GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Genome-wide identification and expression analysis of the *CPP-like* gene family in soybean

L. Zhang[1]*, H.K. Zhao[2]*, Y.M. Wang[1], C.P. Yuan[1], Y.Y. Zhang[1], H.Y. Li[1], X.F. Yan[3], Q.Y. Li[4]* and Y.S. Dong[1]*

[1]Agro-Biotechnology Research Institute, Jilin Academy of Agricultural Sciences, Changchun, Jilin, China
[2]Crop Germplasm Institute, Jilin Academy of Agricultural Sciences, Changchun, Jilin, China
[3]College of Agriculture, Shenyang Agricultural University, Shenyang, Liaoning, China
[4]Institute of Plant Protection, Jilin Academy of Agricultural Sciences, Changchun, Jilin, China

*These authors contributed equally to this study.
Corresponding authors: Y.S. Dong / Q.Y. Li
E-mail: ysdong@cjaas.com / qyli1225@126.com

Genet. Mol. Res. 14 (1): 1260-1268 (2015)
Received February 18, 2014
Accepted August 27, 2014
Published February 13, 2015
DOI http://dx.doi.org/10.4238/2015.February.13.4

ABSTRACT. Cysteine-rich polycomb-like protein (*CPP-like*) genes are a group of transcription factors with highly conserved cysteine-rich domains and are widely distributed in animals and plants, but do not present in yeast. Previous studies have shown that members of this family play important roles in the development of reproductive tissue and in the control of cell division in plants. In this study, whole genome identification of soybean CPP transcription factors was performed using bioinformatic methods. The results showed that there were 20 CPP transcription factors in the soybean genome,

which encoded for 28 distinct CPP proteins. These transcription factors were distributed on 16 of 20 chromosomes. Phylogenetic relationship analysis showed that expression of CPP gene family members occurred before the differentiation of monocotyledons and dicotyledons. RNA-Seq analysis showed that 5 genes were highly expressed in all tissues, including Glyma10g39080, Glyma01g44670, Glyma101g66920, Glyma02g01540, and Glyma20g28740. One gene (Glyma14g14750) was specifically expressed in young leaves, while 2 genes (Glyma02g01540 and Glyma10g01580) were highly expressed in root nodules. Quantitative reverse transcriptase-polymerase chain reaction analysis revealed that the expression levels of most genes increased in the roots under high temperature stress. Our findings indicate that these genes are not only involved in growth and development, but also in the responses to high temperature stress in soybean roots.

Key words: CPP gene family; CXC domain; Expression analysis; Phylogenetic tree; Soybean

INTRODUCTION

Transcription factors can regulate gene expression at the mRNA transcription level. The cysteine-rich polycomb-like protein *(CPP)*-*like* gene family is a group of transcription factors present in animals and plants, but not in yeast. CPP transcription factors are a small gene family and the typical sequence consists of 1 or 2 cysteine-rich domains, known as CXC domains. CXC domains and the sequences connecting them are highly conserved in nearly all species (Riechmann et al., 2000; Andersen et al., 2007). CPP transcription factors have been found to have various functions in many plant species. The first CPP transcription factor, *TSO1*, was identified using map-based cloning in the model plant *Arabidopsis thaliana* and its functions were detected through mutant screening (Hauser et al., 1998, 2000; Song et al., 2000). The transcript of the *TSO1* gene was most abundant in flowers, although the highest level was observed in developing ovules and microspores. *tso1* mutants show a loss of control of directional cellular expansion and coordination of adjacent cell growth, as well as defects in karyokinesis and cytokinesis (Hauser et al., 1998, 2000; Song et al., 2000). In soybean, *CPP1* interacts with the promoter of the soybean leghemoglobin gene *Gmlbc3* to regulate its gene expression (Cvitanich et al., 2000). In addition, previous studies showed that the CPP gene family can bind DNA via their CXC domains to regulate gene expression (Hauser et al., 2000).

Soybean is a crop of global importance and is one of the most widely cultivated crops worldwide. Genome sequencing of soybean was recently completed (Schmutz et al., 2010). This provides a basis for identifying and analyzing the gene family at the genomic level. In the present study, the CPP gene family was identified using bioinformatic approaches and analyzed using a combination of chromosome mapping, phylogenetic relationship analysis, and gene expression analysis. These results will provide a foundation for further identification of plant CPP gene function.

MATERIAL AND METHODS

Plant material, extraction of RNA, and cDNA synthesis

A soybean inbred line (0086-2) was used in the present study. Seedlings were planted in potted trays after accelerating germination. Heat treatment was conducted on the soybean plants at 42°C during 4 leaf stages and samples were collected 0, 1, 2, 6, 12, and 24 h after heat treatments. Three biological replicates were performed for each time point. Roots were collected and immediately frozen in liquid nitrogen and stored at -80°C for subsequent RNA extraction. Total RNA from all samples was isolated using TRIzol (Invitrogen, Carlsbad, CA, USA) according to manufacturer instructions and then treated with DNaseI (Promega, Madison, WI, USA) to remove any traces of genomic DNA. First-strand cDNA syntheses were performed using the PrimeScript 1st Strand cDNA Synthesis Kit (TaKaRa, Shiga, China).

Sequence retrieval and chromosomal distribution of soybean CPP genes

CPP gene family members in soybean and *Arabidopsis thaliana* were identified from databases (http://www.soybase.org/ and http://datf.cbi.pku.edu.cn/). To identify soybean CPP genes, the amino acid sequence of *A. thaliana* CPP gene family members was used as a query, and the soybean genome database was retrieved using BLASTP. CXC domains were identified via the Pfam website (http://pfam.janelia.org/) to verify the accuracy of the obtained genes, and the candidates were harvested. Protein molecular weight, isoelectric point, and other parameters were estimated using ProtParam tools (http://www.expasy.org/tools/protparam). Distributions of these CPP genes on each chromosome were determined according to the soybean chromosome genome information (Liu and Meng, 2003).

Phylogenetic tree

Sequence alignments were conducted using the ClustalX software (Chenna et al., 2003). The phylogenetic tree was constructed by neighbor-joining using the MEGA 5 software (Tamura et al., 2011). Bootstrapping (1000 replicates) was used to evaluate the degree of support for a particular grouping pattern in the phylogenetic tree, and nodes with a bootstrap rate less than 50% were removed to display the length of each branch. Branch lengths were assigned by pairwise calculations of the genetic distances, and missing data were treated by pairwise deletions of the gaps.

Expression analysis of GmCPP genes based on RNA-Seq data-Atlas

RNA-Seq data-Atlas from the soybean database (http://www.soybase.org/) was downloaded. The RNA Seq-Atlas presented here provides high-resolution gene expression in a diverse set of 14 tissues, including young leaf, flower, 1-cm pod, pod shell 10 DAF (days after flowering), pod shell 14 DAF, seed 10 DAF, seed 14 DAF, seed 21 DAF, seed 25 DAF, seed 28 DAF, seed 35 DAF, seed 42 DAF, root, and nodule. Expression levels for each GmCPP were obtained and MeV was used for the heat map (Saeed et al., 2003).

Expression analysis of *GmCPP* genes under drought treatment

Specific primers were designed using the BioXM2.6 software according to the encoded

domain sequences of the CPP transcription factors (Table 1). Elongation factor-1α (Go479260) was selected as an internal control. Real-time quantitative reverse transcription-polymerase chain reaction (qPCR) was conducted using an Applied Biosystems StepOne Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) and SsoFast EvaGreen Supermix (Bio-Rad, Hercules, CA, USA). The total reaction volume was 20 μL, which included 10 μL SYBR premix Ex Taq™ (2X) mixture, 1 μL cDNA (diluted 10 times), 0.4 μL upstream primer (10 μM), 0.4 μL downstream primers (10 μM), and 8.2 μL ddH_2O. The reaction was performed with the following cycling profile: 95°C for 30 s, 40 cycles at denaturation at 95°C for 5 s, and 60°C for 30 s. Three technical replicates were performed for each sample. The calculation of gene expression levels followed the $2^{-\Delta\Delta Ct}$ method described by Livak and Schmittgen (2001).

RESULTS

Identification of CPP gene family members in soybean

To identify soybean CPP gene family members, systematic analysis was performed in the soybean genome downloaded from a database (http://www.soybase.org/) using the CPP gene family members of the 2 model plants rice and *A. thaliana* as queries. A total of 20 CPP gene family members were identified (GmCPP1-GmCPP20), which encoded 28 proteins. Large length variations in the amino acids (aa) of these genes were found from 108 aa (GmCPP18) to 897 aa (GmCPP17), and the molecular weights were between 11.9-97.61 kDa. However, variations in the isoelectric points of these proteins were relatively small, between 5.21 (Glyma14g14750.3) and 8.73 (Glyma01g39490.1) (Table 1).

Table 1. List of CPP genes in the soybean genome.

Name	Gene ID	Protein ID	Chromosome	Locus	No. of exons	Protein length	MW (kDa)	pI
GmCPP1	Glyma01g39490	Glyma01g39490.1	1	51412887..51422885	18	861	96.31	8.73
GmCPP2.1	Glyma01g44670	Glyma01g44670.1	1	55251010..55256148	10	774	84.48	6.04
GmCPP2.2		Glyma01g44670.2	1		9	759	82.76	6.1
GmCPP3	Glyma02g01540	Glyma02g01540.2	2	1100819..1110176	17	868	97.11	6.84
GmCPP4	Glyma03g38321	Glyma03g38321.1	3	44663318..44672260	17	814	93.59	8.33
GmCPP5	Glyma04g08670	Glyma04g08670.2	4	6791246..6797768	8	585	64.54	8.47
GmCPP6	Glyma05g35561	Glyma05g35561.1	5	39550797..39556496	8	514	55.52	7.06
GmCPP7	Glyma06g08774	Glyma06g08774.1	6	6393896..6399620	8	583	64.38	8.59
GmCPP8.1	Glyma07g09380	Glyma07g09380.2	7	7815899..7825023	8	559	60.18	6.79
GmCPP8.2		Glyma07g09380.3	7		8	496	53.75	7.94
GmCPP9.1	Glyma08g04180	Glyma08g04180.1	8	2947000..2953184	8	495	53.52	7.37
GmCPP9.2		Glyma08g04180.4	8		8	551	59.13	7.67
GmCPP10	Glyma09g32420	Glyma09g32420.2	9	38954446..38963061	8	559	60.36	6.82
GmCPP11	Glyma10g39080	Glyma10g39080.2	10	46797012..46802865	10	760	82.34	5.84
GmCPP12.1	Glyma10g01580	Glyma10g01580.3	10	1150091..1160317	17	870	97.27	6.83
GmCPP12.2		Glyma10g01580.4	10		17	869	97.14	6.83
GmCPP12.3		Glyma10g01580.5	10		17	810	90.42	6.17
GmCPP13	Glyma11g00920	Glyma11g00920.2	11	478506..483371	10	774	84.81	6.39
GmCPP14	Glyma11g05760	Glyma11g05760.1	11	4067398..4077158	18	861	96.14	8.72
GmCPP15	Glyma11g07150	Glyma11g07150.2	11	5011600..5018948	17	753	85.65	8.53
GmCPP16.1	Glyma14g14750	Glyma14g14750.1	14	15182849..15189259	12	876	95.78	5.41
GmCPP16.2		Glyma14g14750.2	14		12	874	95.55	5.41
GmCPP16.3		Glyma14g14750.3	14		12	770	84.49	5.21
GmCPP17.1	Glyma17g31399	Glyma17g31399.1	17	Gm17:34555932..34562163	12	897	97.61	5.8
GmCPP17.2		Glyma17g31399.2	17		12	897	97.61	5.8
GmCPP18	Glyma18g45736	Glyma18g45736.1	18	Gm18:55439735..55440323	1	108	11.99	6.49
GmCPP19	Glyma19g40430	Glyma19g40430.2	19	Gm19:46838288..46845575	16	662	75.59	8.45
GmCPP20	Glyma20g28740	Glyma20g28740.3	20	Gm20:37670760..37676598	10	774	84.29	6.02

MW = molecular weight; pI = isoelectric point.

Construction of the phylogenetic tree of soybean CPP proteins

To analyze the evolutionary relationships between members of the soybean CPP gene family, CPP proteins of 2 model plant species (rice and *Arabidopsis*) were selected for the construction of a phylogenetic tree (Figure 1). The CPP gene family members of the 3 species were divided into Group A, Group B, and Group C. Group A and Group B included members from the 3 species, while Group C only contained soybean CPP proteins. No orthologous proteins were identified in the clustering tree, but 15 pairs of paralog proteins were found between species, including AtCPP1, AtCPP7, AtCPP4, and AtCPP6 of *A. thaliana*, OsCPPP2, OsCPPP7, OsCPPP6, OsCPPP4, OsCPPP5, OsCPPP10, OsCPPP3, and OsCPPP11 of rice, and GmCPP6, GmCPP9.2, GmCPP8.1, GmCPP10, GmCPP5, GmCPP7, GmCPP16.1, GmCPP17.1, GmCPP2.1, GmCPP13, GmCPP11, GmCPP20, GmCPP1, GmCPP14, GmCPP3, GmCPP12.1, GmCPP4, and GmCPP15 from soybean. Nine pairs of paralog genes of soybean were located on different chromosomes.

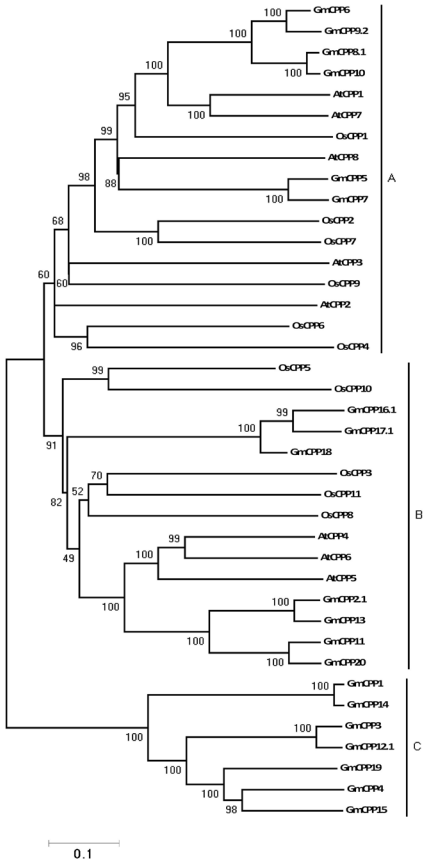

Figure 1. Neighbor-joining phylogenetic tree of CPP proteins from soybean, *Arabidopsis*, and rice. The phylogenetic tree was constructed using the Mega5.0 software based on full-length amino acid sequences from soybean, *Arabidopsis*, and rice. The tree was divided into 3 classes with a total of 47 proteins, including Group A, Group B, and Group C. The numbers are bootstrap values based on 1000 replicates.

Chromosomal locations of GmCPP gene family members

Distribution pattern diagrams of CPP genes on chromosomes were drawn according to soybean genome information (Figure 2). All soybean CPP genes were located on soybean chromosomes, but no CPP genes were found on chromosomes Chr12, Chr13, Chr15, and Chr16. There was only 1 CPP gene distributed on Chr02, Chr03, Chr04, Chr05, Chr06, Chr07, Chr08, Chr09, Chr14, Chr17, Chr18, Chr19, and Chr20. Two CPP genes were located on chromosome 01 and 10. Three CPP genes were distributed on chromosome 11. Most genes were located at the ends of the chromosomes.

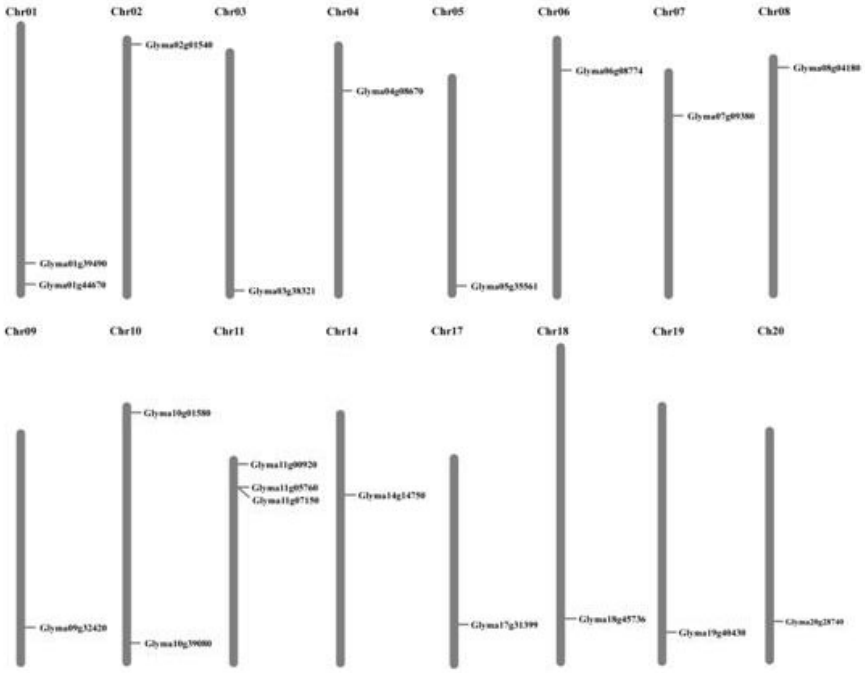

Figure 2. Chromosome location of 20 GmCPP genes on the 16 soybean chromosomes.

Expression analysis of soybean CPP gene based on RNA-Seq datas

To analyze the expression pattern of CPP genes, RNA-Seq expression data of soybean CPP genes were used. A total of 14 tissues were obtained, including young leaf, flower, 1-cm pod, pod shell 10 DAF, pod shell 14 DAF, seed 10 DAF, seed 14 DAF, seed 21 DAF, seed 25 DAF, seed 28 DAF, seed 35 DAF, seed 42 DAF, root, and nodule. Figure 3 shows that 5 genes (Glyma03g38321, Glyma04g08670, Glyma07g09380, Glyma11g07150, and Glyma19g40430) were not expressed in various tissues. Five genes (Glyma10g39080, Glyma01g44670, Glyma11g66920, Glyma02g01540, and Glyma20g28740) were highly expressed in young leaves, flowers, 1-cm pod, 10 DAF pod, and seeds. Glyma14g14750 showed low expression levels in young leaves, but was not expressed in other tissues. The expression of Glyma20g28740 in seeds was high. Root nodules showed high expression levels of Glyma02g01540, Glyma10g01580, and Glyma20g28740.

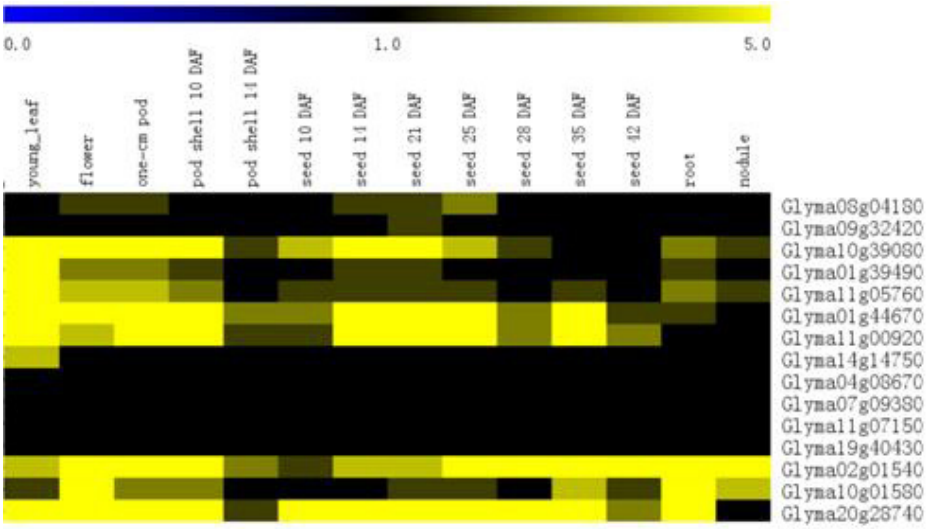

Figure 3. Expression profile cluster analysis of the soybean CPP proteins. The expression values of each CPP gene identified in the study was download from RNA-Seq data, including 14 organs, i.e., young leaf, flower, 1-cm pod, pod shell 10 DAF (days after flowering), pod shell 14 DAF, seed 10 DAF, seed 14 DAF, seed 21 DAF, seed 25 DAF, seed 28 DAF, seed 35 DAF, seed 42 DAF, root, and nodule.

Reverse transcription-qPCR expression analysis of GmCPP genes

Reverse transcription-qPCR was employed to analyze the expression of the 20 GmCPP to determine whether the soybean CPP transcription factors were expressed in the roots and to determine the effects of CPP transcription factor expression on the drought stress response (Figure 4). Our results showed that 18 GmCPP transcription factors had transcriptional activity in soybean roots, except for GmCPP03 and GmCPP07. Under drought stress, these transcription factors had different expression patterns. Expressions of most genes were up-regulated. Expression of GmCPP04, GmCPP15, and GmCPP19 was the highest after 24 h drought treatment. Expression of GmCPP09 peaked at 6-h drought treatment, while the expression levels of other genes were relatively low.

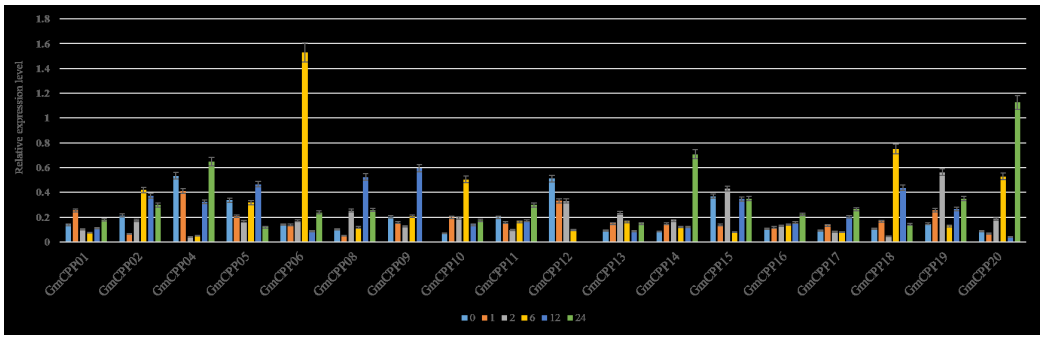

Figure 4. Expression levels of soybean CPP genes in root under drought stress.

DISCUSSION

In the present study, 20 CPP transcription factor family members were identified from the soybean genome database using bioinformatic methods. These members encoded 28 CPP proteins and all contained highly conserved CXC domains. Construction of a phylogenetic tree is helpful for not only analyzing evolutionary relationships of genes, but also researching orthologous genes between species and paralog genes within species. In this study, a total of 47 CPP proteins in *Arabidopsis*, rice, and soybean were used to construct a phylogenetic tree (Figure 1), which showed that the CPP gene families in each species could be divided into 3 groups (A, B, C). The CPP transcription factors of soybean were distributed in the 3 groups, while CPP transcription factors in *Arabidopsis* and rice were present only in Groups A and B, indicating that Group C may be specific for soybean CPP transcription factors, and that formation of the basic features of this gene family preceded the differentiation between monocotyledons and dicotyledons.

No orthologous genes were identified in the present study. However, 15 pairs of paralog CPP genes were found in soybean, *Arabidopsis*, and rice genomes, indicating that these gene families were extended in each species via specific-species manners. This phenomenon has been widely verified in other gene families in plants (Bai et al., 2002; Zhang et al., 2005; Jain et al., 2006). Paralogous genes typically exhibit different functions, while orthologous genes may retain the same functions (Tatusov et al., 1997). Many pairs of paralogous genes were found in soybean, suggesting that gene duplications played an important role in the evolution process. Gene duplications play an important role in plant evolution and are the main driving forces of genome evolution (Kent et al., 2003; Zhang, 2003). In particular, 9 pairs of paralogous genes found in soybean were located on different chromosomes, indicating that the soybean CPP genes might have undergone genome duplication events during evolution.

In addition, this study revealed that soybean CPP genes are not only involved in the growth and development of leaves, roots, root nodules, and other vegetative organs, but also involved in the growth and development of flowers, pods, and seeds, suggesting that these genes play important roles in the growth and development of soybean. Expression of Glyma1g39490 and Glyma1g05760 in various tissues was very similar, indicating that genes in similar groups have similar functions; this provides a foundation for increasing the understanding of CPP gene functions.

The RNA-Seq database was used to analyze the expression of soybean CPP transcription factor genes. Our results revealed that most genes were expressed in different tissues and organs of soybean, indicating that they participate in growth and development. Moreover, except for GmCPP03 and GmCPP07, the remaining 18 CPP genes were all induced by heat shock under drought stress conditions, indicating that these genes are involved in the responses of soybean root systems to high temperature stress and play important roles in regulating heat shock responses.

ACKNOWLEDGMENTS

Research supported by the National “863” Program (#2012AA101106-3), the Natural Science Foundation of China (#31200240), and the Science and Technology Development Program of Jilin Province (#20130102050JC).

REFERENCES

Andersen SU, Algreen-Petersen RG, Hoedl M, Jurkiewicz A, et al. (2007). The conserved cysteine-rich domain of a tesmin/TSO1-like protein binds zinc *in vitro* and TSO1 is required for both male and female fertility in *Arabidopsis thaliana*. *J. Exp. Bot.* 58: 3657-3670.

Bai J, Pennill LA, Ning J, Lee SW, et al. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.* 12: 1871-1884.

Chenna R, Sugawara H, Koike T, Lopez R, et al. (2003). Multiple sequences alignment with the Clustal series of programs. *Nucleic Acids Res.* 31: 3497-3500.

Cvitanich C, Pallisgaard N, Nielsen KA, Hansen AC, et al. (2000). CPP1, a DNA-binding protein involved in the expression of a soybean leghemoglobin c3 gene. *Proc. Natl. Acad. Sci. U. S. A*. 97: 8163-8168.

Hauser BA, Villanueva JM and Gasser CS (1998). *Arabidopsis* TSO1 regulates directional processes in cells during floral organogenesis. *Genetics* 150: 411-423.

Hauser BA, He JQ, Park SO and Gasser CS (2000). TSO1 is a novel protein that modulates cytokinesis and cell expansion in *Arabidopsis*. *Development* 127: 2219-2226.

Jain M, Tyagi AK and Khurana JP (2006). Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*). *Genomics* 88: 360-371.

Kent WJ, Baertsch R, Hinrichs A, Miller W, et al. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A*. 100: 11484-11489.

Liu RH and Meng JL (2003). MapDraw: a microsoft excel macro for drawing genetic linkage maps based on given genetic linkage data. *Hereditas* 25: 317-321.

Livak KJ and Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C (T)) method. *Methods* 25: 402-408

Riechmann JL, Heard J, Martin G, Reuber L, et al. (2000). *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105-2110.

Saeed AI, Sharov V, White J, Li J, et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34: 374-378.

Schmutz J, Cannon SB, Schlueter J, Ma J, et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178-183.

Song JY, Leung T, Ehler LK, Wang C, et al. (2000). Regulation of meristem organization and cell division by TSO1, an *Arabidopsis* gene with cysteine-rich repeats. *Development* 27: 2207-2217.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739.

Tatusov RL, Koonin EV and Lipman DJ (1997). A genomic perspective on protein families. *Science* 278: 631-637.

Zhang J (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18: 292-298.

Zhang S, Chen C, Li L, Meng L, et al. (2005). Evolutionary expansion, gene structure, and expression of the rice wall-associated kinase gene family. *Plant Physiol.* 139: 1107-1124.