
MtrCLO5	Medtr7g115490		6.66	27222.20	chr7:47728374..47730461	H

a. NA means no available value.

b. Glyma.20G200900* is a gene re-analyzed by FGENESH using gene models Glyma.20G200900 and Glyma.20G201000.

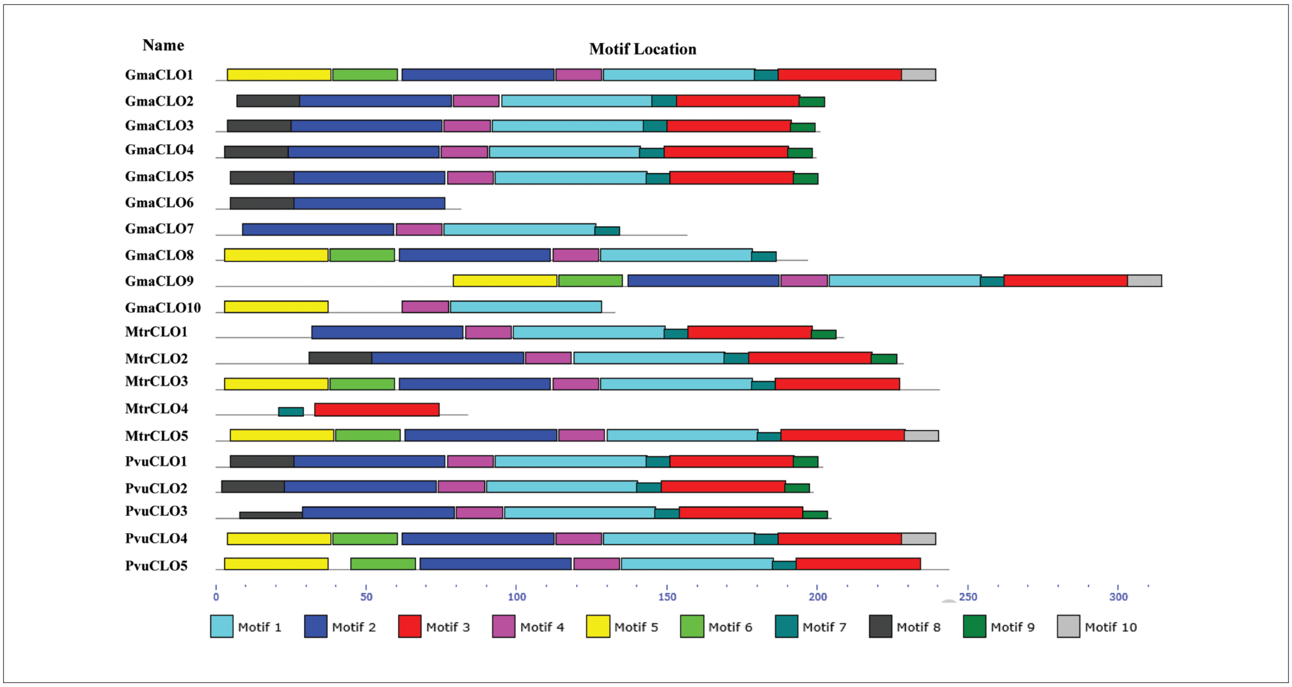

Fig. 1. Distribution of conserved motifs in 20 caleosin proteins of legumes. Sequences and annotations of the motifs are shown in Table S1.

apart from MtrCLO4 and GmaCLO6, 9 caleosin proteins were classified as both H-types and L-types.

Alignment results also showed that the *N*-terminal of GmaCLO9 was especially long (Fig. S1), which is characteristic of caleosin proteins in monocot [30]. The EF-hand domain and proline knot, which are important to caleosin proteins, were examined. The EF-hand domain was constant, except for GmaCLO10 (Fig. S1), which suggested that GmaCLO10 may lose its calcium-binding ability. And proline knots of the 18 caleosin proteins were highly conserved. In addition, previous research indicated that His^{70} and His^{133} in AtCLO1 were found to be crucial for peroxygenase activity; a similar result was also obtained in AtCLO3 and AtCLO4 [7-9]. These two sites were found to be highly conserved in all 18 leguminous caleosin proteins, except for GmaCLO10. GmaCLO10 lost His^{70} because of the medium sequence loss (Fig. S1).

Potential phosphorylation sites of serine, threonine and tyrosine were analyzed by NetPhos 2.0 and are given in Fig. S1. S^{118}, T^{136} and T^{181} (according to GmaCLO2) were conserved in L-caleosin with an exception of PvuCLO3, while T^{85} (according to GmaCLO1) was conserved in H-caleosin. Y^{143} (according to GmaCLO1) was conserved in all the examined caleosin proteins, except GmaCLO2 and GmaCLO4 in L-caleosin (Fig. S1).

In addition, 12 residues were constant in the tested sequences, indicating these sites might play unique roles in the function of leguminous caleosin proteins (Fig. S1).

Property and motif analysis of leguminous caleosin proteins

After exclusion of the two truncated members (GmaCLO6 and MtrCLO4), the predicted pIs of L-caleosin proteins were all above 8.6, except GmaCLO4, GmaCLO5, PvuCLO1; and predicted pIs of H-caleosin proteins were all below 7.1, except GmaCLO9 (Table 1).

MEME was performed to analyze the motif configurations of leguminous caleosin proteins and the lost motifs in the truncated proteins. The results showed that all motifs were in accord with the char-

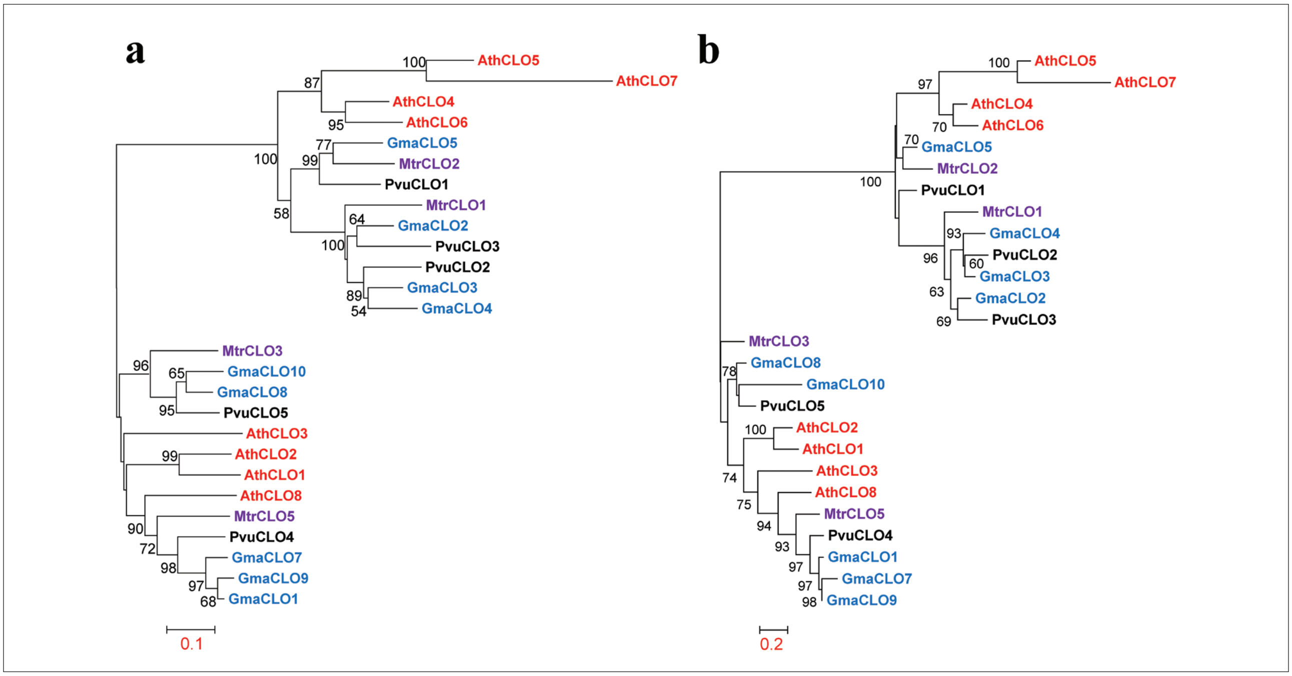

Fig. 2. Phylogenetic trees of 18 caleosin proteins in legumes and 7 caleosin proteins in Arabidopsis by the neighbor-joining method in MEGA 6.06 based on 1000 replications (a) and the maximum likelihood method in PhyML 3.0 based on 100 replications (b). Blue, black, purple and red represent caleosin proteins from soybean, common bean, barrel medic and Arabidopsis, respectively. Numbers next to a node represent percent bootstrap values (>50%). The bars at the bottom represent the average number of amino acid substitutions per site.

acters of "caleosin" (Fig. 1, Table S1). GmaCLO6 had only *N*-terminal motifs 2, 8, while MtrCLO4 had only *C*-terminal motifs 3, 7 (Fig. 1). Besides, in H-caleosin, motifs 5, 6, 3 and 10 were lost in GmaCLO7, motifs 3 and 10 were lost in GmaCLO8, and motifs 6, 2, 7, 3 and 10 were absent in GmaCLO10; in L-caleosin, motif 8 was absent in MtrCLO1 (Fig. 1). Moreover, GmaCLO6 contained motif 8, which was the *N*-terminal unique to L-caleosin. Thus, GmaCLO6 belonged to L-caleosin. This provided supplementary information for the classification based on multiple alignment.

Phylogenetic analysis of the leguminous caleosin proteins

To investigate the evolutionary relationships of caleosin proteins in the legumes, phylogenetic trees were built using the NJ method by MEGA and the ML method by PhyML. Two trees had similar topologies with high bootstrap values (mostly >80%), suggesting that the trees were reliable (Fig. 2). Based on the trees, 18 caleosin proteins (excluding the two truncated caleosin proteins) were assigned into two groups, which was identical to the two groups classified by the alignment.

Expansion analysis of caleosin genes in the three legumes

According to the chromosomal architecture of the leguminous caleosin genes (Table 1), one cluster including five genes (50%) was tandemly located on chromosome 9 in soybean; one cluster including three genes (60%) was tandemly located on chromosome 3 in common bean, while a pair of genes (40%) was tandemly located on chromosome 1 in barrel medic. It was noteworthy that L-caleosin underwent tandem duplication. Moreover, two pairs containing four genes (40%) were identified to be a segmental duplication in soybean, while no gene was found to be a segmental duplication in common bean and barrel medic (Table 2). These four segmentally duplicated caleosin genes belong to H-caleosin. According to Ks values, *GmaCLO1* and *GmaCLO9* might have been duplicated ~11.61 million years ago (MYA), while *GmaCLO8* and *GmaCLO10*, might have been duplicated ~58.99MYA (Table 2).

Evolution of caleosin genes in three legumes and Arabidopsis

By combining phylogenetic trees and evolution age, evolution trees were drawn (Fig. 3). The duplications of H-caleosin and L-caleosin in the legumes and Arabidopsis occurred after the speciation of legumes and crucifers. In H-caleosin, three copies (HF1, HF2, HF3, short for H-caleosin 1, 2, 3 of the Fabaceae) were formed by segmental duplication, while in L-caleosin, three copies (LF1, LF2, LF3, short for L-caleosin 1, 2, 3 of fabaceaes) (Fig. 3a) or two copies (LF1, LF2, short for L-caleosin 1, 2 of fabaceaes) (Fig. 3b) were formed by tandem duplication.

Table 2. Segmentally duplicated caleosin genes in soybean

Duplicated caleosin 1	Duplicated caleosin 2	flanking genes	Ks	Age(MYA)
GmaCLO10	*GmaCLO8*	14	0.61	58.99
GmaCLO9	*GmaCLO1*	18	0.12	11.61

Expression patterns of leguminous caleosin genes in different tissues

The expression patterns of legume caleosin genes were investigated by searching public resources. Corresponding ESTs were identified in three legumes. RNA-seq data were collected in soybean and common bean, while microarrays were obtained for barrel medic. The ESTs of common bean and barrel medic were limited from these resources. The data showed that every caleosin gene had at least one expression, except *MtrCLO4*, suggesting that these genes were all expressed except *MtrCLO4* (Table S2). It was further suggested that *MtrCLO4* may lose its function in the examined tissues.

The expression profiles from soybean ESTs and RNA-seq data were similar (Table S2). The expression of *GmaCLO2* was relatively high in the examined tissues. Meanwhile, the expressions of *GmaCLO4*, *GmaCLO6* and *GmaCLO7* were particularly low, close to zero. *GmaCLO1*, *GmaCLO2* and *GmaCLO9* displayed high expression in flowers (Fig. 4a).

The common bean ESTs of seeds were abundant and ESTs of other tissues were scarce; the RNA-seq data of seeds were absence (S2). Therefore, the data

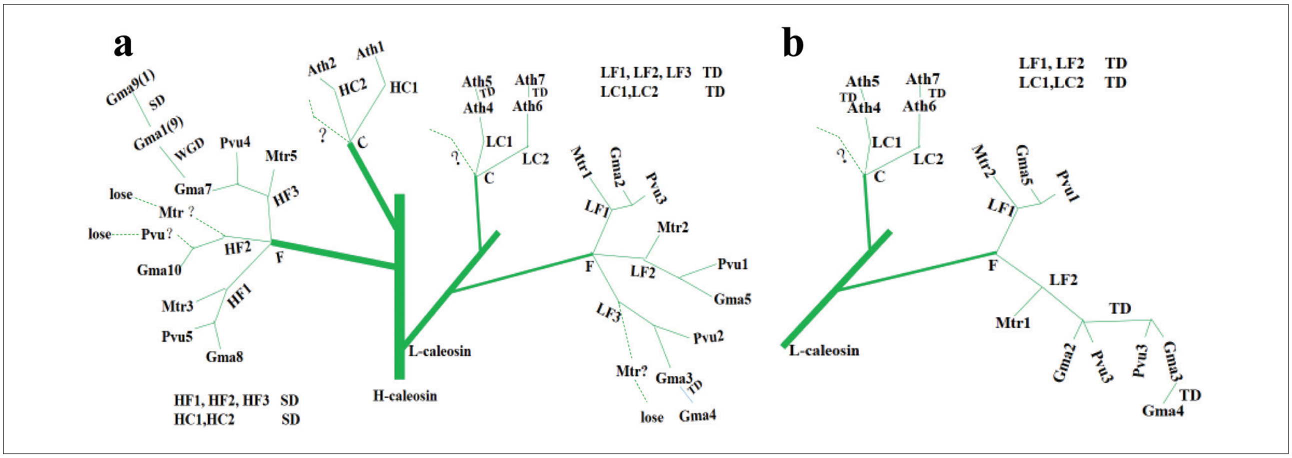

Fig. 3. A deduced evolutionary tree in caleosin genes in legumes and Arabidopsis. In L-caleosin, three copies (LF1, LF2, LF3, short for L-caleosin 1, 2, 3 of fabaceaes) (**a**) or two copies (LF1, LF2, short for L-caleosin 1, 2 of fabaceaes) (**b**) were formed by tandem duplication. In H-caleosin, three copies (HF1, HF2, HF3, short for H-caleosin 1, 2, 3 of fabaceaes) are formed by segmental duplication. In Fig. 3a and Fig. 3b, H-caleosin part is the same and H-caleosin part is omitted in Fig. 3b. Proven events are shown by solid line, speculated events by dotted line. L – L-caleosin; H – H-caleosin; TD – tandem duplication; SD – segmental duplication; WGD – whole genome duplication; C – crucifer; F – fabaceae=legume; Gma1-Gma10, GmaCLO1-GmaCLO10; Pvu1-Pvu5, PvuCLO1-PvuCLO5; Mtr1-Mtr5, MtrCLO1-MtrCLO5.

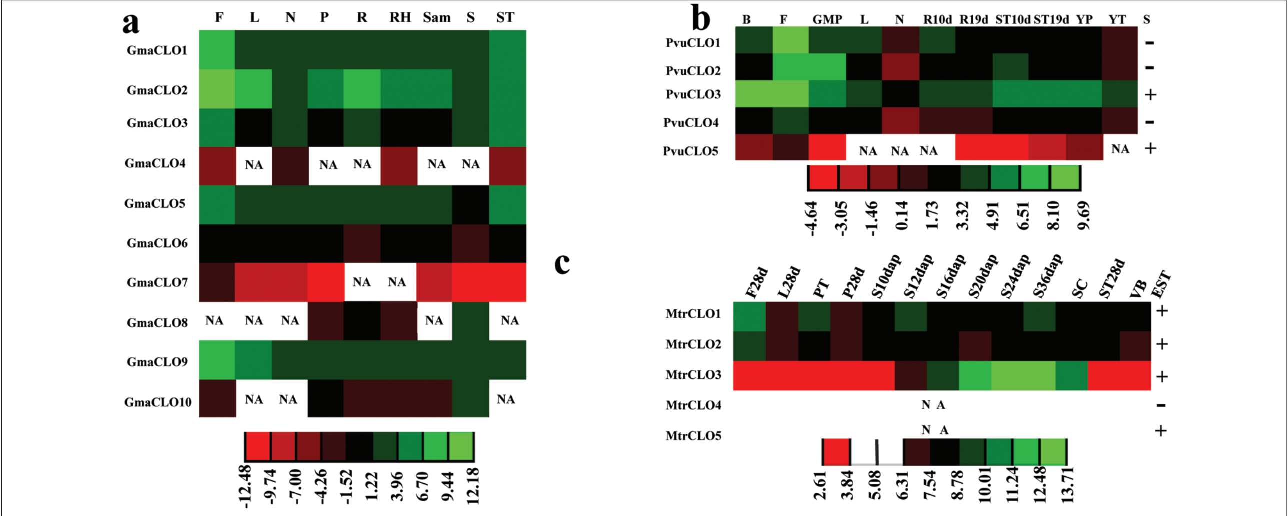

Fig. 4. Expression profiles of caleosin genes. Profiles in soybean (**a**), common bean (**b**) and barrel medic (**c**). The heat maps were generated by log2 transformed normalized expression values of caleosin genes. The color scale representing log2 transformed values is shown at the bottom. F – flower; L – leaf; N – nodules; P – pod; R – root; RH – root hairs; S – seed; ST – stem; B – flower bud; GMP – green and mature pods; YP – young pod; YT – young trifoliates; PT – petiole; SC – seed coat; VB – vegetative buds; d – day; dap – day after pollination; NA – not available; + – data exist; - – data do not exist.

from two sources were combined to survey the expression pattern in common bean. As shown in Fig. 4b, the expressions of *PvuCLO1*and *PvuCLO3* were relatively high in the examined tissues. *PvuCLO1* was specifically expressed in flowers.

In barrel medic, the expression patterns from microarray data and ESTs were similar (Table S2). The expression of *MtrCLO3* was relatively high and specifically expressed in seeds (Fig. 4c). Evidence for *MtrCLO5* expression from ESTs was very low (Table S2). Moreover, the data of ESTs also showed that *GmaCLO1*, *GmaCLO2*, *GmaCLO5*, *GmaCLO8* and *MtrCLO1* were expressed under stress conditions (Table S2). In addition, some caleosin genes were expressed in nodules. Whether caleosin genes play a role in nodules remains to be determined.

Cis-regulatory element analysis of important leguminous caleosin genes

According to the New PLACE analysis (Table S3), ABRERATCAL (S000507) was identified in *GmaCLO2* and *PvuCLO3*. DOFCOREZM (S000265) and RAV1AAT (S000314) were identified in *GmaCLO2*, *PvuCLO1*, *PvuCLO3* and *MtrCLO3*.

DISCUSSION

Caleosin genes and coding proteins in three legumes

In this study, ten caleosin genes are identified in the latest soybean genome *Glycine max Wm82.a2.v1*. Song et al. [11] showed that there are seven caleosin genes in soybean genome *Glycine max*. This may be because the latest soybean genome version (978Mb, 56,044 protein-coding loci) is more complete than the older version (975Mb, 54,175 protein-coding loci). Meanwhile, the 27 KDa and 29 KDa caleosin proteins have recently been identified in extracted soybean LDs [10], which does not match the molecular weights predicted in this study, which may be due to the post-translational modifications of caleosin proteins [31].

Compared with Arabidopsis [6], phosphorylation sites T^{136} and T^{181} (according to GmaCLO2) are unique in the L-caleosin of the legumes. These phosphorylation sites may play import roles in these species. In addition, analysis of protein motif configuration indicates that H-caleosins are more likely to lose motifs. Besides, all L-caleosin genes undergo tandem duplication, and two H-caleosin genes undergo segmental duplication, suggesting this also may be the result of different ways of expansion.

Expansion patterns of the caleosin gene family in three legumes

Segmental duplication, tandem duplication and transposition can form members of a gene family. In 20 leguminous caleosin genes, ten genes (50%) have evolved by tandem duplication and four (20%) by segmental duplication. Thus, tandem duplication is the main reason for caleosin expansion in the three legumes.

Evolution of caleosin genes in three legumes and Arabidopsis

In the H-caleosin type, the segmental duplicated pair, *GmaCLO8* and *GmaCLO10*, is speculated to have occurred ~58.99 MYA, which is in agreement with the emergence of papilionoids (59 MYA) [12]. The pair, *AthCLO1* and *AthCLO2*, might have evolved 26.33 MYA [6], which is consistent with the origin of crucifers (24-40 MYA) [32]. This suggests that the duplications of H-caleosin genes in legumes and Arabidopsis occurred just after the speciation of legumes and crucifers. The phylogeny trees show that each clade of H-caleosin contains genes from the three legumes, implying that the H-caleosin in legumes duplicated before the split between barrel medic and milletioids (including soybean and common bean) (54MYA) [14]. Another gene pair in H-caleosin genes, *GmaCLO1* and *GmaCLO9*, is due to a recent duplication ~11.61MYA after the whole genome duplication of soybean ~13MYA [12].

In the L-caleosin type, the central hydrophobic domain, PX3PSX3P, develops into PX3FSX3P in Arabidopsis [6] but is invariable in legumes. This indicates that the duplications of L-caleosin in legumes and Arabidopsis occurred after the speciation. The duplicated pair, *AthCLO4* and *AthCLO5*, is speculated to have occurred ~40.33MYA [6], which fits into the origin of crucifers (24-40 MYA) well [32], providing support of the above viewpoints. However, numbers of clades (two (Fig. 3b) or three (Fig.3a)) of the L-caleosin classification in legumes remain to be determined (Fig. 3). The duplicated pair, *GmaCLO1* and *GmaCLO9*, is speculated to have evolved ~11.61MYA, following the split of common bean and soybean ~19.2 [13].

Expression of caleosin genes in the three legumes

Motif analysis show that most H-caleosin genes with lower expression are truncated, suggesting that H-caleosin genes may lose or undermine their functions through loss of essential motifs or sequences. Moreover, the expressions of L-caleosin are much higher than H-caleosin in soybean and common bean, while the result is just the opposite in barrel medic. This may be due to the divergence of Hologalegina (including barrel medic) and milletioids (including soybean and common bean) 54 MYA [14].

In order to improve the fitness of plants, genes with higher expression abundances may be more important [33]. Thus, *GmaCLO2*, *PvuCLO1*, *PvuCLO3* and *MtrCLO3* merit further study, while *GmaCLO4*, *GmaCLO6*, *GmaCLO7*, *GmaCLO10*, *MtrCLO4* and *MtrCLO5* may be less important in the examined tissues (Fig. 4). Expressions of *GmaCLO4*, *GmaCLO6*, *GmaCLO7*, *GmaCLO10*, *MtrCLO4* and *MtrCLO5* are low (Fig. 2), and combining with motif analyses (Fig. 1, Fig. S1) suggests that *GmaCLO6*, *GmaCLO10* and *MtrCLO4* may have lost their function.

Potential functions of important leguminous caleosin genes

The motifs DOFCOREZM (S000265) and RAV1AAT (S000314), which bind DOF and RAV1 proteins, are found in *GmaCLO2*, *PvuCLO1*, *PvuCLO3*, and *MtrCLO3*. The RAV1 protein contains AP2-like and B3-like domains, which include transcription factors

WRI1, FUS3, ABI3 and LEC2 related to lipid metabolisms [34]. A study shows that the soybean DOF proteins are associated with lipid accumulation [35]. Thus, these caleosin proteins may play roles in lipid accumulation.

ABRERATCAL (S000507), which responds to calcium signaling [36], is also found in *GmaCLO2* and *PvuCLO3*. GmaCLO2 and PvuCLO3 have a calcium-binding domain, which suggests that these two caleosin proteins may be associated with calcium signal transduction.

GmaCLO2 displays expression under stress and specific expression in the flower. Therefore, GmaCLO2 may be involved in lipid accumulation in the flower, or calcium signal transduction in flower and the process under stress.

PvuCLO3 presents upregulated expression in flower and bud. Thus, PvuCLO3 may participate in calcium signal transduction and lipid accumulation in flower and bud.

PvuCLO1 and *MtrCLO3* show specific expression in flower and seed, respectively. Therefore, PvuCLO1 and MtrCLO3 may be related to lipid accumulation in flower and seed, respectively.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (31270295) and the starting funding of Northwest A&F University, China.

Authors' contributions: YS, ZWL, LLW performed the experiments. YS, MZL analyzed the data. QLJ, CZZ coordinated and helped to draft the manuscript. YS, ZXL and MZ designed the experiment and prepared the manuscript. All the authors have read and approved the final manuscript.

Conflict of interest disclosure: The authors declare no conflict of interest..

REFERENCES

1. USDA. Oilseeds: World Markets and Trade. 2014. Available: http://www.fas.usda.gov/data/oilseeds-world-markets-and-trade.
2. Chapman KD, Dyer JM, Mullen RT. Biogenesis and functions of lipid droplets in plants. Thematic Review Series: Lipid Droplet Synthesis and Metabolism: from Yeast to Man. J Lipid Res. 2012;53(2):215-26.
3. Chen JC, Tsai CC, Tzen JTC. Cloning and secondary structure analysis of caleosin, a unique calcium-binding protein in oil bodies of plant seeds. Plant Cell Physiol. 1999;40(10):1079-86.
4. Hernandez-Pinzon I, Patel K, Murphy DJ. The *Brassica napus* calcium-binding protein, caleosin, has distinct endoplasmic reticulum- and lipid body-associated isoforms. Plant Physiol Biochem. 2001;39(7):615-22.
5. Murphy DJ. The dynamic roles of intracellular lipid droplets: from archaea to mammals. Protoplasma. 2012;249(3):541-85.
6. Shen Y, Xie J, Liu R-d, Ni X-f, Wang X-h, Li Z-x, Zhang M. Genomic analysis and expression investigation of caleosin gene family in Arabidopsis. Biochem Biophys Res Commun. 2014;448(4):365-71.
7. Blée E, Boachon B, Burcklen M, Le Guédard M, Hanano A, Heintz D, Ehlting J, Herrfurth C, Feussner I, Bessoule J-J. The reductase activity of the Arabidopsis caleosin RESPONSIVE TO DESSICATION20 mediates gibberellin-dependent flowering time, abscisic acid sensitivity, and tolerance to oxidative stress. Plant Physiol. 2014;166(1):109-24.
8. Blée E, Flenet M, Boachon B, Fauconnier ML. A non-canonical caleosin from Arabidopsis efficiently epoxidizes physiological unsaturated fatty acids with complete stereoselectivity. FEBS J. 2012;279(20):3981-95.
9. Hanano A, Burcklen M, Flenet M, Ivancich A, Louwagie M, Garin J, Blée E. Plant seed peroxygenase is an original heme-oxygenase with an EF-hand calcium binding motif. J Biol Chem. 2006;281(44):33140-51.
10. Zhao L, Chen Y, Cao Y, Kong X, Hua Y. The integral and extrinsic bioactive proteins in the aqueous extracted soybean oil bodies. J Agr Food Chem. 2013;61(40):9727-33.
11. Song W, Qin Y, Zhu Y, Yin G, Wu N, Li Y, Hu Y. Delineation of plant caleosin residues critical for functional divergence, positive selection and coevolution. BMC Evol Biol. 2014;14(1):124.
12. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178-83.
13. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46(7):707-13.
14. Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. 2011;480(7378):520-4.
15. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(D1):D1202-10.
16. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N. Phytozome:

a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178-D1186.

17. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40.
18. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003;31(13):3784-8.
19. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294(5):1351-62.
20. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006;34(suppl 2):W369-W373.
21. Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8.
22. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-9.
23. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.
24. Ma H, Zhao J. Genome-wide identification, classification, and expression analysis of the arabinogalactan protein gene family in rice (*Oryza sativa L.*). J Exp Bot. 2010;61(10):2647-68.
25. Maher C, Stein L, Ware D. Evolution of Arabidopsis microRNA families through duplication events. Genome Res. 2006;16(4):510-9.
26. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(suppl 2):W609-W612.
27. He J, Benedito VA, Wang M, Murray JD, Zhao PX, Tang Y, Udvardi MK. The Medicago truncatula gene expression atlas web server. BMC Bioinformatics. 2009;10(1):441.
28. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
29. Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 1999;27(1):297-300.
30. Chen D-H, Chyan C-L, Jiang P-L, Chen C-S, Tzen JTC. The same oleosin isoforms are present in oil bodies of rice embryo and aleurone layer while caleosin exists only in those of the embryo. Plant Physiol Biochem. 2012;60:18-24.
31. Vermachova M, Purkrtova Z, Santrucek J, Jolivet P, Chardot T, Kodicek M. New protein isoforms identified within Arabidopsis thaliana seed oil bodies combining chymotrypsin/trypsin digestion and peptide fragmentation analysis. Proteomics. 2011;11(16):3430-4.
32. Blanc G, Hokamp K, Wolfe KH. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 2003;13(2):137-44.
33. Cheng F, Wu J, Wang X. Genome triplication drove the diversification of Brassica plants. Hortic Res. 2014;1:14024.
34. McGlew K, Shaw V, Zhang M, Kim RJ, Yang W, Shorrosh B, Suh MC, Ohlrogge J. An annotated database of Arabidopsis mutants of acyl lipid metabolism. Plant Cell Rep. 2015;34(4):519-32.
35. Wang HW, Zhang B, Hao YJ, Huang J, Tian AG, Liao Y, Zhang JS, Chen SY. The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic Arabidopsis plants. Plant J. 2007;52(4):716-29.
36. Kaplan B, Davydov O, Knight H, Galon Y, Knight MR, Fluhr R, Fromm H. Rapid transcriptome changes induced by cytosolic Ca^{2+} transients reveal ABRE-related sequences as Ca^{2+}-responsive cis elements in Arabidopsis. Plant Cell. 2006;18(10):2733-48.

Supplementary Files

Fig. S1. Multiple alignment of 18 caleosins in soybean, common bean and barrel medic. H-form insertion, calcium binding motif and proline-knot domain are boxed with red lines. Putative phosphorylation sites of H-caleosins are marked by a blue box. Putative phosphorylation sites of L-caleosins are marked by a green box. Putative phosphorylation sites of all caleosins are marked by a yellow box. Eleven fully conserved residues are shown by an upper star. The highly conserved histidine sites of haem-binding motif are indicated by red arrows.
Available on: http://serbiosoc.org.rs/sup/1/FigS1.pdf

Table S1. Motif sequences and annotations. Numbers refer to the motifs identified in Fig. 1. NH – no hits found.
Available on: http://serbiosoc.org.rs/sup/1/TableS1.xlsx

Table S2. Expression data of caleosin genes in different tissues using RNA-seq and ESTs in soybean and common bean, and microarray data and ESTs in barrel medic. EST libraries used to make the synthetic libraries in three legumes are also shown.
Available on: http://serbiosoc.org.rs/sup/1/TableS2.xlsx

Table S3. Potential *cis*-regulated elements of important legume caleosins identified by PLACE.
Available on: http://serbiosoc.org.rs/sup/1/TableS3.xlsx