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

Genome-wide analysis of the homeodomain-leucine zipper (HD-ZIP) gene family in peach (*Prunus persica*)

C.H. Zhang[1], R.J. Ma[1], Z.J. Shen[1], X. Sun[2], N.K. Korir[3] and M.L. Yu[1]

[1]Institute of Horticulture, Jiangsu Academy of Agricultural Sciences,
Nanjing, Jiangsu, China
[2]College of Horticulture, Nanjing Agricultural University, Nanjing,
Jiangsu, China
[3]Department of Agricultural Technology, Kenyatta University, Nairobi, Kenya

Corresponding author: M.L. Yu
E-mail: mly1008@aliyun.com

Genet. Mol. Res. 13 (2): 2654-2668 (2014)
Received August 5, 2013
Accepted December 6, 2013
Published April 8, 2014
DOI http://dx.doi.org/10.4238/2014.April.8.8

ABSTRACT. In this study, 33 homeodomain-leucine zipper (HD-ZIP) genes were identified in peach using the HD-ZIP amino acid sequences of *Arabidopsis thaliana* as a probe. Based on the phylogenetic analysis and the individual gene or protein characteristics, the HD-ZIP gene family in peach can be classified into 4 subfamilies, HD-ZIP I, II, III, and IV, containing 14, 7, 4, and 8 members, respectively. The most closely related peach HD-ZIP members within the same subfamilies shared very similar gene structure in terms of either intron/exon numbers or lengths. Almost all members of the same subfamily shared common motif compositions, thereby implying that the HD-ZIP proteins within the same subfamily may have functional similarity. The 33 peach HD-ZIP genes were distributed across scaffolds 1 to 7. Although the primary structure varied among HD-ZIP family proteins, their tertiary structures were similar. The results from this study will

©FUNPEC-RP www.funpecrp.com.br

be useful in selecting candidate genes from specific subfamilies for functional analysis.

Key words: Peach; Homeodomain-leucine zipper family;
Phylogenetic analysis

INTRODUCTION

Homeodomain-leucine zipper (HD-ZIP) genes are the most abundant group of homeodomain (HD) genes in plants. HD-ZIP proteins are mostly characterized by the presence of an HD that is closely linked to a leucine zipper (LZ) motif. The HD is responsible for specific DNA binding, whereas the LZ motif mediates protein dimerization (Sessa et al., 1998; Johannesson et al., 2001). Based on the DNA binding specificities, additional conserved motifs, and their physiological functions, HD-ZIP genes are further divided into 4 subfamilies (HD-ZIP I, II, III, and IV) in *Arabidopsis thaliana* (Sessa et al., 1998; Aso et al., 1999).

HD-ZIP I has 17 members (*ATHB1/HAT5*, *ATHB3/HAT7*, *ATHB5-7*, *ATHB12*, *ATHB13*, *ATHB16*, *ATHB20-23*, *ATHB40*, and *ATHB51-54*) in *Arabidopsis* (Henriksson et al., 2005). The HD-ZIP I proteins are characterized by the presence of an HD that is closely linked to an LZ motif. Some HD-ZIP I genes not only are reported to be involved in abscisic acid and sucrose signaling pathways but are also critical to plant abiotic stress responses, embryogenesis, and cotyledon and leaf development (Himmelbach et al., 2002; Johannesson et al., 2003). The HD-ZIP II subfamily consists of 9 members (*ATHB2/HAT4*, *ATHB4*, *HAT1-3*, *HAT9*, *HAT14*, *HAT17*, and *HAT22*) (Ciarbelli et al., 2008). All the 9 members possess a set of conserved cysteine molecules in and outside the LZ motif (Tron et al., 2002). Most of these genes are mainly implicated in phytochrome-mediated organ development, such as leaf morphogenesis (Ciarbelli et al., 2008), and some also respond to light quality changes, shade avoidance, and auxin as revealed by genetic and biochemical analyses (Morelli and Ruberti, 2002; Sawa et al., 2002).

The HD-ZIP III and IV subfamilies are defined by the presence of 2 additional domains, the steroidogenic acute regulatory protein related lipid transfer (START) domain and the START-adjacent domain (SAD). These 2 families can be distinguished by a fifth domain, the C-terminal MEKHLA motif, which is present in the HD-ZIP III subfamily and absent in the HD-ZIP IV subfamily (Mukherjee and Bürglin, 2006). The HD-ZIP III subfamily comprises only 5 genes in *Arabidopsis*, *ATHB8*, *PHAVOLUTA* (*PHV*)/*ATHB9*, *PHABULOSA* (*PHB*)/*ATHB14*, *REVOLUTA* (*REV*)/*INTERFASCICULARFIBERLESS1* (*IFL1*), and *CORONA* (*CNA*)/*ATHB15*/*INCURVATA4* (*ICU4*) (Prigge et al., 2005), but they are the key developmental regulators of *Arabidopsis* apical embryo and shoot radial patterning, shoot meristem formation, vascular differentiation, lateral organ polarity development, and auxin transportation (Baima et al., 2001; McConnell et al., 2001; Emery et al., 2003; Ohashi-Ito and Fukuda, 2003; Prigge et al., 2005).

The class IV HD-ZIP subfamily is also known as HDGL2 after the first identified gene *GLABRA2* (*GL2*). HD-ZIP IV transcription factors are likely present in all land plants. They have been described specifically in the dicots apple (*Malus domestica*; Dong et al., 1999) and tomato (*Solanum lycopersicum*; Isaacson et al., 2009), and in the monocots maize (*Zea mays*; Ingram et al., 2000), rice (*Oryza sativa*; Ito et al., 2003), sorghum (*Sorghum bicolor*; Swigonová et al., 2004). In *Arabidopsis*, HD-ZIP IV constitutes a large subfamily of genes that is composed of 16 members: *GLABRA2* (*GL2*)/*ATHB10*, *ARABIDOPSIS THALIANA MERISTEM LAYER1* (*ATML1*), *ANTHOCYANINLESS2* (*ANL2*), *PROTODERMAL FACTOR2* (*PDF2*), *HOMEODOMAIN GLABROUS 1-5* (*HDG1-5*), *HDG6*/*FWA*, and *HDG7-12* (Nakamura et al., 2006). Genetic analysis shows that HD-ZIP IV proteins function in epidermal processes, trichome formation, root development, and anthocyanin accumulation (Javelle et al., 2011).

Peach (*Prunus persica* L.) is considered to be one of the most widely grown and economically important stone fruit species in the Rosaceae family, comprising more than 3000 species in approximately 110 genera that are distributed worldwide. In addition to its ecological and high economic importance, peach is also emerging as a model tree species for comparative genomic studies, evolutionary studies, and plant development research because of its small genome size of 220-230 Mbp (about twice the size of the *A. thaliana* genome) and the relatively short reproductive time. Since the release of the peach genome sequence in 2010, current estimates indicate that peach has 28,689 transcripts and 27,852 genes (Jung et al., 2008). The functional annotation of these genes predicted that peach also contains some HD-ZIP genes. However, the biological functions of these genes in peach are not verified yet. The identification of the members and characterization of HD-ZIP genes is the first step to understand their function in the transcriptional regulation of a variety of biological processes related to growth and development, as well as various plant biotic and abiotic stress responses.

Compared to the largely investigated functions of *Arabidopsis* HD-ZIPs, only 1 peach HD-ZIP gene (ppa001386m, *ppe-ATHB8*) has recently been characterized (Zhang et al., 2012). Until now, no detailed systematic analysis of HD-ZIPs in peach has been performed, including member identification, genome organization, and gene. In this study, the identification and characterization of HD-ZIP family genes based on the current availability of a large public Genome Database for Rosaceae (GDR) and The *Arabidopsis* Information Resource (TAIR) database were attempted. According to phylogenetic and protein motif structural analyses, the subfamilies of the HD-ZIP family genes in peach were classified using the HD-ZIP family classification in *Arabidopsis* as a reference. Our results presented here may provide a subset of potential candidate HD-ZIP genes for further functional characterization of these HD-ZIP genes and future engineering modifications by utilizing these genes to examine development and stress tolerance in peach.

MATERIAL AND METHODS

Isolation of predicted HD-ZIP genes in peach

According to the identified proteins of HD-ZIP family genes in *Arabidopsis* that were downloaded from TAIR database (Huala et al., 2001), the hidden Markov model (HMM) profile of the HD-ZIP family was extracted from the Pfam database as described by Wang et al. (2010). With the aid of the HMM profile, searching against the peach proteins database in GDR was performed using identified proteins of the HD-ZIP family in *Arabidopsis* as query sequences. As a result, 33 amino acid sequences of HD-ZIP proteins were obtained after removing the redundant sequences. In order to confirm these predicted HD-ZIP family proteins, these protein sequences were then searched for specific conserved domains that each subfamily contained using the InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/) and CD search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) web servers. The numbers, names, and locations of conserved domains that were contained in each HD-ZIP protein sequence of peach were also recorded. Finally, coding DNA sequences (CDSs) corresponding to the HD-ZIP family proteins that were predicted above were extracted from total CDSs of peach and downloaded from the GDR database (Jung et al., 2008).

Motif display and phylogenetic analysis of predicted HD-ZIP proteins in peach

The online tool of MEME (Version 4.9.0) was used to search the conserved motifs that were shared by HD-ZIP proteins (http://meme.nbcr.net/meme/cgi-bin/meme.cgi, Bailey et al., 2006) by uploading the file of 33 amino acid sequences from the HD-ZIP family in peach. The parameters were set as follows: 0 or 1 occurrence of a single motif per sequence, motif width ranges of 2 to 300 aa, and 20 as the maximum number of motifs to find. All other parameters were set at default. The amino acid sequences of the HD-ZIP family in *Arabidopsis* were downloaded from TAIR database (http://www.arabidopsis.org/browse/genefamily/index.jsp), and a multiple alignment analysis among peach, *Arabidopsis*, *Vitis vinifera*, and *O. sativa* HD-ZIP proteins (amino acid sequences) was conducted using the Clustal W program built in the MEGA 5.0 software (Tamura et al., 2011). *V. vinifera* and *O. sativa* HD-ZIP protein sequences were downloaded from a previous report (Hu et al., 2012). The phylogenetic tree of HD-ZIP family proteins in peach, *Arabidopsis*, *V. vinifera*, and *O. sativa* was also constructed using the Clustal W tool in conjunction with the MEGA5.0 software using the neighbor-joining method and a bootstrap of 1000 replicates. Using a combination of phylogenetic trees and conserved domain analysis with characteristics and structure of genes, the HD-ZIP family in peach was classified into 4 subfamilies.

Gene characteristics and structure analysis of the predicted HD-ZIP genes in peach

Genomic sequences (peach v1.0), ID number, gene distribution on scaffold, and genome location of peach HD-ZIP genes were downloaded and obtained from the Phytozome database (http://www.phytozome.net/peach.php). The open reading frame (ORF) length of peach HD-ZIP genes was analyzed using the ORF finder from the National Center for Biotechnology Information. A structural figure of peach HD-ZIP genes, including exon and intron numbers and locations, was constructed and displayed using the gene structure display server (GSDS) web-based bioinformatic tool (http://gsds.cbi.pku.edu.cn/).

Characteristics of predicted HD-ZIP proteins in peach

The basic physical and chemical characteristics (primary structure) of peach HD-ZIP proteins were calculated by the online ProtParam tool (http://www.expasy.org/tools/protparam.html), including the number of amino acids, molecular weight, theoretical isoelectric point (pI), aliphatic index, and grand average of hydropathicity (GRAVY), among others. Analysis of the tertiary structure was performed using the online server ExPaSy Swiss-Model (http://swissmodel.expasy.org), which automatically makes amino acid sequences to form tertiary protein structures through homology modeling.

RESULTS

Identification of the predicted HD-ZIP genes in peach

With the aid of the HMM profile, after removing the redundant sequences, 33 HD-ZIP genes were identified in the peach genome (Table 1). All HD-ZIP candidates were manually analyzed using the InterProScan program and online tool CD-search to verify the presence of conserved domains. InterProScan and CD-search results demonstrated that all genes were predicted to encode proteins containing an HD. Among them, 21 genes were predicted to encode proteins containing both an HD and an LZ domain (later, they were classified into HD-ZIP I and II subfamilies). Four genes were predicted to encode a single START domain together with 1 LZ domain at the N terminus, 1 MEKHLA domain, and 1 HD (they were later classified into the HD-ZIP III subfamily). The remaining 8 genes contained both an HD and START domain but no MEKHLA domain (they were classified into the HD-ZIP IV subfamily).

Table 1. Characteristics of the predicted *HD-ZIP* family genes in peach.

Gene	Genome location	Scaffold distribution	ORF (bp)	Identity	E value	Highest homolog	Match ID	Exon No.	Intron No.
Ppa015952m	14838046-14838861	7	558	54.64	1E-44	AT5G03790.1	HB51	2	1
Ppa009419m	41865691-41866822	1	882	40.4	4E-43	AT2G22430.1	ATHB6	3	2
Ppa009498m	26863556-26865405	1	873	63.07	5E-94	AT1G69780.1	ATHB13	3	2
Ppa020465m	19802617-19804218	3	921	45.34	9E-58	AT5G15150.1	HAT7	3	2
Ppa008344m	18200468-18203048	6	1011	62.23	2E-54	AT2G22430.1	ATHB6	3	2
Ppa011221m	26090250-26091573	2	660	44.06	2E-36	AT3G61890.1	ATHB12	2	1
Ppa018002m	26504568-26506376	1	966	54.68	3E-33	AT3G01470.1	HAT5/ATHB1	3	2
Ppa011343m	16540922-16542630	7	645	55.19	7E-52	AT4G36740.1	ATHB40	3	2
Ppa008318m	18200468-18203048	6	1014	61.7	1E-54	AT2G22430.1	ATHB6	3	2
Ppa008495m	15413212-15415522	5	990	49.42	3E-36	AT3G01470.1	HAT5/ATHB1	3	2
Ppa009747m	20399064-20401045	3	840	65.22	9E-44	AT3G01470.1	HAT5	3	2
Ppa010647m	19683542-19685010	2	726	51.54	9E-43	AT2G46680.1	ATHB7	2	1
Ppa008984m	35211385-35212952	1	936	45.43	5E-54	AT4G37790.1	HAT22	3	2
Ppa017711m	10289845-10290360	4	516	43.56	2E-26	AT5G53980.1	ATHB52	1	0
Ppa008774m	7820745-7822915	5	960	61.61	5E-80	AT4G16780.1	ATHB2/HAT4	4	3
Ppa024161m	11762418-11764150	7	903	57.42	3E-73	AT3G60390.1	HAT3	4	3
Ppa008075m	11116950-11119040	7	1041	54.08	5E-67	AT5G06710.1	HAT14	4	3
Ppa007421m	10204755-10206514	2	1107	71.59	1E-53	AT5G06710.1	HAT14	4	3
Ppa009614m	20286988-20288415	6	858	81.41	1E-59	AT4G37790.1	HAT22	3	2
Ppa016510m	4995209-4996982	1	690	52.63	4E-43	AT5G06710.1	HAT14	4	3
Ppa011006m	30022178-30024745	1	684	66.32	1E-62	AT2G01430.1	ATHB17	4	3
Ppa001343m	4481032-4487670	4	2550	81.27	0	AT2G34710.1	ATHB14/PHB	18	17
Ppa001378m	7111056-7117525	6	2529	83.14	0	AT5G60690.1	REV	18	17
Ppa001405m	3503422-3509619	3	2514	87.26	0	AT1G52150.1	ATHB15	18	17
Ppa001386m	36826758-36831876	1	2523	82.99	0	AT4G32880.1	ATHB8	18	17
Ppa014792m	3416210-3420889	5	2496	55.07	0	AT5G46880.1	ATHB7	11	10
Ppa001894m	24898453-24902697	2	2238	60.44	0	AT4G00730.1	ANL2	8	7
Ppa002343m	225769-229529	3	2058	48.64	0	AT1G73360.1	HDG11	10	9
Ppa015345m	18156038-18159444	5	1734	66.09	7E-162	AT1G73360.1	HDG11	10	9
Ppa019198m	16655934-16660264	3	2271	79.97	0	AT1G05230.2	HDG2	11	10
Ppa001875m	1177592-1181051	4	2253	80.0	0	AT4G04890.1	PDF2	10	9
Ppa001436m	17450801-17456118	2	2490	66.9	0	AT4G00730.1	ANL2	9	8
Ppa001840m	3982925-3987373	3	2274	64.98	0	AT1G79840.1	GL2	11	10

The full-length coding sequences of the HD-ZIP genes in peach ranged from 516 (Ppa017711m) to 2550 bp (Ppa001343m) with an average of 1385 bp. To further investigate the relationship between the genetic divergence within the HD-ZIP family and gene duplication in peach, the scaffold location of each HD-ZIP gene was determined from the peach genomic sequences. The 33 peach HD-ZIP genes were distributed across scaffolds 1 to 7. Scaffold 1 had the largest number (7) of HD-ZIP genes followed by 6 on scaffold 3 and 5 on scaffold 2. In contrast, only 4 HD-ZIP genes were found on scaffolds 5, 6, and 7, and 3 HD-ZIP genes were found on scaffold 4. The peach HD-ZIP family contains a pair of genes that underwent a duplication event, Ppa008344m and Ppa008318m. Ppa008318m and Ppa008344m almost have the same nucleotide sequences except that the former has 3 tandem nucleotides that are missing in the latter. In most cases, there are 2 or more peach HD-ZIP genes for the orthologs in *Arabidopsis*, but in some cases, there are no orthologous peach HD-ZIP genes in *Arabidopsis*. The detailed information of HD-ZIP family genes in peach, including accession numbers, the highest homolog, and identity percent to their *Arabidopsis* orthologs, is listed in Table 1.

Phylogenetic analysis and motif display of predicted HD-ZIP proteins in peach

To further categorize the HD-ZIP protein subfamilies, as well as to analyze the phylogenetic relationships, a phylogenetic tree was constructed based on the alignment of the HD-ZIP amino acid sequences from eudicots (*A. thaliana* and *V. vinifera*), monocots (*O. sativa*), and 33 HD-ZIP proteins in peach (Figure 1). Following the subfamily classification of genes in *Arabidopsis* and phylogenetic tree from 4 species, 33 HD-ZIP proteins in peach were further determined to form 4 groups that were categorized as groups I to IV (Table 2), which correspond to groups I-IV as classified in previous reports (Henriksson et al., 2005; Prigge et al., 2005; Nakamura et al., 2006; Ciarbelli et al., 2008) in *Arabidopsis*. HD-ZIP I genes consisted of 14 members and was the largest subfamily in peach. In contrast, HD-ZIP III genes included the fewest HD-ZIP genes, only 4 members. In addition, there are 7 and 8 members in HD-ZIP II and IV subfamilies in peach, respectively.

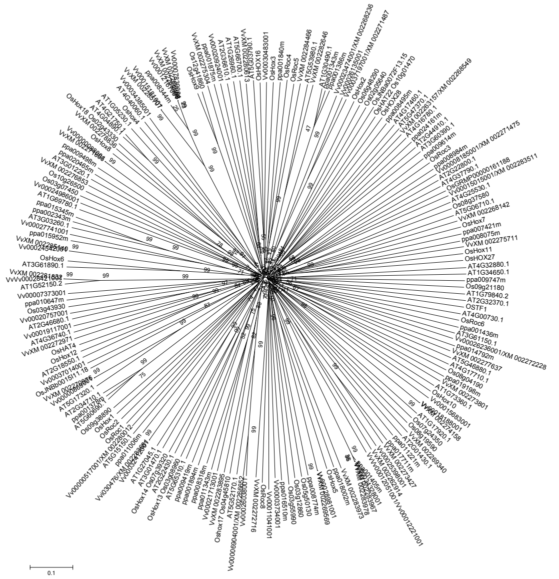

Figure 1. The phylogenetic tree of the Homeodomain-leucine zipper (HD-ZIP) protein family of peach, *Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa*. The tree was created by the bootstrap option of the CLUSTAL W multiple alignment packages and the neighbor-joining method using the 47, 63, 38, 33 HD-ZIP amino acid sequences of *Arabidopsis*, *Vitis vinifera*, *Oryza sativa*, and peach, respectively.

Table 2. Physical-chemical analysis of predicted HD-ZIP family protein in peach.

Subfamily	Gene ID	No. of amino acid	Molecular weight (Da)	Theoretical pI	Aliphatic index	Gravy
HD-ZIP I	ppa015952m	185	21259.8	8.44	69.51	-0.924
	ppa009419m	293	33288.7	5.73	57.61	-0.990
	ppa009498m	290	32771.6	5.72	66.00	-0.798
	ppa020465m	306	34762.6	6.43	61.21	-0.986
	ppa008344m	336	37977.1	5.21	62.65	-0.822
	ppa011221m	219	25272.2	6.33	68.58	-0.889
	ppa018002m	321	36193.9	4.69	64.74	-0.824
	ppa011343m	214	24771.8	6.41	66.50	-1.019
	ppa008318m	337	38048.2	5.21	62.76	-0.814
	ppa008495m	329	36763.2	4.67	66.05	-0.798
	ppa009747m	279	31813.7	4.80	62.87	-0.946
	ppa010647m	241	27725.6	4.93	65.60	-0.967
	ppa008984m	311	35033.9	6.09	63.73	-0.840
	ppa017711m	171	19880.6	7.10	85.56	-0.752
HD-ZIP II	ppa008774m	319	35757.9	7.69	63.26	-0.849
	ppa024161m	300	33391.3	7.53	66.37	-0.894
	ppa008075m	346	38204.8	8.69	65.23	-0.750
	ppa007421m	368	40137.1	6.32	51.74	-0.861
	ppa009614m	285	31537.3	8.62	61.30	-0.778
	ppa016510m	229	26386.1	9.21	66.46	-0.938
	ppa011006m	227	25495.9	9.13	65.73	-0.754
HD-ZIP III	ppa001343m	849	93018.8	5.92	83.30	-0.163
	ppa001378m	842	92159.9	6.02	87.61	-0.093
	ppa001405m	837	91800.4	6.06	88.71	-0.076
	ppa001386m	840	92410.7	5.93	84.68	-0.158
HD-ZIP IV	ppa014792m	831	92217.2	5.42	79.43	-0.393
	ppa001894m	745	81863.1	5.85	81.80	-0.283
	ppa002343m	685	76060.1	6.06	79.26	-0.357
	ppa015345m	581	64328.0	6.47	79.19	-0.326
	ppa019198m	756	82420.1	5.61	79.19	-0.323
	ppa001875m	750	82035.7	5.87	80.72	-0.326
	ppa001436m	829	90225.2	5.97	78.09	-0.309
	ppa001840m	757	84381.5	5.91	73.82	-0.509

Given the classification above, the 33 protein sequences of the HD-ZIP family in peach were subjected to the MEME web server to analyze conserved motif distribution (Figure 2). Almost all members in the same subfamily shared common motif compositions with each other, suggesting functional similarities among the HD-ZIP proteins within the same subfamily (Figure 2). Meanwhile, this also illustrates that the subfamily classification according to the phylogenetic tree above is logical. As illustrated in previous studies, most of the HD-ZIP proteins possessed an HD and LZ domain at the N terminus. In this study, motifs 1, 3, and 19 were present in all of the HD-ZIP family members in peach except for members of the HD-ZIP III subfamily. Motif 5, corresponding to the START domain, and conserved motif 2, representing MEKHLA, were found to be distributed in the C terminus of the HD-ZIP III subfamily proteins. Members in the HD-ZIP I and II subfamilies possessed a significantly reduced number (3-4) of conserved motifs compared to those in the HD-ZIP III and IV subfamilies (6-12).

Gene structure of the HD-ZIP genes in peach

Schematic structures of the HD-ZIP genes were as shown by the GSDS utility in Figure 3 and Table 1. Most closely related peach HD-ZIP members within the same subfamilies shared very similar gene structures in terms of either intron numbers or exon lengths (Figure 3), i.e., all the peach HD-ZIP I genes had 2 or 3 exons, except ppa017711m, and all 4 HD-ZIP III genes possessed the same number of introns (17) and exons (18) in their coding sequence. Similarly, among the 7 gene members in the peach HD-ZIP II subfamily, 6 genes had 4 exons and 3 introns. Besides, ppa009614m, which had 3 exons and 2 introns, was an exception to the HD-ZIP II subfamily. Nonetheless, the gene structures in peach HD-ZIP IV subfamily appeared to be more variable and displayed the largest number of exon/intron structure variants, i.e., HD-ZIP IV members had a large variation from 8 to 11 exons. We also investigated intron phases with respect to codons. Although the intron phases were remarkably well conserved within the same subfamilies, there were striking distinctions in the arrangement of introns and intron phases among subfamilies of peach HD-ZIP I-IV (Figure 3). The conservation of intron phases within peach HD-ZIP subfamilies and the striking dissimilarity between subfamilies may reciprocally provide support to the results from phylogenetic analysis and genome duplication.

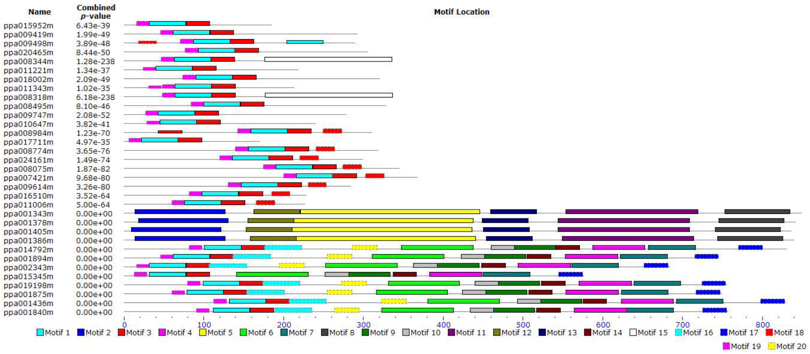

Figure 2. Motif distribution of HD-ZIP family proteins in peach. Motifs of HD-ZIP family proteins were elucidated by MEME web server. Each motif is represented by a number in the colored box. Non-overlapping sites with a P value better than 0.0001. The height of the motif 'block' is proportional to -log (P value), truncated at the height for a motif with a P value of 1e-10.

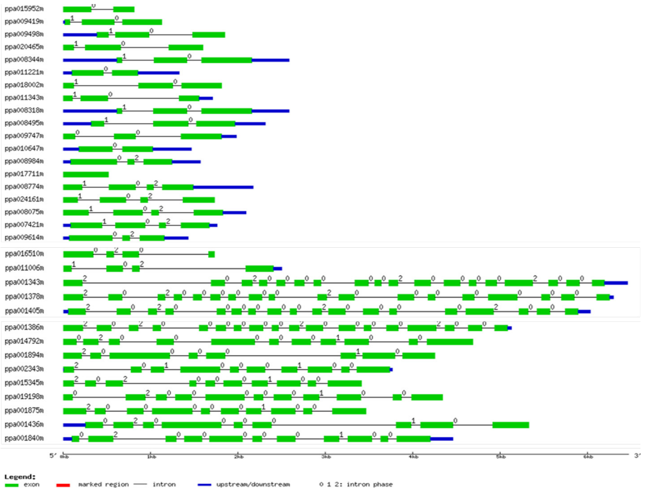

Figure 3. Exon/intron organization of genes of peach HD-ZIP family. Exons and introns were shown by filled green boxes and single lines, respectively. The sizes of exons and introns are proportional to their sequence lengths. The numbers indicate the splicing phases of the HD-ZIP genes, 0 refers to phase 0, 1 to phase 1, and 2 to phase 2. UTRs were displayed by thick blue lines at the ends.

Characteristics and structure of predicted HD-ZIP proteins in peach

Among the 33 amino acid sequences of peach HD-ZIP proteins, the shortest and longest sequences were 171 (ppa017711m) and 849 aa (ppa001343m), respectively. The primary structure of peach HD-ZIP proteins was calculated by the online ProtParam tool as shown in Table 2. A negative GRAVY index indicated that proteins in the HD-ZIP subfamily I and II were hydrophilic proteins, while the members of the HD-ZIP III and IV subfamilies were medium proteins. The ProtParam results revealed that the number of amino acids was positively correlated with the molecular weight of HD-ZIP family proteins. In addition, most of the proteins in HD-ZIP subfamily I and III-IV were acidic peptides and all the proteins in HD-ZIP subfamily II were alkalescent peptides based on the theoretical pI except ppa007421m.

The tertiary structures of representative predicted HD-ZIP family proteins in peach were built through the SWISS-MODEL web-based tool (http://swissmodel.expasy.org/). As shown in Figure 4, each tertiary structure of the 33 peach HD-ZIP family proteins contains 3 α-helices (red) and 4 random coils (gray). Especially, the 3 α-helices of each member of the HD-ZIP III subfamily contain 13, 9, and 15 amino acids according to the Swiss-Pdb Viewer, respectively, except ppa001405m, whose α-helices contain 13, 10, and 13 amino acids. Furthermore, the 4 random coils of the tertiary structure of each member in the HD-ZIP III subfamily contain 4, 5, 10, and 1 amino acids, respectively. However, ppa001405m is an exception, which has 7, 5, 9, and 4 amino acids. In all of them, the number of amino acids in each α-helix or random coil of some tertiary structure differed, creating slight variations in the length of the α-helix or random coils of each tertiary structure. In conclusion, the HD-ZIP family proteins generally have similar tertiary structures with slight differences in the length and amino acid composition of units that make up the tertiary structures.

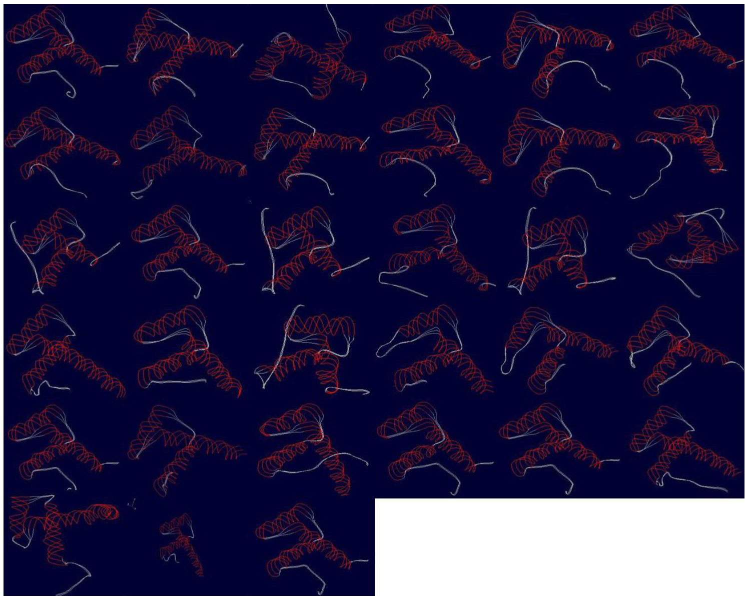

Figure 4. Tertiary structures of all 33 HD-ZIP family proteins in peach. Each tertiary structure picture was arranged according to the order of HD-ZIP family proteins as shown in Table 2. Red: Alpha helix; Gray: Random coil.

DISCUSSION

In this study, a comprehensive analysis of genes encoding HD-ZIP proteins in the peach genome was carried out, resulting in the identification of 33 HD-ZIP family genes. The availability of the complete genome sequences of peach and the previous identification of HD-ZIP family genes from some plant species (Aso et al., 1999; Agalou et al., 2008; Övernäs, 2010; Hu et al., 2012) enabled a comparison of individual families and groups of these genes with those in *Arabidopsis* (Henriksson et al., 2005; Prigge et al., 2005; Nakamura et al., 2006; Ciarbelli et al., 2008). Although peach has a large genome of 220-230 Mbp (Jung et al., 2008), which is larger than that of *A. thaliana* (145 Mbp) (Huala et al., 2001) and smaller than that of *O. sativa* (430 Mbp) (Goff et al., 2002), the number of genes in the HD-ZIP family in peach (33) is not more than that in *A. thaliana* (47) and is not much less than that in *O. sativa* (38). This expansion to more abundant HD-ZIP genes in the *A. thaliana* (47) and *V. vinifera* (63) genomes suggests a great need of HD-ZIP genes to participate in more complicated transcriptional regulation of these 2 species. Taken as a whole (Table 3), the HD-ZIP family has a similar number of genes in peach (33) and *O. sativa* (38), which is less than that in *Arabidopsis* (47) and *V. vinifera* (63) (Hu et al., 2012).

Table 3. Summary of the numbers of HD-ZIP family genes in peach

Species	Genome size (Mbp)	I		II		III		IV		Total
		No.	Percentage	No.	Percentage	No.	Percentage	No.	Percentage	
A. thaliana	145	17	36.17	9	19.15	5	10.64	16	34.04	47
V. vinifera	430	27	42.86	12	25.53	12	25.53	12	25.53	63
O. sativa	475	12	31.58	12	31.58	5	13.16	9	23.68	38
Peach	220-230	14	42.42	7	21.21	4	12.12	8	24.24	33

Like in *Arabidopsis*, *V. vinifera*, and *O. sativa*, the HD-ZIP family proteins in peach were also classified into 4 subfamilies. Because all the 4 subfamilies presented in monocot as well as dicot systems, it is probable that major diversification of this family predates the monocot-dicot divergence. These similar traits indicate that the HD-ZIP family of genes is highly conserved during evolution or separation processes of various plant species. Although the overall number of genes belonging to each subfamily of the HD-ZIP family was different among the species, the percentage of the number of genes in the HD-ZIP III subfamily relative to the total number of genes in the HD-ZIP family for each plant species was the lowest, while that in the HD-ZIP I subfamily was the highest. *V. vinifera* is an exception because each of its HD-ZIP subfamilies I-III has the same number of members. The uneven distributions of the HD-ZIP I and III subfamilies in these plants may have originated from some gene duplication events. Övernäs (2010) concluded that the members of HD-ZIP I in several species may play vital roles in abscisic acid and sucrose signaling pathways and are critical to plant embryogenesis and de-etiolation. Therefore, we deduced that a relatively large number of genes in these groups might be the consequence of evolution to adapt to various environmental changes.

It is widely accepted that the intron/exon position pattern provides clues to evolutionary relationships (Hu and Liu, 2011). Hu et al. (2012) reported that 4 (18%) of 22 *Populus* HD-ZIP I members had no introns in their coding regions, and this phenomenon did not exist in other *Populus* HD-ZIP subfamilies. Our schematic structure analysis indicates that having no intron in 1 gene of the HD-ZIP I subfamily is also a feature in peach (Table 1 and Figure 3). We also found that the genes of the HD-ZIP III subfamily in peach have the same and highest numbers of introns (17) in comparison with those in the HD-ZIP I-II and IV subfamilies. These findings in peach are consistent with those in *Populus*, where the number of introns for all genes in HD-ZIP III subfamily is the same (17) (Hu et al., 2012). In addition, introns of HD-ZIP IV members in peach varied from 7 to 10 in this study. This is also in consistent with findings in *Populus*, where the number of introns varied from 7 to 11 in the HD-ZIP IV subfamily. Besides further validating the subfamily classification of the HD-ZIP family, this observation indicates that the similar characteristics of HD-ZIP family genes among plants (both dicots and monocots) may contribute to their functional similarity within the same subfamily.

Based on the scaffold distribution and genome location, the phenomenon of gene duplication was also observed in the peach genome. For example, the unique presence of the duplication is Ppa008344m and Ppa008318m, which belong to the HD-ZIP I subfamily and are located on scaffold 6 (Table 1). Like in *Arabidopsis*, the duplication events of these genes may have a great impact on the amplification of members of a gene family in the peach genome. In addition, the pairwise evolution of duplicated genes also suggests that these genes might coordinately regulate certain biological processes that are common to this species, such as signal transduction and transcription. This is supported by previous findings that demonstrate that the duplicated genes that are involved in signal transduction and transcription are preferentially retained (Blanc and Wolfe, 2004).

The comparative research on the phylogenetic relationships among the HD-ZIP genes of peach, *Arabidopsis*, *V. vinifera*, and *O. sativa* was performed. The results revealed a great deal about the diversification and conservation of the HD-ZIP family in peach, where segment or whole-gene duplication, as well as a more ancient transposition and homing, might have contributed to the expansion of the HD-ZIP gene family. During this expansion, many groups have evolved, resulting in a high level of functional divergence in the HD-ZIP gene family. In a few cases, there are 2 or more peach HD-ZIP genes that are orthologs of a single gene in *Arabidopsis*; for example, peach has 3, 2, 3, and 2 orthologs of *ATHB6*, *ATHB1/HAT5*, *HAT14*, and *HDG11* in *Arabidopsis*, respectively. However, in some cases, there are no orthologs of *Arabidopsis* HD-ZIP genes in peach. For example, 4 HD-ZIP genes, including *ATHB5*, *ATHB4*, *PHAVOLUTA* (*PHB*)/*ATHB9*, and *ATHB10* (*GL2*), which belong to HD-ZIP I-IV, respectively, are present in *Arabidopsis* (AT5G65310, AT2G44910, AT1G30490, and AT1G79840) but not in peach species (Table 1). Therefore, we deduced that these members may have only evolved in *Arabidopsis* after functional divergence. Because peach is a woody species, selection either during domestication from its wild ancestor or during subsequent agricultural improvement may have been important for the evolution of the peach HD-ZIP family. Otherwise, it suggests that these HD-ZIP genes (missing in peach) were probably lost in the peach lineage after the diversification of the peach from the seed plant lineage, although we cannot deny the possibility that some of these genes may be found by further screening.

MEME is widely used to analyze similarities among DNA or protein sequences and produce a motif for each pattern that it discovers. The combined P value represents the combined best matches of a sequence to a group of motifs. The P value of a match of a sequence to a group of motifs is defined as the probability of a randomly generated sequence of the same length having sequence P values whose product is at least as small as the product of the sequence P values of the matches of the motifs to the given sequence. As for peach in this study, the P value of 12 proteins in the HD-ZIP III and IV subfamilies (36%) appeared to be zero (Figure 2), indicating that the sequence similarity among these genes was very high. This, in some ways, confirms that members within a given subfamily may have common and recent evolutionary origins. On the other hand, among the 33 proteins in the peach HD-ZIP family, no motif was repeated twice or thrice with the exception of ppa008984, which had a double motif 3.

In terms of each HD-ZIP subfamily in peach, the members in the same subfamily almost have the same number of motifs. For example, all genes in the peach HD-ZIP II and III subfamily have 4 and 6 motifs, respectively. The genes in the HD-ZIP I and IV subfamily have 3 and 12 motifs, respectively, except ppa015345m in the HD-ZIP IV subfamily and 5 exceptions in HD-ZIP I. Remarkably, all the motif positions for each member of the same subfamily were invariant or highly conserved without any insertion or deletion. This indicates that HD-ZIP proteins have the potential to recognize the same target genes. On the other hand, the highly conserved motif number and position among members in each HD-ZIP subfamily in peach also implies that they may have similar functions or a complex pattern of overlapping functions. As previously concluded, proteins within a subfamily that share motifs are likely to share similar functions (Hu and Liu, 2011). In *Arabidopsis*, the 5 members of HD-ZIP III show a close relationship in vascular development (Baima et al., 2001; McConnell et al., 2001; Ohashi-Ito et al., 2002; Zhong and Ye, 2004; Prigge et al., 2005). Furthermore, all of the HD-ZIP III genes except *ATHB8* play roles in patterning the apical portion of embryos (Emery et al., 2003; Prigge et al., 2005). Even though the functions of HD-ZIP genes in peach are not known, the gene tree of HD-ZIP genes suggests that HD-ZIP genes were present in the common ancestor of peach and angiosperms. Thus, it is conceivable that some may also play vital roles in vascular differentiation and patterning the apical portion of embryos in peach, which need further experimental verification.

The HD-ZIP genes or conserved motifs in peach were highly conserved with those of other species. The genes or proteins of the HD-ZIP I and II subfamilies in peach have similar lengths, but they are far shorter than those in HD-ZIP III and IV. Furthermore, HD-ZIP IV members are consistently somewhat shorter than HD-ZIP III members in peach. These characteristics were also observed in *Populus* (Hu et al., 2012) and *Arabidopsis* (Figure 5). The MEME result revealed that HD-ZIP III proteins have a C-terminal extension of about 150 amino acids termed MEKHLA. This region was also well conserved in the HD-ZIP III proteins from both monocots (*O. sativa*) and dicots (*Arabidopsis*) (Prigge et al., 2005; Agalou et al., 2008).

The differences in physical-chemical properties of side chains (amino acid sequences or primary structure) can result in the diversity of 3-dimensional protein folds that are observed in nature. Experiments that were performed decades ago demonstrated that the information specifying the 3-dimensional structure of a protein is contained in its amino acid sequence (Anfinsen, 1973). Although the protein sequences of the HD-ZIP family were different, the protein tertiary structures were similar overall. For example, no members had beta-sheet extended strands, which are present in the transcription factor families such as the ARF family. This point can be explained somewhat by the observations that the protein structure is more conserved than the protein sequence, and 2 sequences that share more than 30% sequence identity are likely to have similar structures (Mona, 2001).

In conclusion, this research has taken us a step further in understanding the basic information of the HD-ZIP family in peach. Phylogenetic and comparative analyses of HD-ZIP genes in peach will act as a first step toward a comprehensive functional characterization of the HD-ZIP gene family by reverse genetic approaches in the future. These results will assist in mining candidate genes for detailed characterization, provide useful information regarding the breeding of new cultivars that may be adaptable to less favorable environmental conditions, and illustrate mechanisms of development in peach.

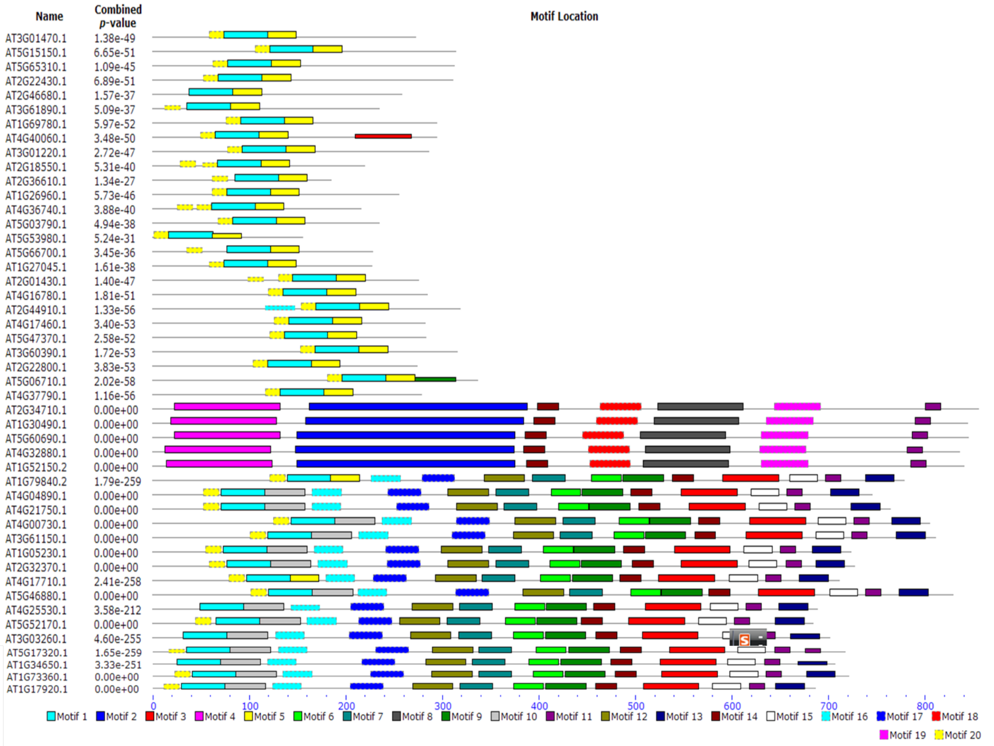

Figure 5. Motif distribution of HD-ZIP family proteins in *Arabidopsis thaliana*. Motifs of HD-ZIP family proteins were elucidated by MEME web server. Each motif is represented by a number in the colored box. Non-overlapping sites with a P value better than 0.0001. The height of the motif 'block' is proportional to -log (P value), truncated at the height for a motif with a P value of 1e-10.

ACKNOWLEDGMENTS

Research supported by the Natural Scientific Research Foundation (Youth Foundation) of Jiangsu Province (#BK20130700).

REFERENCES

Agalou A, Purwantomo S, Overnas E, Johannesson H, et al. (2008). A genome-wide survey of HD-Zip genes in rice and analysis of drought-responsive family members. *Plant Mol. Biol.* 66: 87-103.

Anfinsen CB (1973). Principles that govern the folding of protein chains. *Science* 181: 223-230.

Aso K, Kato M, Banks JA and Hasebe M (1999). Characterization of homeodomain-leucine zipper genes in the fern *Ceratopteris richardii* and the evolution of the homeodomain-leucine zipper gene family in vascular plants. *Mol. Biol. Evol.* 16: 544-552.

Bailey TL, Williams N, Misleh C and Li WW (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34: W369-W373.

Baima S, Possenti M, Matteucci A, Wisman E, et al. (2001). The *Arabidopsis* ATHB-8 HD-zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. *Plant Physiol.* 126: 643-655.

Blanc G and Wolfe KH (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16: 1679-1691.

Ciarbelli AR, Ciolfi A, Salvucci S, Ruzza V, et al. (2008). The *Arabidopsis* homeodomain-leucine zipper II gene family: diversity and redundancy. *Plant Mol. Biol.* 68: 465-478.

Dong YH, Yao JL, Atkinson RG and Morris BA (1999). Mdh3 encoding a phalaenopsis O39-like homeodomain protein expressed in ovules of *Malus domestica*. *J. Exp. Bot.* 50: 151-142.

Emery JF, Floyd SK, Alvarez J, Eshed Y, et al. (2003). Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Curr. Biol.* 13: 1768-1774.

Goff SA, Ricke D, Lan TH, Presting G, et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.

Henriksson E, Olsson AS, Johannesson H, Johansson H, et al. (2005). Homeodomain leucine zipper class I genes in *Arabidopsis*. Expression patterns and phylogenetic relationships. *Plant Physiol.* 139: 509-518.

Himmelbach A, Hoffmann T, Leube M, Hohener B, et al. (2002). Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in *Arabidopsis*. *EMBO J.* 21: 3029-3038.

Hu LF and Liu SQ (2011). Genome-wide identification and phylogenetic analysis of the ERF gene family in cucumbers. *Genet. Mol. Biol.* 34: 624-633.

Hu R, Chi X, Chai G, Kong Y, et al. (2012). Genome-wide identification, evolutionary expansion, and expression profile of homeodomain-leucine zipper gene family in poplar (*Populus trichocarpa*). *PLoS One* 7: e31149.

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, et al. (2001). The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29: 102-105.

Ingram GC, Boisnard-Lorig C, Dumas C and Rogowsky PM (2000). Expression patterns of genes encoding HD-ZipIV homeo domain proteins define specific domains in maize embryos and meristems. *Plant J.* 22: 401-414.

Isaacson T, Kosma DK, Matas AJ, Buda GJ, et al. (2009). Cutin deficiency in the tomato fruit cuticle consistently affects resistance to microbial infection and biomechanical properties, but not transpirational water loss. *Plant J.* 60: 363-377.

Ito M, Sentoku N, Nishimura A and Hong SK (2003). Roles of rice *GL2*-type homeobox genes in epidermis differentiation. *Breed. Sci.* 53: 245-253.

Javelle M, Klein-Cosson C, Vernoud V, Boltz V, et al. (2011). Genome-wide characterization of the HD-ZIP IV transcription factor family in maize: preferential expression in the epidermis. *Plant Physiol.* 157: 790-803.

Johannesson H, Wang Y and Engström P (2001). DNA-binding and dimerization preferences of *Arabidopsis* homeodomain-leucine zipper transcription factors *in vitro*. *Plant Mol. Biol.* 45: 63-73.

Johannesson H, Wang Y, Hanson J and Engström P (2003). The *Arabidopsis thaliana* homeobox gene ATHB5 is a potential regulator of abscisic acid responsiveness in developing seedlings. *Plant Mol. Biol.* 51: 719-729.

Jung S, Staton M, Lee T, Blenda A, et al. (2008). GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.* 36: D1034-D1040.

McConnell JR, Emery J, Eshed Y, Bao N, et al. (2001). Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature* 411: 709-713.

Mona S (2001). Predicting Protein Secondary and Supersecondary Structure. Princeton University, CRC Press, Boca Raton, 3-5.

Morelli G and Ruberti I (2002). Light and shade in the photocontrol of *Arabidopsis* growth. *Trends Plant Sci.* 7: 399-404.

Mukherjee K and Bürglin TR (2006). MEKHLA, a novel domain with similarity to PAS domains, is fused to plant homeodomain-leucine zipper III proteins. *Plant Physiol.* 140: 1142-1150.

Nakamura M, Katsumata H, Abe M, Yabe N, et al. (2006). Characterization of the class IV homeodomain-Leucine Zipper gene family in *Arabidopsis*. *Plant Physiol.* 141: 1363-1375.

Ohashi-Ito K and Fukuda H (2003). HD-zip III homeobox genes that include a novel member, ZeHB-13 (Zinnia)/ATHB-15 (*Arabidopsis*), are involved in procambium and xylem cell differentiation. *Plant Cell Physiol.* 44: 1350-1358.

Ohashi-Ito K, Demura T and Fukuda H (2002). Promotion of transcript accumulation of novel Zinnia immature xylem-specific HD-Zip III homeobox genes by brassinosteroids. *Plant Cell Physiol.* 43: 1146-1153.

Övernäs E (2010). Characterization of Members of the HD-Zip I and DREB/ERF Transcription Factor Families and Their Functions in Plant Stress Responses. Acta Universitatis Upsaliensis: Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology, Uppsala.

Prigge MJ, Otsuga D, Alonso JM, Ecker JR, et al. (2005). Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *Plant Cell* 17: 61-76.

Sawa S, Ohgishi M, Goda H, Higuchi K, et al. (2002). The HAT2 gene, a member of the HD-Zip gene family, isolated as an auxin inducible gene by DNA microarray screening, affects auxin response in *Arabidopsis*. *Plant J.* 32: 1011-1022.

Sessa G, Steindler C, Morelli G and Ruberti I (1998). The *Arabidopsis* Athb-8, -9 and -14 genes are members of a small gene family coding for highly related HD-ZIP proteins. *Plant Mol. Biol.* 38: 609-622.

Swigonová Z, Lai J, Ma J, Ramakrishna W, et al. (2004). Close split of sorghum and maize genome progenitors. *Genome Res.* 14: 1916-1923.

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.

Tron AE, Bertoncini CW, Chan RL and Gonzalez DH (2002). Redox regulation of plant homeodomain transcription factors. *J. Biol. Chem.* 277: 34800-34807.

Wang Y, Deng D, Bian Y, Lv Y, et al. (2010). Genome-wide analysis of primary auxin-responsive *Aux/IAA* gene family in maize (*Zea mays*. L.). *Mol. Biol. Rep.* 37: 3991-4001.

Zhang CH, Zhang YP, Guo L and Han J (2012). Characterization of the miR165 family and its target gene *Pp-ATHB8* in *Prunus persica*. *Sci. Hortic.* 146: 21-28.

Zhong R and Ye ZH (2004). Amphivasal vascular bundle 1, a gain-of-function mutation of the IFL1/REV gene, is associated with alterations in the polarity of leaves, stems and carpels. *Plant Cell Physiol.* 45: 369-385.