Biochemical and Biophysical Research Communications 448 (2014) 365–371

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Genomic analysis and expression investigation of caleosin gene family in Arabidopsis

Yue Shen [a,b], Jun Xie [b], Rui-dan Liu [b], Xue-feng Ni [b], Xue-hao Wang [b], Zhi-xi Li [a,*], Meng Zhang [b,*]

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

ARTICLE INFO

Article history:
Received 16 April 2014
Available online 2 May 2014

Keywords:
Caleosin
Classification
Evolution
Expression
Function

ABSTRACT

Caleosin is a common lipid-droplet surface protein, which has the ability to bind calcium. Arabidopsis (*Arabidopsis thaliana*) is considered a model organism in plant researches. Although there are growing researches about caleosin in the past few years, a systemic analysis of caleosins in Arabidopsis is still scarce. In this study, a comprehensive investigation of caleosins in Arabidopsis was performed by bioinformatics methods. Firstly, eight caleosins in Arabidopsis are divided into two types, L-caleosin and H-caleosin, according to their molecular weights, and these two types of caleosin have many differences in characteristics. Secondly, phylogenetic tree result indicates that L-caleosin may evolve from H-caleosin. Thirdly, duplication pattern analysis shows that segmental and tandem duplication are main reasons for Arabidopsis caleosin expansion with the equal part. Fourthly, the expression profiles of caleosins are also investigated in silico in different organs and under various stresses and hormones. In addition, based on promoter analysis, caleosin may be involved in calcium signal transduction and lipid accumulation. Thus, the classification and expression analysis of caleosin genes in Arabidopsis provide facilities to the research of phylogeny and functions in this gene family.

1. Introduction

Plant oils are very important not only to supply essential nutrition for people, but also to provide energy and material for seed germination and seedling development. Triacylglycerols (TAGs) are the most abundant neutral lipids in the plant oils. Most of TAGs are partitioned in small spherical particles called lipid droplets (LDs, also known as oil bodies, lipid bodies, spherosomes, and oleosomes). LDs contain a hydrophobic core of TAGs, embraced by a monolayer of phospholipids, and unique proteins called oleosins, caleosins and steroleosins [1].

Caleosins have been discovered ubiquitously in plants and fungi [2]. As the structure of oleosin, caleosins consist of three domains: an N-terminal hydrophilic domain including EF-hand calcium-binding motif, a central hydrophobic domain containing proline knot for anchoring LDs, and a C-terminal hydrophilic domain with several phosphorylation sites [3].

Arabidopsis thaliana is considered a model organism in plant researches for its small genome (125 Mb) and short life cycle [4]. Eight caleosin genes have been identified in the Arabidopsis genome [5]. However, only limited analysis of their functions has been reported to date. At4g26740 has been found to be involved in post-germinative activity [6] and At5g55240 was believed to be associated with dormancy [7]. And more recently, At4g26740, At5g55240, At2g33380, At1g70670 are verified to be act as peroxygenases [5,8,9]. The term "caleosin" is given by their ability to bind calcium and similarity to oleosin in structure. Calcium signal is involved in various aspects of plant development. A study showed that caleosin expression was up-regulated in high yield and quality rapeseed [10]. Thus, whether caleosin has other functions, like regulation of oil content, still remains to be investigated. Potential functions of caleosin are proposed in this study.

The completion of *A. thaliana* genome sequence and development of various databases of gene expression provide a great opportunity to identify and analyze genes and predict gene function quickly by bioinformatics tools. *Cis*-acting elements are important switches taking part in the transcriptional regulation of activities of a gene in various biological processes. To the best of our knowledge, little attention has been devoted to the promoter elements analysis of caleosins.

In this work, a systemic analysis of eight caleosin genes in *A. thaliana* was performed. According to molecular weights (Mw), caleosin could be divided into two classes, H-caleosin and L-caleosin,

* Corresponding authors. Fax: +86 29 8708 2845 (M. Zhang).
E-mail addresses: shenyue1020@163.com (Y. Shen), xiejun1029@gmail.com (J. Xie), liurd1991@gmail.com (R.-d. Liu), nixuefeng123456@126.com (X.-f. Ni), xuehaowang1226@gmail.com (X.-h. Wang), lizhxi@nwsuaf.edu.cn (Z.-x. Li), zhangm@nwsuaf.edu.cn (M. Zhang).

http://dx.doi.org/10.1016/j.bbrc.2014.04.115
0006-291X/

for the first time. This classification was supported by the results of motif analysis, multiple alignment and phylogenetic tree. The properties, evolution and expression profiles of caleosins in Arabidopsis were also investigated. Moreover, potential functions of caleosin were speculated based on promoter analysis. Further studies on verifying the functions of caleosin in experiment will be reported in our next study.

2. Materials and methods

2.1. Identification and characteristics of caleosin genes in Arabidopsis

In order to survey whether there were other caleosin genes, exact name and HMM (Hidden Markov Model) searches were performed to search the putative caleosins in *A. thaliana* genome. Exact name searches were executed by using "caleosin" as a query in the Arabidopsis Information Resource (TAIR10.0, http://www.arabi-dopsis.org/) [11]. The raw HMM of caleosin domain (PF05042) was obtained from Pfam 26.0 (http://pfam.sanger.ac.uk/) [12] and as a query to retrieve the protein database in TAIR using "hmmsearch" in local HMMER 3.0 software package [13], with *E*-value $<10^{-10}$. After that, the gene sequences of predicted caleosin were collected and the redundant sequences were removed manually. All the deduced amino acid sequences of the putative caleosins were submitted to the InterProScan 4 (http://www.ebi.ac.uk/Tools/pfa/iprscan/) [14] to verify the presence of caleosin domain (IPR007736).

Physical and chemical properties of caleosins, Mw, isoelectric points (p*I*), and amino acid lengths, were performed on ExPASy (http://web.expasy.org/compute_pi/) [15]. The hydropathic plot was drawed in ProtScale (http://web.expasy.org/protscale/) [16] with Kyte and Doolittle method and the default setting. Ser, Thr and Tyr phosphorylation sites in proteins were retrieved by program NetPhos 2.0 (http://www.cbs.dtu.dk/services/NetPhos/) [17]. Protein stability changes for single-site mutations were predicted by MUpro (http://www.ics.uci.edu/~baldig/mutation.html) [18].

To investigate whether the caleosin protein owns a conserved domain other than "caleosin", the sequences were analyzed by MEME (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) [19] with 5 different motifs. The annotations of the identified motifs were carried out by InterProScan 4 (http://www.ebi.ac.uk/Tools/InterProScan/) [14].

2.2. Multiple alignment and phylogenetic tree construction

Multiple sequence alignment of the complete protein sequences of eight Arabidopsis caleosins and one cycad (*Cycas revoluta*) caleosin was done by software ClustalX 2.1 [20]. The cycad caleosin sequence was extract from GenBank with accession number FJ455154. Pretty output and shading of the alignment results were operated using online software Boxshade 3.21 in ExPASy (http://www.ch.embnet.org/software/BOX_form.html) [15].

Then, in order to build a neighbor-joining (NJ) tree, the alignment was submitted to MEGA 6.06 [21] using NJ method, Poisson model and pairwise deletion of gaps. The phylogeny was tested by bootstrap replication 1000 and the branch length count on phylogenetic distances.

To construct a maximum-likelihood (ML) tree, the alignment was employed to PhyML 3.0 on website (http://www.atgc-montpellier.fr/phyml/) [22] using LG model, 6 for substitution rate categories and SPR for tree improvement. The reliability of the tree was assessed with bootstrap replication 500.

The visualization of the two phylogenetic trees was both generated using MEGA 6.06.

2.3. Chromosome distribution and gene expansion analysis

Genes were mapped on five Arabidopsis chromosomes by Chromosome map tool on TAIR, using locus names.

The caleosin genes separated by a maximum of five genes were defined as tandem duplication [23]. Searching for segmental duplications were performed according to the study before [24], and also gained by searching locus name against the PGDD (plant genome duplication database, http://chibba.agtec.uga.edu/duplication/) [25].

Ka and Ks values were extracted from PGDD. The Ks values were used to estimate the date of the duplication occurrence following the formula $T = Ks/2\lambda$. For Arabidopsis, λ is 1.5×10^{-8} [26].

2.4. Upstream sequence element analysis

The 2000 bp of eight caleosins genomic sequences upstream of the start condon were obtained from Arabidopsis genome in TAIR. Putative transcription factor binding sites were predicted by using web-based program TFSEARCH 1.3 (http://www.cbrc.jp/research/db/TFSEARCH.html) in plant [27]. Potential *cis*-acting regulatory elements were searched by PLACE 30.0 (http://www.dna.affrc.go.jp/PLACE/) [28].

2.5. Expression patterns of caleosin genes

Expression patterns of caleosin genes were performed by analyzing ESTs (expressed sequence tags), full-length cDNAs, microarray data and MPSS (massively parallel signature sequencing) tags in the public databases.

ESTs and cDNAs were obtained from DFCI database (http://compbio.dfci.harvard.edu/tgi/plant.html) [29]. Caleosin gene sequences were used to search Arabidopsis DFCI database, using BLASTN, with *E*-value $<10^{-10}$, identity >95% and length >200 bp. The most matched one was selected when a gene corresponds to multiple TCs. To analysis gene expression in different tissues, all the libraries were classified into eight synthetic libraries (Table S2). Then, for each gene, the number of corresponding ESTs in each synthetic library was counted and was normalized to TPM (Transcripts per million).

MPSS tags of caleosin genes were acquired from Arabidopsis MPSS Database (http://mpss.udel.edu/at/mpss_index.php) with normalized signature MPSS (TPM) of 17-base signatures.

Microarrays analysis and visualization were done through AtGenExpression Visualization Tool (AVT, http://jsp.weigelworld.org/expviz/expviz.jsp) [30,31], using locus search with mean-normalized value.

The heatmap was generated for displaying expression by R script [32] on website http://www.hiv.lanl.gov/content/sequence/HEATMAP/heatmap.html.

3. Result

3.1. Identification, classification and characteristics of caleosins in Arabidopsis

Eight caleosins in Arabidopsis were identified through exact name search in TAIR and HMM search in local, which is consistent with the reports before [5]. Since only names AtCLO1 to AtCLO5 were given in the literatures, additional caleosin genes were named from AtCLO6 to AtCLO8 (Table 1). The size of ORF and deduced amino acid, calculated Mw and p*I* of eight caleosin genes varied in a wide range (Table 1). According to Mw, eight caleosins were classified into two groups, high-Mw and low-Mw caleosin, named as H-caleosin and L-caleosin.