Journal of Animal & Plant Sciences, 32(5): 2022, Page: 1394-1402
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2022.5.0546

IN SILICO STRUCTURAL, EVOLUTIONARY, AND EXPRESSION ANALYSIS OF SMALL HEAT SHOCK PROTEIN (SHSP) ENCODING GENES IN COCOA (*THEOBROMA CACAO* L.)

P. B. Cao

[1]Department of Biology, Faculty of Natural Sciences, Hung Vuong University, Nong Trang Wards, Viet Tri City, Phu Tho Province, Vietnam
*Corresponding author's Email: phibang.cao@hvu.edu.vn

ABSTRACT

The small heat shock proteins (sHSP) are the molecular chaperones that play important roles in the development and stress responses of plants. However, the sHSP family has not been well investigated in cocoa (*Theobroma cacao* L.). In this work, a total of 32 full-length genes encoding putative sHSP proteins were predicted in the cocoa genome. Predicted proteins were ranging from 130 to 269 amino acids. Most of them were intronless or single intron. The cocoa sHSPs were divided into 12 subclasses including seven cytoplasmic (CI–CVII) and five organelles localized subclasses. The 32 cocoa *sHSP* (*TcsHSP)* genes were randomly distributed in the entire cocoa genome but chromosome 9 appeared as the hot spot region for *TcsHSPs*. Tandem duplication events made a contribution to the expansion of *sHSP* genes in cocoa. The Ka/Ks values from three CII *sHSP* genes were ranging from 0.12 to 0.24. Differently, most of the Ka/Ks ratios from 11 CI *sHSP* genes were greater than 1.0, suggesting the driving change in this group. *TcsHSPs* were found expressed in all developmental stages with different profiles. This work contributes to providing valuable information on the evolutionary relationship of the *sHSP* gene family in cocoa which is useful for future investigation of the functional characteristics of *TcsHSP* genes.

Keywords: cocoa (*Theobroma cacao* L.), gene family, *in silico*, small heat shock proteins (sHSP)

Published first online April 30, 2022 Published final October 05, 2022

INTRODUCTION

Heat shock protein (HSP) was firstly found in a common fruit fly (*Drosophila melanogaster*) by Ritossa (Italia) in 1960 and was called "Heat shock protein" in the report of Tissieres *et al.* (1974). HSPs function as chaperone molecules. They play an important role in folding, translocating, and degrading protein in the normal cellular process as well as stabilizing protein molecules and biological membranes under stress conditions. The major HSPs are classified into five conserved families named HSP100, HSP90, HSP70, HSP60, and sHSP/HSP20, respectively, based on their approximate molecular weight (Krishna, 2004). The small heat shock proteins (sHSPs) belong to an ancient and omnipresent family of proteins recorded in all taxonomical domains (Waters, 2013). These proteins are molecular chaperones like most other HSPs (Waters, 2013). By binding to denatured proteins, the sHSPs can prevent their irreversible aggregation (Nakamoto and Vigh, 2007). The sHSPs have a major role in both the plant stress response and normal development (Waters, 2013; Waters *et al.*, 2008). Plant *sHSPs* are expressed under various types of stress including thermal stress. Most of the *sHSPs* were strongly upregulated during heat stress. The expression of some of the plant *sHSPs* is stimulated in response to stress such as heavy metals, drought, UV, salinity, cold, osmotic, and oxidative stress. Additionally, some plant sHSPs play a part in normal growth and development processes like embryogenesis, seed germination, pollen development, and fruit maturation (Waters, 2013; Waters *et al.*, 2008). Recently, *sHSPs* were considered to involving to plant immunity (Park and Seo, 2015).

All members of the sHSP family share a conserved domain of ~90 amino acid residues found near the C terminus of the protein and known as the α-crystallin domain (ACD) (de Jong *et al.*, 1998). This conserved domain is flanked by an N- and a C-terminal region (de Jong *et al.*, 1998). The ACD domain includes a compact β-sheet sandwich structure, b2-strand to the b9-strand. These sheets can dimerize to form large oligomers as well as involve substrate interaction (Jaya *et al.*, 2009). The N-terminus takes part in the binding to denatured proteins, whereas the C-terminus is associated with both homooligomerization and the formation of heat stress granules (Jaya *et al.*, 2009). The amino acid sequences of both N- and C-terminal regions are highly variable leading to sHsp molecular sizes ranging from approximately 15–30 kDa (de Jong *et al.*, 1998; Waters, 2013; Waters *et al.*, 2008). In angiosperms, the sHSPs are classified into at least 11 classical subclasses, cytosolic I, II, III, IV, V, VI, ER, PX, CP, MTI, and MTII (Waters, 2013; Waters *et al.*, 2008). Compared to other HSP

families, sHSP family size is the largest and highest divergent. Gene duplication has been important in gene expansion and in generating functional diversity within the *sHSP* family (de Jong *et al.*, 1998; Waters, 2013; Waters *et al.*, 2008). The whole-genome analysis allowed to identify 19, 36, and 23 *sHSP* genes including orphan *sHSP* in *Arabidopsis thaliana*, poplar (*Populus tricocarpa*), and rice (*Oryza sativa*) respectively (Waters *et al.*, 2008). Recent genome analysis permitted to discover 35 sHSP genes in pepper (*Capsicum annuum* L.) (Guo *et al.*, 2015), 27 in Chinese cabbage (*Brassica rapa* ssp *pekinensis*) (Tao *et al.*, 2015), 48 in grape (*Vitis vinifera*) (Ji *et al.*, 2019), and 47 in *Sorghum bicolor* L. (Nagaraju *et al.*, 2020). However, to our knowledge, genome-wide analysis of the sHSP family in cocoa, a tropical tree, is still lacking.

Theobroma cacao L. (2n = 2x = 20), a neotropical species, was native to Amazonian lowland rainforests and was domesticated over 1,500 years ago (Motamayor *et al.*, 2002). To date, this evergreen tropical tree has been grown in more than 50 countries in the world. Cocoa beans are used for the chocolate, confectionery, and cosmetic industries (Figueira *et al.*, 2005). Recently, the medicinal benefits of cocoa were reported (Pucciarelli, 2013). Cocoa is essential to the livelihoods of 40-50 million people worldwide (World Cocoa Foundation, 2012). Furthermore, cocoa plantation has provided sustainable economic and environmental benefits to some of the poorest and most ecologically sensitive areas of the world because this tree is often grown in agroforestry-type ecosystems alongside other fruit and commodity crops (Guiltinan *et al.*, 2008). In the research domain, cocoa was considered as an experimental organism having several limitations (Figueira *et al.*, 2005), however, its genome was a good resource allowing to accelerate progress in cocoa breeding and plantation as well as to support the understanding of its biochemistry (Motamayor *et al.*, 2013).

Thanks to the Cacao Genome Project, the availability of the cocoa genome (Motamayor *et al.*, 2013) permitted us to discover the *sHSP* family of this tropical plant by using *in silico* methods. In this study, we identified and analyzed 32 classical *sHSP* genes from the cocoa genome, focusing on their structure, distribution, classification, and evolution. These results provide a foundation for further functional analysis of the cocoa *sHSP* family.

MATERIALS AND METHODS

Identification of *TcsHSP* in cocoa genome: To identify the *sHsp* family members in the cocoa genome, a basic local alignment search tool (BLASTP) was performed against the cocoa proteome database (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Tcacao) using known *Arabidopsis* sHSP (Siddique *et al.*, 2008; Waters *et al.*, 2008) as querries. All output putative cocoa sHSP were verified by using Pfam (http://pfam.xfam.org/) to confirm the presence of the ACD/HSP20 domain. The predicted genes lacking an ACD domain and/or the genes whose molecular weight out of the range from 15 to 30 kDa were rejected. Then, the selected cocoa genes were used for a second TBLASTN round on the cocoa genome. This additional step permitted identifying the cocoa paralogs that had been excluded by their dissimilarity to the *Arabidopsis* orthologs or not annotated.

Sequence analysis and Chromosomal Localization: The predicted protein sequences were analyzed to obtain the number of amino acids, molecular weight, theoretical isoelectric point (pI), and GRAVY index by using EXPASY Protoparam (http://www.expasy.org/tools/protparam.html) (Gasteiger *et al.*, 2005). The chromosomal location data and intron numbers of putative *TcsHSP* were retrieved from the cocoa genome database. Graphical presentation of gene localization was performed using MapChart V2.1 (Voorrips, 2002) based on their chromosomal position and the relative distance between these genes on the same chromosome. Subcellular locations of deduced amino acids were predicted using the ProtCompv.9.0program(http://linux1.softberry.com/berry.phtml?topic=protcomppl&group=programs&subgroup=proloc) and TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson *et al.*, 2007).

Phylogenetic analysis of cocoa sHSPs: The deduced full-length amino acid sequences of sHSP members from cocoa, *Arabidopsis*, poplar, grape and rice were aligned by the MAFFT program (Katoh and Standley, 2013). The phylogenetic tree was conducted using MEGA X software (Kumar *et al.*, 2018) with the bootstrap test replicated 1000 times, and a JTT model.

Gene duplication and evolutionary analysis: The *TcsHSP* gene duplication was defined according to the previous description (Guo *et al.*, 2015). Ka (nonsynonymous substitutions per nonsynonymous site) and Ks (synonymous substitutions per synonymous site) values were calculated which furnishes the mean number Ks of synonymous substitutions per synonymous site between pairs of duplicated genes by using MegaX (Kumar *et al.*, 2018)

***In silico* gene expression analysis:** *In silico* expression profiles of *TcsHSP* genes were analyzed at different developmental stages of zygotic and somatic embryo based on Gene Expression Omnibus dataset GSE55476 (Maximova *et al.*, 2014). Relative expression values of *TcsHSP* genes were estimated by calculating the ratio of their expression level per that of the *Acyl carrier protein*

B (ACP B) gene which was the most stable expressed gene in various cocoa tissues (Pinheiro *et al.*, 2011).

RESULTS AND DISCUSSION

Identification and characteristics of *sHSP* genes in cocoa: A total of 43 full-length sequences containing the HSP20/alpha-crystallin domain (PF00011) were found. Two of them were manual annotated (TcsHSP24 and TcsHSP25). Sequences whose molecular weights out of the range from 15 to 30 kDa were rejected. Finally, a total of 32 full-length genes encoding putative sHSP proteins were validated by joint phylogenetic analysis with sHSP proteins of *A. thaliana*, rice, poplar, and grape.

Table 1. *Theobroma cacao* small heat shock proteins

Gene name	Locus name	Sub-class	Genomic (bp)	Peptide length (aa)	Peptide MW (Da)	pI	GRAVY	Introns number	Sub-cellular Loc
TcsHSP01	Thecc1EG001120	CVI	459	152	17.12	4.68	-0.135	0	C/N
TcsHSP02	Thecc1EG006853	CP	1705	214	24.10	8.34	-0.428	2	CP
TcsHSP03	Thecc1EG010703	CVII	811	232	26.36	8.31	-0.636	1	C
TcsHSP04	Thecc1EG014924	MTI	736	211	23.89	5.56	-0.69	1	M
TcsHSP05	Thecc1EG015499	CP	1137	236	26.30	7.66	-0.639	2	CP
TcsHSP06	Thecc1EG015982	MTI	1259	214	24.27	5.98	-0.584	1	M
TcsHSP07	Thecc1EG017043	CI	417	138	15.89	6.33	-0.68	0	C/N
TcsHSP08	Thecc1EG017468	PXI	516	141	15.80	6.92	-0.489	1	Px
TcsHSP09	Thecc1EG021809	CI	435	144	16.59	6.19	-0.681	0	C/M
TcsHSP10	Thecc1EG021810	CI	465	154	17.44	7.84	-0.618	0	C/N
TcsHSP11	Thecc1EG027146	ER	519	172	19.49	5.61	-0.642	0	C
TcsHSP12	Thecc1EG027161	MTII	827	235	26.91	6.93	-0.781	1	M
TcsHSP13	Thecc1EG029219	CIII	549	152	17.14	8.53	-0.561	1	C
TcsHSP14	Thecc1EG030200	CIV	501	138	15.82	4.93	-0.276	1	C
TcsHSP15	Thecc1EG031122	CV	922	194	22.32	5.05	-0.551	1	C
TcsHSP16	Thecc1EG036433	ER	576	191	21.73	6.46	-0.441	0	C/N
TcsHSP17	Thecc1EG037345	CII	471	156	17.57	5.79	-0.428	0	C
TcsHSP18	Thecc1EG037346	CII	474	157	17.81	6.17	-0.661	0	C
TcsHSP19	Thecc1EG037347	CII	480	159	17.94	5.6	-0.534	0	C
TcsHSP20	Thecc1EG041713	CI	453	150	17.29	6.18	-0.739	0	C
TcsHSP21	Thecc1EG041714	CI	471	156	17.97	5.7	-0.706	0	C
TcsHSP22	Thecc1EG041717	CI	471	156	17.81	5.97	-0.601	0	C
TcsHSP23	Thecc1EG041718	CI	474	157	18.07	5.98	-0.658	0	C
TcsHSP24	Nd	CI	471	140	16.11	8.96	-0.509	1	C
TcsHSP25	Nd	CI	595	130	15.30	8.73	-0.92	1	C/N
TcsHSP26	Thecc1EG041722	CI	468	155	17.97	5.97	-0.669	0	C/N
TcsHSP27	Thecc1EG041723	CI	477	158	18.22	5.98	-0.692	0	C
TcsHSP28	Thecc1EG041724	CI	591	140	16.30	6.53	-0.754	1	C/N
TcsHSP29	Thecc1EG041856	CI	480	159	17.98	5.57	-0.689	0	C
TcsHSP30	Thecc1EG041857	CI	462	149	16.81	5.55	-0.625	1	C
TcsHSP31	Thecc1EG041941	CI	477	158	18.35	5.98	-0.677	0	C
TcsHSP32	Thecc1EG042632	CI	483	160	18.25	6.34	-0.639	0	C/N

C*:* Cytoplasmic; CP: Chloroplast; MT: Mitochondria; N: Nuclear; Px: Peroxisome

In the literature, 23, 19, 36, and 48 sHSP were found in rice, *A. thaliana,* poplar, and grape respectively (Waters *et al.*, 2008). Regarding this gene family in the recent genome version in three angiosperm plants by eliminating the unexistance any more, we found only 23, 19, 29 and 42 classical sHSP in rice, *A. thaliana*, poplar, and grape, respectively. As a result, *sHSP* family in *T. cacao* contained more members than in these three angiosperm plant models as well as in other plants, including Chinese cabbage (27 genes) (Tao *et al.*, 2015) but fewer than in *S. bicolor* (47 genes) (Nagaraju *et al.*, 2020) and grape (42 genes) (Ji *et al.*, 2019). The abundance of cocoa *sHSP* genes may be involved in their diverse role in tropical woody plant, which is constantly exposed to high temperatures.

Details of the physiological and biochemical properties of all the cocoa *sHSP* genes were analyzed (Table 1). The genomic sequence lengths of *TcsHSP* genes were quite variable, ranging from 417 to 1705 bp. A comparison of the full-length cDNA sequences with the genomic DNA allowed the determination of the exons and introns content of each *sHSP* gene (Fig. 1). The data

showed that the 18 out of 32 *TcsHSP* genes were intronless. Twelve members contained a single intron. However, only two genes, *TcsHSP02* and *TcsHSP05* (chloroplast subclass) possessed two introns, respectively. The cocoa *sHSP* genes encoded the proteins ranging from 130 to 236 amino acids. Bioinformatic analyses showed that the molecular weight average of the cocoa predicted sHSP proteins were between 15.3 and 26.91 kDa (Table 1). These deduced proteins had theoretical pI ranging in value from 4.68 to 8.96. Based on theoretical pI value, 32 sHSP proteins could be divided into two subgroups. Twenty-three sHSP protein sequences were rich in acidic amino acids while nine possessed more alkaline amino acids. sHSPs sequence showed a wide range of GRAVY values that were all hydrophilic (ranging from -0.92 to -0.135). Subcellular localization of proteins could supply important information about their function (Emanuelsson *et al.*, 2007). The ProtComp and TargetP software were used to investigate the subcellular localization of TcsHSP proteins. The results showed that 32 sHSPs are predicted to be ubiquitously distributed across cellular compartments and organelles (Table 1). Briefly, 25 of them are predicted for the cytoplasm (CI-CVI), two for the endoplasmic reticulum (ER), two for the chloroplasts (CP), three for the mitochondria (MT), and one for the peroxisomes (PX). Similar results have been observed in *Arabidopsis*, rice (Waters *et al.*, 2008), and grape (Ji *et al.*, 2019), cytoplasm and these organelles were described as targeted location of sHSP proteins.

Classification of the TcsHSPs: To evaluate the evolutionary relationships of the cocoa sHSPs, phylogenetic analysis was further conducted based on multiple sequence alignment of full-length sequences of TcsHSP proteins together with rice, *Arabidopsis,* poplar, and grape orthologs. The accordance with classification in other plants, the phylogenetic analysis allowed classifying all TcsHSPs into 12 classical subclasses including seven subclasses that are cytoplasmic localized (CI–CVII) and five sHSP subclasses that localize to organelles. The results of phylogenetic analysis confirmed the subcellular localization prediction. The organelle subclasses consisted of two endoplasmic reticulum genes (*TcsHSP11* and *TcsHSP16*), one peroxisome gene (*TcsHSP08*), two chloroplast genes (*TcsHSP02* and *TcsHSP05*), and three mitochondria genes (*TcsHSP04* and *TcsHSP06* belonging to MTI subclass, and *TcsHSP12* of MTII subclass). The classification of cocoa sHSP is similar to *Arabidopsis*, poplar, and grape but different from rice. In detail, the CI-CVII subclasses were reported in these four species, except CVI and CVII subclasses which were not detected in the rice (monocot plant) (Waters *et al.*, 2008). In addition, four new nuclear subgroups (CVIII, CIX, CX, and CXI) were reported in the genome of rice (Sarkar *et al.*, 2009). However, these four subgroups were not identified in the cocoa genome (Table 2). Most of the TcsHSPs, including 25 out of 32, were classified into CI–CVII, which suggested that the cytosol might be the primary functional site of the cocoa sHSPs. This result confirmed the reported situation in grape (Ji *et al.*, 2019).

Chromosomal localization and gene duplication: The 32 *TcsHSP* genes were randomly distributed in the entire cocoa genome (Figure 2). Indeed, fourteen genes (45.2%) were located on chromosome 9. On this chromosome, 12 out of 14 genes condensed into two blocks, three genes (*TcsHSP17, TcsHSP18* and *TcsHSP19*) in position from 39218847 to 39227025 and ten (*TcsHSP20, TcsHSP21, TcsHSP22, TcsHSP23, TcsHSP24, TcsHSP25, TcsHSP26, TcsHSP27, TcsHSP28, TcsHSP29*) in position from 39218847 to 39256026. Two other genes (*TcsHSP30* and *TcsHSP31*) were sparsely located. Seventeen genes were distributed on remaining chromosomes, four genes on chromosome 6, three genes on chromosome 3, two genes on each of chromosome 2, 4 and 5, and only one gene on each of chromosome 1, 7, 8, and 10.

The duplications were analyzed to investigate the expansion and evolutionary change of the *TcsHSP* gene family. The *TcsHSP* gene duplication was considered based on the described criteria: (1) the length of the sequence alignment covered ≥70% of the longer gene; and (2) the similarity of the aligned gene regions ≥70% and (3) only one event of duplication is counted for tightly linked genes. A block of duplication was defined if more than one gene was involved in the duplication (Yang *et al.*, 2008). Tandem duplicated (TD) genes were defined as those separated by five or fewer genes in a 100-kb region (Wang *et al.*, 2010); genes were considered as segmental duplication (SD) if blocks of DNA that map to different loci in the same chromosome and whole-genome duplication (WGD) meant that the paralogs were located on different chromosomes (Li *et al.*, 2015).

According to the above definition, two blocks of duplication were found, a CII block including three genes (*TcsHSP17, TcsHSP18,* and *TcsHSP19*) and a CI block including 11 genes (*TcsHSP20 to TcsHSP30*) located in chromosome 9. Each of these blocks underwent tandem duplication events during cocoa chromosome evolution. These results suggested that tandem duplication events made a contribution to cocoa sHSP genes expansion. This observation is in agreement with that in grape, tandem duplication had been reported to have shaped the expansion of the sHSP family in this species (Ji *et al.*, 2019).

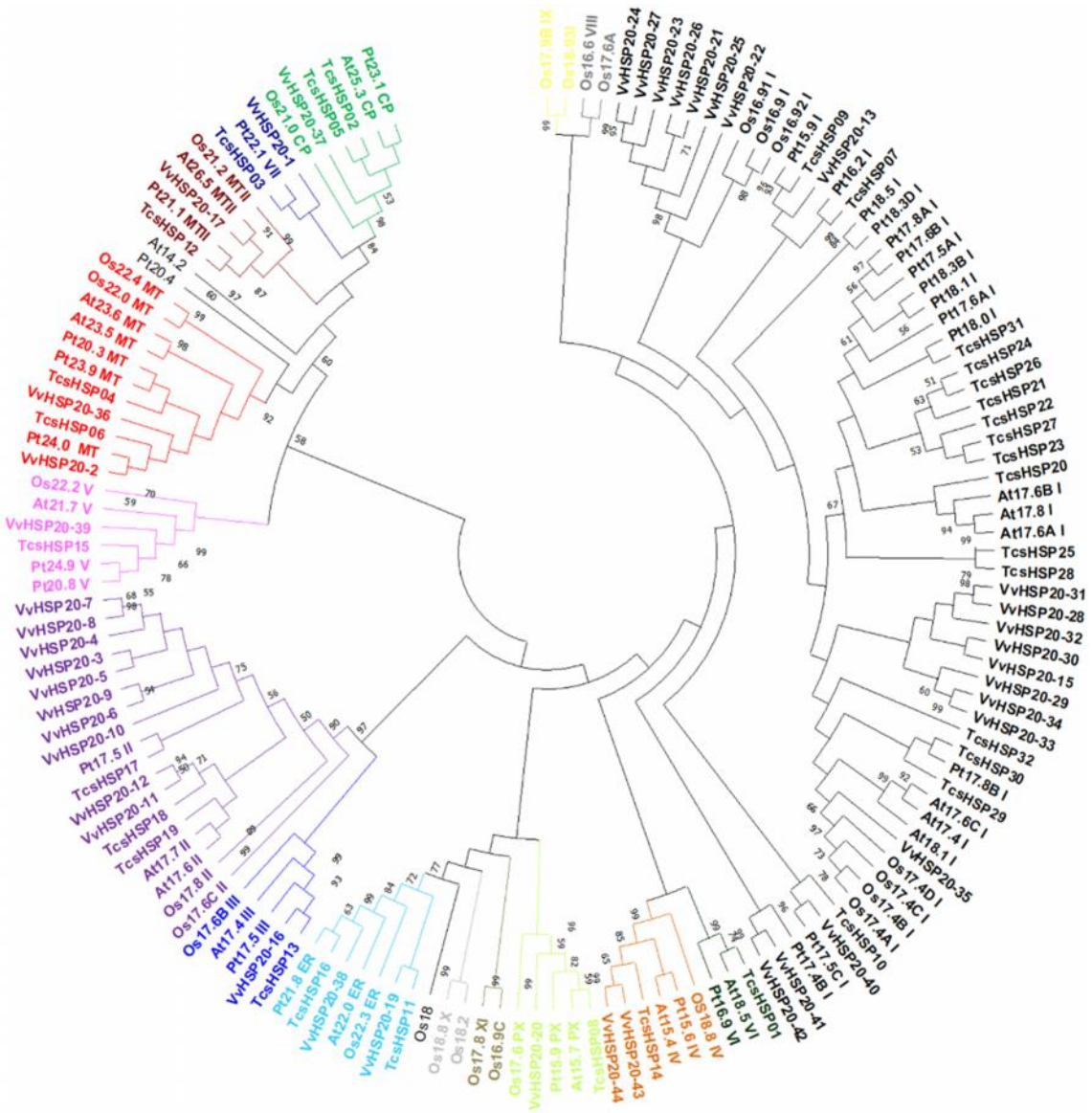

Figure 1. The evolutionary history was inferred using the Neighbor-Joining method. The evolutionary distances were computed using the JTT matrix-based method and were in the units of the number of amino acid substitutions per site. This analysis involved 149 amino acid sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 610 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar *et al.*, 2018). P: green; MTI: red, MTII: maroon, CI: black, CII: purple, CIII: blue, CIV: orange, CV: pink, CVI: teal, CVII: navy blue, CVIII: gray, CIX: golden, CX: silver, CXI: olive, PX: lime, ER: cyan.

Table 2. Distribution of genes within *sHSP* family from *O. sativa (Os)*, *A. thaliana (At)*, *P. tricocarpa (Pt)*, *V. Vinifera* (Vv), and *T. cacao (Tc)*

	CI	CII	CIII	CIV	CV	CVI	CVII	CVIII	CIX	CX	CXI	MTI	MTII	PX	P	ER
Os	33.33	8.33	4.17	4.17	4.17	0.00	0.00	4.17	4.17	4.17	4.17	8.33	4.17	4.17	4.17	8.33
At	31.58	10.53	5.26	5.26	5.26	5.26	5.26	0.00	0.00	0.00	0.00	10.53	5.26	5.26	5.26	5.26
Pt	41.38	3.45	6.90	3.45	6.90	6.90	3.45	0.00	0.00	0.00	0.00	10.34	3.45	3.45	6.90	3.45
Tc	**50.00**	9.38	3.13	3.13	3.13	3.13	3.13	0.00	0.00	0.00	0.00	6.25	3.13	3.13	6.25	6.25
Vv	37.50	20.83	2.08	0.00	4.17	2.08	4.17	0.00	0.00	0.00	0.00	6.25	2.08	6.25	4.17	4.17

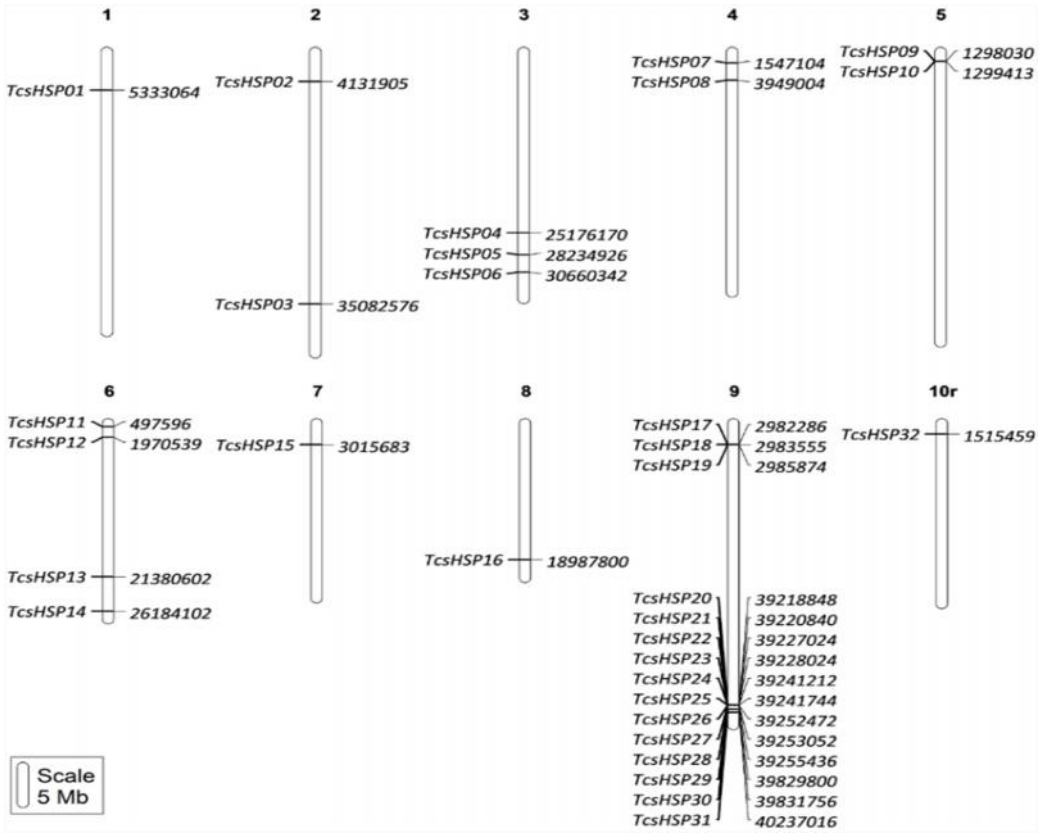

Figure 2. Localization of the identified *sHSP* genes on the cocoa chromosomes by using MapChart v.2.2. The chromosome number is shown at the top of each chromosome. The chromosome size, chromosomal position and the relative distance between these genes on the same chromosome were retrieved from Argout *et al.* (2011)

Table 3. Synonymous (Ks) and Nonsynonymous (Ka) substitutions for the CI and CII duplicated genes. Analyses were conducted using the Pamilo-Bianchi-Li model (Pamilo and Bianchi, 1993). Evolutionary analyses were conducted in MEGAX (Kumar *et al.*, 2018)..

Duplicated gene 1	Duplicated gene 2	Ka (dN)	Ks (dS)	Ka/Ks	Selection	Duplication
TcsHSP17	TcsHSP18	0.279	1.256	0.222	Purifying	TD
TcsHSP18	TcsHSP19	0.019	0.030	0.636	Purifying	TD
TcsHSP20	TcsHSP21	0.250	0.094	2.662	Positive	TD
TcsHSP21	TcsHSP22	0.166	0.109	1.520	Positive	TD
TcsHSP22	TcsHSP23	0.155	0.090	1.723	Positive	TD
TcsHSP23	TcsHSP24	0.229	0.057	4.025	Positive	TD
TcsHSP24	TcsHSP25	0.371	0.170	2.180	Positive	TD
TcsHSP25	TcsHSP26	0.291	0.156	1.862	Positive	TD
TcsHSP26	TcsHSP27	0.185	0.081	2.266	Positive	TD
TcsHSP27	TcsHSP28	0.071	0.094	0.747	Purifying	TD
TcsHSP28	TcsHSP29	0.391	0.293	1.332	Positive	TD
TcsHSP29	TcsHSP30	0.218	0.152	1.439	Positive	TD

To estimate the sequence evolution between the sequences of each of two duplication blocks, the nonsynonymous (Ka) and synonymous substitutions (Ks) were calculated using MegaX (Kumar *et al.*, 2018). The analysis of synonymous and nonsynonymous substitutions was presented in Table 3. The high level of Ks and relatively low Ka were found among CII genes. Contrastly, a pattern of high Ka and lower Ks was seen among the CI genes, except for one gene pair (*TcsHSP27* and *TcsHSP28*). Differently, in *A. thaliana* and poplar, the pattern of low Ka and higher Ks among the cytosolic I genes was reported (Waters *et al.*, 2008). To explore the

selective constraints on duplicated cocoa *sHSP* genes, a ratio of nonsynonymous versus synonymous substitutions (Ka/Ks) for each pair of duplicated *sHSP* genes was further calculated. In general, a ratio of 1.0 shown that both genes were drifting neutrally; a Ka/Ks ratio > 1.0 indicated the accelerated evolution with positive selection, while a ratio < 1.0 indicated functional constraint, with a negative or purifying selection of the gene (Nekrutenko *et al.*, 2002). The Ka/Ks values from three CII *sHS*P genes were ranging from 0.222 to 0.636, which implied purifying/stabilizing selection. Differently, most of the Ka/Ks ratios from 11 duplicated CI *sHSP* genes were greater than 1.0 (Table 3), suggesting that the subclass CI *sHSP* genes were under a strong positive selection pressure with functional divergence occurring after tandem duplication. The expansion and evolution of CI and CII *sHSP* genes in the cocoa genome might be associated with the environmental adaptation of this tropical plant.

***In silico* gene expression analysis in different developmental stages of zygotic and somatic embryo:** In this study, *TcsHSPs* were found expressed in all developmental stages (Figure 3). In compasion with the expression level of *Acyl carrier protein B (ACP B)* which was the most stably expressed gene in various cocoa tissues (Pinheiro *et al.*, 2011), most of *TcsHSP* genes exhibited a relatively equal or higher expression level, except *TcsHSP07, TcsHSP10, TcsHSP24* (in mature somatic embryo) and *TcsHSP02* (in late Torpedo and mature somatic embryo). All *TcsHSP* genes displayed higher expression levels in mature zygotic embryo than in other stages of zygotic and somatic embryos. As compared with other *TcsHSPs*, eight members of the cocoa *sHSPs* such as *TcsHSP01, TcsHSP21, TcsHSP22, TcsHSP31* (CI group), *TcsHSP18* (CII group), *TcsHSP04, TcsHSP06* (MTI group), and *TcsHSP08* (PX group) were found highly expressed in embryo at different developmental stages. This finding is consistent with the observation described in previous works. It was reported that expression of all *sHSPs* genes in all developmental stages with the different patterns had been recorded in *S. bicolor* (Nagaraju *et al.*, 2020). Similarly, *HSP20s* were highly expressed in the development stages of the zygotic embryo in Chinese cabbage (Tao *et al.*, 2015), and during fruit development in grape (Ji *et al.*, 2019). Similar expression profiles were displayed within both tandem duplicated *TcsHSP* gene groups (CI and CII duplicated gene groups). This observation suggested that these duplicated genes retained similar structure and function. This finding confirmed the similar expression patterns recorded within the tandem duplicated *sHSP* gene groups in grape (Ji *et al.*, 2019).

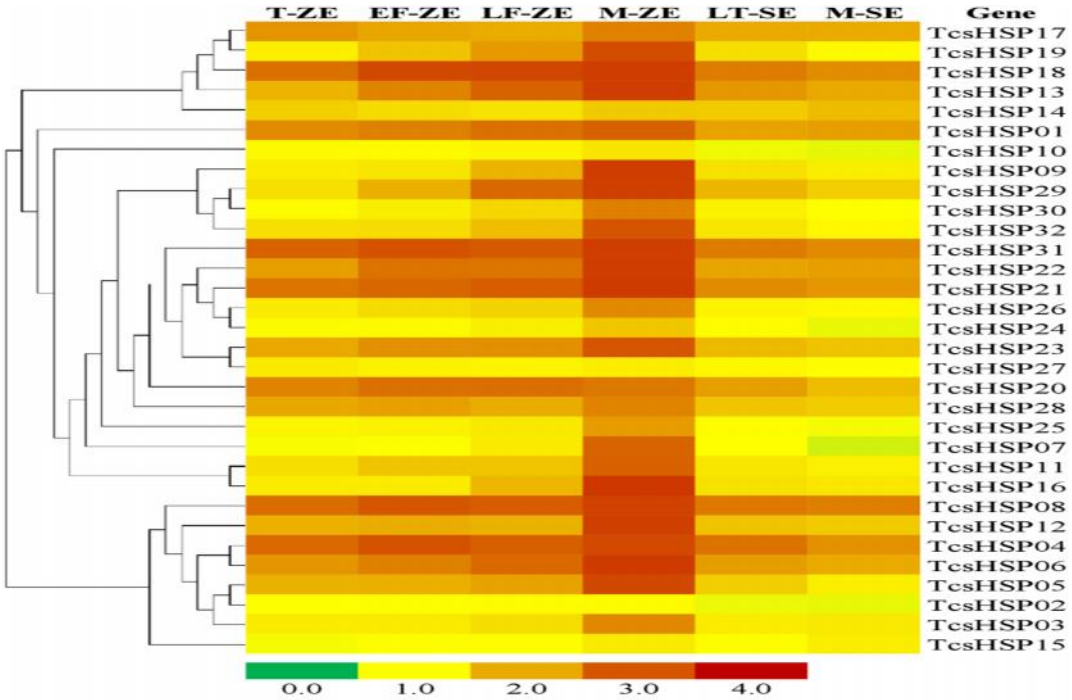

Figure 3. Expression profiles of the *T. cacao sHSP* genes during Zygotic (ZE) and Somatic Embryo (SE) maturation. Values represented relative expression level of TcsHSP genes per expression level of Acyl carrier protein B (ACP B) gene which was the most stable expressed gene in various tissues (Pinheiro *et al.*, 2011). T-ZE: Torpedo zygotic embryo, EF-ZE: Early-full zygotic embryo, LF-ZE: Late-full zygotic embryo, M-ZE: Mature zygotic embryo, LT-SE: Late Torpedo somatic embryo, M-SE: Mature somatic embryo

Conclusions: By using *in silico* methods, a total of 32 putative sHSP encoding genes were found in the cocoa genome. The detailed analysis disclosed their physico-chemical properties, structural organization, subcellular localizations, phylogenetic and evolutionary relations. Based on protein structure and phylogeny analysis, these cocoa *sHSP* genes were classified into 12 groups, including seven cytoplasmic and five organelle subgroups. The *sHSP* genes were randomly distributed in the entire cocoa genome with two condensed blocks on chromosome 9. Tandem duplication events resulted in the expansion of CI *sHSP* genes in cocoa genome. All *TcsHSP* genes revealed expression in different developmental stages of the zygotic and somatic embryo.

REFERENCES

Argout, X., J. Salse, J.M. Aury, M.J. Guiltinan, G. Droc, J. Gouzy, M. Allegre, C. Chaparro, T. Legavre, S.N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Poulain, M. Ruiz, Y. Roguet, M. Rodier-Goud, J.F. Barbosa-Neto, F. Sabot, D. Kudrna, J. S.S. Ammiraju, S.C. Schuster, J.E. Carlson, E. Sallet, T. Schiex, A. Dievart, M. Kramer, L. Gelley, Z. Shi, A. Bérard, C. Viot, M. Boccara, A.M. Risterucci, V. Guignon, X. Sabau, M.J. Axtell, Z. Ma, Y. Zhang, S. Brown, M. Bourge, W. Golser, X. Song, D. Clement, R. Rivallan, M. Tahi, J.M. Akaza, B. Pitollat, K. Gramacho, A. D'Hont, D. Brunel, D. Infante, I. Kebe, P. Costet, R. Wing, W.R. McCombie, E. Guiderdoni, F. Quetier, O. Panaud, P. Wincker, S. Bocs and C. Lanaud (2011). The genome of *Theobroma cacao*. *Nat. Genet*. 43(2):101-8. DOI:10.1038/ng.736

de Jong, W.W., G.J. Caspers and J.A. Leunissen (1998). Genealogy of the alpha-crystallin-small heat-shock protein superfamily. *Int. J. Biol. Macromol*. 22(3-4):151-62. DOI: 10.1016/s0141-8130(98)00013-0.

Emanuelsson, O., S. Brunak, G. von Heijne and H. Nielsen (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc*. 2(4):953-71. DOI:10.1038/nprot.2007.131

Figueira, A., L. Alemanno and R.E. Litz (2005). *Theobroma cacao* cacao. in Richard E.L. eds. *Biotechnology of Fruit and Nut Crops*, Cabi Publishing:639-669. DOI: 10.1079/9780851996622.0639

Gasteiger, E., C. Hoogland, A. Gattiker, M.R. Wilkins, R.D. Appel and A. Bairoch (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*. Springer: 571-607. DOI: 10.1385/1-59259-890-0:571

Guiltinan, M.J., J. Verica, D. Zhang and A. Figueira (2008). Genomics of *Theobroma cacao*, "the Food of the Gods". In: Moore, P.H., and R. Ming, eds. *Genomics of Tropical Crop Plants*. Springer New York:145-170. DOI: 10.1007/978-0-387-71219-2_6

Guo, M., J.H. Liu, J.P. Lu, Y.F. Zhai, H. Wang, Z.H. Gong, S.B. Wang and M.H. Lu (2015). Genome-wide analysis of the CaHsp20 gene family in pepper: comprehensive sequence and expression profile analysis under heat stress. *Front. Plant. Sci*. 6:806. DOI:10.3389/fpls.2015.00806

Ji, X-R, Y-H. Yu, P-Y. Ni, G-H. Zhang and D-L. Guo (2019). Genome-wide identification of small heat-shock protein (HSP20) gene family in grape and expression profile during berry development. *BMC Plant Biol*. 19(1):433. DOI:10.1186/s12870-019-2031-4

Jaya, N., V. Garcia and E. Vierling (2009). Substrate binding site flexibility of the small heat shock protein molecular chaperones. *Proc. Nat. Acad. Sci. USA*. 106(37):15604-9. DOI:10.1073/pnas.0902177106

Katoh, K. and D.M. Standley (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. 30(4):772-80. DOI:10.1093/molbev/mst010

Krishna, P. (2004). Plant responses to heat stress. In: Hirt, H and K. Shinozaki, eds. *Plant Responses to Abiotic Stress*. Springer Berlin Heidelberg; 73-101. DOI: 10.1007/978-3-540-39402-0_4

Kumar, S., G. Stecher, M. Li, C. Knyaz K. Tamura (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol*. 35(6):1547-1549. DOI:10.1093/molbev/msy096

Li, Q., H. Yu, P.B. Cao, N. Fawal, C. Mathé, S. Azar, H. Cassan-Wang, A.A. Myburg, J. Grima-Pettenati, C. Marque, C. Teulières and C. Dunand (2015). Explosive Tandem and Segmental Duplications of Multigenic Families in Eucalyptus grandis. *Genome Biol. Evol*. 4:1068-81. DOI: 10.1093/gbe/evv048

Maximova, S.N., S. Florez, X. Shen, N. Niemenak, Y. Zhang, W. Curtis and M.J. Guiltinan (2014). Genome-wide analysis reveals divergent patterns of gene expression during zygotic and somatic embryo maturation of *Theobroma cacao* L., the chocolate tree. *BMC Plant Biol*.14:185. DOI:10.1186/1471-2229-14-185

Motamayor, J.C., A.M. Risterucci, P.A. Lopez, C.F. Ortiz, A. Moreno and C. Lanaud (2002). Cacao domestication I: the origin of the cacao

cultivated by the Mayas. *Heredity*. 89(5):380-386. DOI: 10.1038/sj.hdy.6800156

Motamayor, J.C., K. Mockaitis, J. Schmutz, N. Haiminen, D. Livingstone 3rd, O. Cornejo, S.D. Findley, P. Zheng, F. Utro, S. Royaert, C. Saski, J. Jenkins, R. Podicheti, M. Zhao, B.E. Scheffler, J.C. Stack, F.A. Feltus, G.M. Mustiga, F. Amores, W. Phillips, J.P. Marelli, G.D. May, H. Shapiro, J. Ma, C.D. Bustamante, R.J. Schnell, D. Main, D. Gilbert, L. Parida and D.N. Kuhn (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*. 14(6):r53. DOI:10.1186/gb-2013-14-6-r53

Nagaraju, M., P.S. Reddy, S.A. Kumar, A. Kumar, G. Rajasheker, D.M. Rao and P.B.K. Kishor (2020). Genome-wide identification and transcriptional profiling of small heat shock protein gene family under diverse abiotic stress conditions in *Sorghum bicolor* (L.). *Int. J. Biol. Macromol*. 142:822-834. DOI: 10.1016/j.ijbiomac.2019.10.023

Nakamoto, H. and L. Vigh (2007). The small heat shock proteins and their clients. *Cell. Mol. Life. Sci*. 64(3):294-306. DOI:10.1007/s00018-006-6321-2

Nekrutenko, A., K.D. Makova and W.H. Li (2002). The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*. 12(1):198-202. DOI:10.1101/gr.200901

Pamilo, P. and N.O. Bianchi, 1993, Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol. Evol*. 10(2):271-81. DOI:10.1093/oxfordjournals.molbev.a040003

Park, C.J. and Y.S. Seo (2015). Heat Shock Proteins: A Review of the Molecular Chaperones for Plant Immunity. *Plant. Pathol. J*. 31(4):323-33. DOI:10.5423/ppj.rw.08.2015.0150

Pinheiro, T.T., C.G. Litholdo, Jr., M.L. Sereno, G.A. Leal, Jr., P.S. Albuquerque and A. Figueira (2011). Establishing references for gene expression analyses by RT-qPCR in *Theobroma cacao* tissues. *Genet. Mol. Res*. 10(4):3291-305. DOI:10.4238/2011.November.17.4

Pucciarelli, D.L. (2013). Cocoa and heart health: a historical review of the science. *Nutrients*. 5(10):3854-70. DOI:10.3390/nu5103854

Sarkar, N.K., K.Y. Kim and A. Grover (2009). Rice *sHsp* genes: genomic organization and expression profiling under stress and development. *BMC Genomics*.10:393. DOI:10.1186/1471-2164-10-393

Siddique, M., S. Gernhard, P. Koskull-Doring, E. Vierling and K.D. Scharf (2008). The plant sHSP superfamily: five new members in *Arabidopsis thaliana* with unexpected properties. *Cell Stress Chaperon*. 13:183-197. DOI: 10.1007/s12192-008-0032-6.

Tao, P., W.L. Guo, B.Y. Li, W.H. Wang, Z.C. Yue, J.L. Lei and X.M. Zhong (2015). Genome-wide identification, classification, and expression analysis of sHSP genes in Chinese cabbage (*Brassica rapa* ssp pekinensis). *Genet. Mol. Res*. 14(4):11975-93. DOI:10.4238/2015.October.5.11

Tissieres, A., H.K. Mitchell, and U.M. Tracy (1974). Protein synthesis in salivary glands of *Drosophila melanogaster*: relation to chromosome puffs. *J. Mol. Biol*.,84(3):389-98. DOI: 10.1016/0022-2836(74)90447-1

Voorrips, R.E. (2002). MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J. Hered*. 93(1):77-78. DOI: 10.1093/jhered/93.1.77

Wang L., K. Guo, Y. Li, Y. Tu, H. Hu, B. Wang, X. Cui and L. Peng (2010). Expression profiling and integrative analysis of the CESA/CSL superfamily in rice. *BMC Plant Biol*. 2010;10:282. DOI:10.1186/1471-2229-10-282

Waters, E.R. (2013). The evolution, function, structure, and expression of the plant sHSPs. *J. Exp. Bot*. 64(2):391-403. DOI:10.1093/jxb/ers355

Waters, E.R., B.D. Aevermann and Z. Sanders-Reed (2008). Comparative analysis of the small heat shock proteins in three angiosperm genomes identifies new subfamilies and reveals diverse evolutionary patterns. *Cell Stress Chaperon*. 13(2):127-42. DOI:10.1007/s12192-008-0023-7

Yang, S., X. Zhang, J.X. Yue, D. Tian and J.Q. Chen (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics*. 280(3):187-98. DOI:10.1007/s00438-008-0355-0