ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 09, Issue, 12, pp. 32139-32152, December, 2019

RESEARCH ARTICLE OPEN ACCESS

GENOME WIDE ANALYSIS OF TRANSCRIPTION FACTORS OF HD-ZIP GENE FAMILY OF ARABIDOPSIS THALIANA

Shaiq Sultan*[1], Muahmmad Shahid Javaid[1], Beenish Naz[1], Sidra tul Muntaha[1], Bilal Saleem[1], Anum Arshad[1], Oreha Sultan[2], Muhammad Amjad Ali[3,4]

[1]Department of Bioinformatics & Biotechnology, Government College University, 38000 Faisalabad, Pakistan
[2]Department of Botany, University of Agriculture, 38040 Faisalabad, Pakistan
[3]Department of Plant Pathology, University of Agriculture, 38040 Faisalabad, Pakistan
[4]Centre of Agricultural Biochemistry & Biotechnology, University of Agriculture, 38040 Faisalabad, Pakistan

ARTICLE INFO

Article History:
Received 27th September, 2019
Received in revised form
16th October, 2019
Accepted 20th November, 2019
Published online 30th December, 2019

Key Words:

HD-ZIP family,
Arabidopsis Thaliana, Homeo box,
Phylogenetic analysis, Motif

***Corresponding author:* Shaiq Sultan**

ABSTRACT

HD-ZIP proteins are a class of transcription factors family that have a conserved homeodomain in all of its factors. Homeo box; conserved part of home domain consists of 56 amino acids residues and involved in plant development from formation pattern to specification into different cell types. We retrieved the protein sequence of all the 48 transcription factors of HD-ZIP family and performed genome wide analysis and grouped them into four subfamilies. A comprehensive genome wide analysis in this study include mapping of all the transcription factors on 5 chromosomes of Arabidopsis thaliana, gene structure analysis by mapping introns and exons, multiple sequence alignment to find out conserved domain, phylogenetic analysis, promoter analysis by taking 1000bp upstream genomic sequence of all these transcription factors and motif analysis. This classification and analysis further categorized the transcription factors of 4 HD-ZIP subfamilies into different classes and revealed a deep evolutionary relationship among them and thus help to explore further functioning of these factors. These results help us to investigate functional homology among these factors on the basis of class grouping, comparison of tree with motifs that further depends upon the number of exons and introns.

Citation: Shaiq Sultan, Muahmmad Shahid Javaid, Beenish Naz, Sidra tul Muntaha *et al.*, 2019. "Genome wide analysis of transcription factors of hd-zip gene family of Arabidopsis Thaliana", *International Journal of Development Research*, 09, (12), 32139-32152.

INTRODUCTION

Transcription factors (TFs) are proteins that affect many biological processes such as development, growth and cell division and show response to stress and environmental stimuli in organisms or cells. TFs bind to DNA and either repress or activate expression of gene at the level of mRNA transcription (Chen *et al*., 2014). The plants are divided into more than 60 families of transcription factors that have different functional activities (Qin *et al*., 2014; Ruth *et al*., 2012). Typical TFs mainly contain two types of domains: DNA binding domain and transcriptional activation domain; the first domain recognizes target DNA sequences while the second domain initiates transcription (Chen *et al*., 2014). This paper discuss about analysis of HD-ZIP transcription factors family in *Arabidopsis thaliana.*

Homeobox (HB), one of the important transcription factors families. Each HB gene encodes a conserved 56 amino acid sequence called the homeodomain (HD) which is responsible for DNA binding at specific site (Chen *et al*., 2014). Homeodomain (HD) proteins play fundamental roles in a multiform set of plant developmental procedures, from formation pattern to specifications of cell types(Hu *et al*., 2012). HD-containing proteins were grouped into four families, comprising HD-Zip (homeodomain-leucine-zipper). HD-Zip proteins are ubiquitous in plants and convey important duty in various procedures of plant growth and development. HD-Zip proteins contain a leucine motif contiguous to the N-terminus of the homeodomain (Chen *et al*., 2014). In plants Homeodomain-leucine zipper (HD-ZIP) genes are the most abundant group of HD genes but in other eukaryotes no HD-ZIP genes present. The presence of a HD domain and an adjacent Leucine Zipper (LZ) motif are the Unique features of

HD-ZIP proteins (Hu *et al*., 2012). The *Arabidopsis* (*Arabidopsis thaliana*) genome comprises 48 genes believed to encode HD-Zips, which are clustered into four subfamilies based on their additional conserved domains, structures and physiological functions. Members of group I and II recognize similar pseudo palindromic binding sites (BSs; CAATNATTG). While HD-Zip III and IV proteins interact with slightly different sequences (GTAAT [G/C] ATTAC and TAAATG[C/T] A, respectively). Members of the HD-Zip gene family contain a special conserved HD and a common conserved LZ domain. The difference between the four subfamilies mainly arose in the region downstream of the LZ domain, which contains different domains(Chen *et al*., 2014). Arabidopsis HD-ZIP I subfamily comprising 17 members (ATHB1/HAT5, ATHB3/HAT7, ATHB5–7, ATHB12, ATHB13, ATHB16, ATHB20–23, ATHB40, and ATHB51–54). Arabidopsis HD-ZIP I genes are does not show response to abscisic acid (ABA) signaling, sugar signaling and abiotic stresses, but also crucial to plant de-etiolation and embryogenesis. HD-ZIP I proteins also play some role in ABA responses (Hu *et al*., 2012).

Arabidopsis HD-ZIP II subfamily containing nine members (ATHB2/HAT4, ATHB4, HAT1–HAT3, HAT9, HAT14, HAT17 and HAT22). All nine members have a cellular redox status perceptive CPSCE (Cys, Pro, Ser, Cys, and Glu) motif at the downstream of LZ motif and most of these genes show response to shading, light and auxin as disclose by biochemical and genetic analyses (Chen *et al*., 2014). Arabidopsis HD-ZIP III subfamily consist of only five genes, PHABULOSA (PHB)/ATHB14, PHAVOLUTA (PHV)/ ATHB9, REVOLUTA(REV)/ INTERFASCICULAR FIBERLESS1 (IFL1), ATHB8 and CORONA (CNA)/ ATHB15/ INCURVATA4 (ICU4). But they are the key developmental regulators of Arabidopsis apical embryo patterning, formation of shoot meristem, determination of organ polarity, vascular differentiation as well as transportation of auxin (Chen *et al*., 2014). In Arabidopsis thaliana, HD-ZIP IV (also known as HD-GL2) a large subfamily of genes containing 16 members: GLABRA2 (GL2)/ATHB10, ARABIDOPSIS THALIANA MERISTEMLAYER1 (ATML1), ANTHOCYANINLESS2 (ANL2), PROTODERMAL FACTOR2 (PDF2), HOMEODOMAIN GLABROUS1 (HDG1)-HDG5, HDG6/FWA and HDG7–HDG12. Genetic analysis shows that HD-ZIP IV proteins play critical roles in cell differentiation of epidermal, formation of trichome, development of root and accumulation of anthocyanin. (Chen *et al*., 2014). A lot of work has been carried out on the functional studies of HD-ZIP family of *Arabidopsis thaliana* but no sufficient work is done on *in silico* genome wide analysis of this family. The purpose of this study was to analyse 58 HD-ZIP (homeodomain-leucine-zipper) transcriptional factors from *Arabidopsis Thaliana*; their gene structure analysis, promoter analysis, phylogenetic analysis, conserved domain analysis, chromosomal mapping and analysis of cis regulatory elements is performed. These have been carried out by applying various bioinformatics tools.

MATERIALS AND METHODS

Chromosome mapping of HD-ZIP genes: The physical location was demonstrated of each transcription factor of HD-ZIP family of *Arabidopsis thaliana*. The chromosomal mapping of all the 48 transcription factors of HD-ZIP family on the 5 chromosomes of *Arabidopsis thaliana* was done by using the online available chromosome map tool of TAIR. *(*http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp)

Identification of conserved domain: The core of Homeo domain was studied in all the 48 genes of HD-ZIP family of *Arabidopsis thaliana* including all the subfamilies. For this purpose, multiple sequence alignment of all the 48 genes was performed by Unipro UGNENE software. Then protein sequences of all the genes were compared to identify the whole domain.

Phylogenetic analysis: For phylogenetic analysis, multiple sequence alignment of the protein sequences of all 48 HD-ZIP genes of *Arabidopsis* was performed by CLUSTALW that was available in MEGA 6.0 software as a built in program. These 48 HD-ZIP genes in *Arabidopsis* were classified into sub families on the basis of their function and domain. The parameters used for alignment were as follow: gap opening penalty: 10; gap extension penalty: 0.2; protein weight matrix: gonnet; residue-specific penalties: on; hydrophilic penalties: on; gap separation distance: 4; end gap separation: off; use negative matrix: off; delay divergent cutoff: 30%. On the basis of this multiple sequence alignment, an unrooted phylogenetic tree was constructed using Neighbour Joining (NJ) method. Phylogenetic tree was divided into 4 sub families on the basis of their functions.

Synteny analysis: Frosynteny analysis, the protein sequences of all the 48 genes of HD-ZIP family of *Arabidopsis thaliana* were retrieved from an online database of plant transcriptional factors (http://planttfdb.cbi.pku.edu.cn/). From these sequences two files were created; in first file there were protein sequences of the genes of HD-ZIP family from gene number 1 to gene number 24 and in second file there were protein sequences form gene number 25 to gene number 48. The sequences of these two files were in FASTA format. Then synteny analysis was performed by using an online tool that is available at http://tools.bat.infspire.org/circoletto/. Files were uploaded and the result was taken by using default parameters.

Gene structure analysis: The Database of Arabidopsis Transcription Factors (DATF)(http://datf.cbi.pku.edu.cn/), TAIR (www.arabidopsis.org/) and Arabidopsis.org were used to extract the information and annotate the gene structure .power point and excel sheets were used to explain the gene structure.

Identification of conserved motifs in proteins domain: Conserved motif analysis was done by using MEME online software Version 4.9.1(http://meme.nbcr.net/meme/cgi-bin/meme.cgi). The complete dataset (protein sequence) of 48 transcriptional factors of HD-ZIP family was pasted in the corresponding box features/properties selected to specify the results were; number of repetitions - any number of repetitions, number of motifs - 15, minimum motif width -10, maximum motif width-56.

Analysis of Cis regulatory elements in promoter: Promoter analysis was done by retrieving 1000bp upstream of transcription start site of all 48 genes from phytozome version 9.1 (http://www.phytozome.net/). The complete list of cis regulatory elements was obtained by using PLACE (Higo *et al*., 1999) (http://www.dna.affrc.go.jp/PLACE/). We choose the most commonly present cis regulatory elements in promoter sequences of all HD-ZIP genes. The location of cis regulatory elements was mapped by using power point.

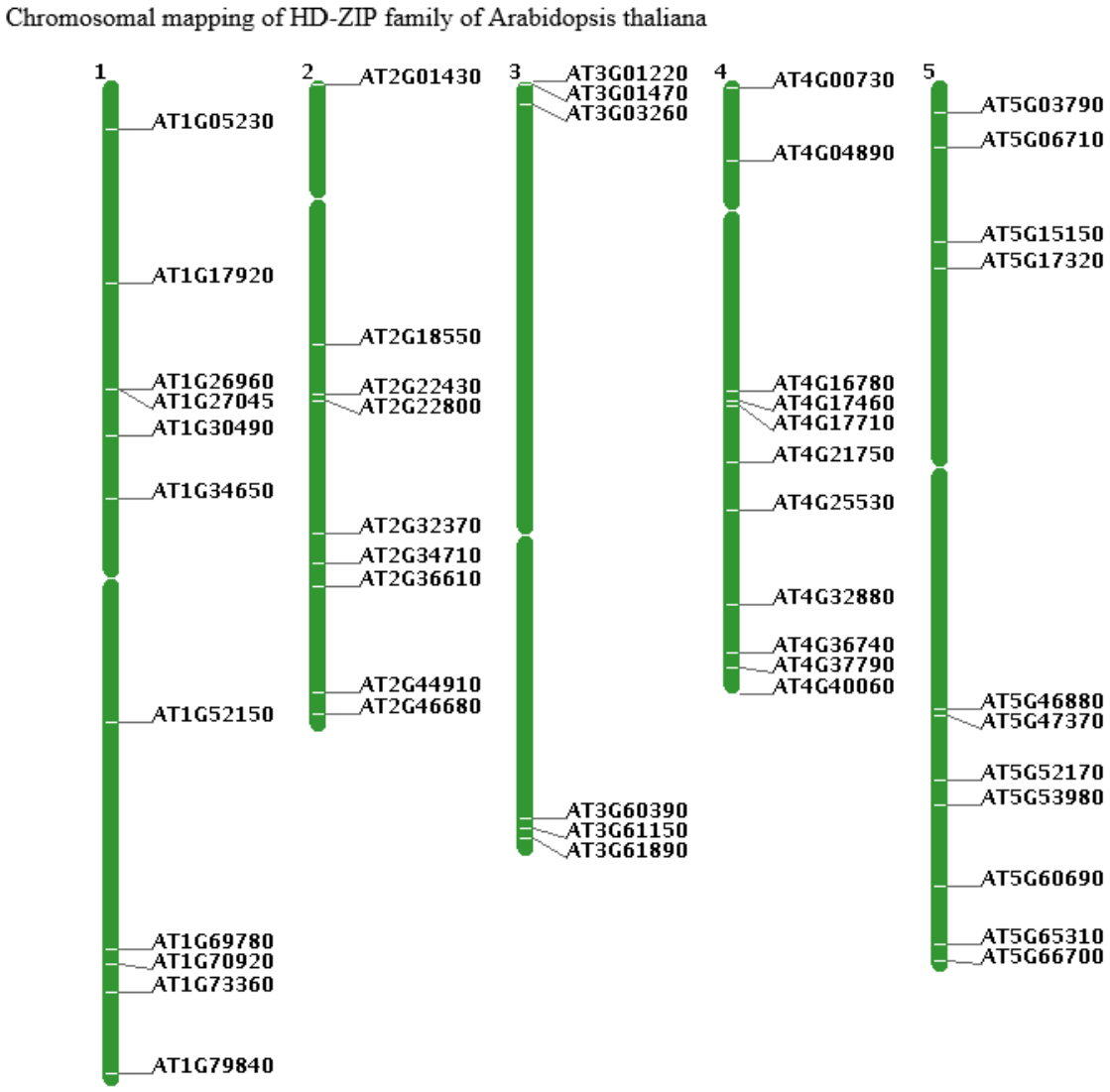

Figure 1. The mapping of 48 transcription factors of *Arabidopsis thaliana* on the chromosomes of *Arabidopsis*. The numbers of chromosomes are shown at the tip of each chromosome and transcription factors are mapped according to their positions on the chromosome. The length of each chromosome is indicating its size. The common names of transcription factors are shown in table

RESULTS AND DISCUSSION

Analysis of HD-ZIP gene family has been performed in model plant *Arabidopsis* (Chen *et al*., 2014). The *Arabidopsis thaliana* expression is highly regulated and show response to environmental stimuli, organ-specific signals, to light, and to the stress condition (Liu *et al*., 1996). All candidates of HD-ZIP family are usually examined to clarify the presence of LZ and HD domain. The *Arabidopsis thaliana* species were divided into four well-conserved subfamilies, HD-ZIP I to IV. HD-ZIP I and IV subfamilies have high number of genes in the plant species while HD-ZIP II subfamily have low number of genes as compared to I and IV. HD-ZIP III subfamily composed of the fewest number of genes except for moss (Hu *et al*., 2012).

Chromosomal mapping: In this study, a total of 48 genes from the genome of *Arabidopsis thaliana* were identified as the member of HD-ZIP transcription factors family. Different variants of these transcription factors were excluded (only those transcription factors were studied that have "1" value after decimal point e.g. AT3G01470.1). These 48 transcription factors were encoded 48 proteins and these factors were mapped on 5 chromosomes, as in figure 1, using online "chromosome map tool" from 'TAIR' database. *(http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp)* The first chromosome has 11 genes; 6 genes on short arm and 5 genes on the long arm. Chromosome 2 has 9 genes; 1 gene on the short arm and 8 genes on the long arm. Chromosome 3 has 6 genes; 3 genes on the short and 3 genes on the long arm. The forth chromosome has 11 genes; 2 genes on the short arm and 7 genes on the long arm and chromosome 5 also has 11 genes; 4 genes on the short arm and 7 genes on the long arm (Figure 1). Based on the information that we retrieved from 'Phytozome', the chromosomal location of all the 48 genes of HD-ZIP family of *Arabidopsis thaliana* are examined. The localization of the genes of this family has showed that these genes are unevenly distributed to some extent throughout all the five chromosomes of *Arabidopsis*. As 11 genes are localized on chromosome number 1, 4 and 5 respectively while 9 genes are on chromosome number 2 and 6 genes are on chromosome number 3. So, distribution of 48 genes seems like uneven to some extent (Figure 1).

Identification of conserved domain: All the 48 genes of HD-ZIP family of *Arabidopsis* were multiply aligned by using Unipro UGENE software and analyzed for the conserved domain in all the genes.

Table 1. Accession number, chromosome number, common names and consensus domain sequence with conserved WFQNrr box of all the transcription factors of HD-ZIP family of *Arabidopsis thaliana* is indicated in the table

Accession No.	Chromos	Common names	Domain Sequence
AT3G01470.1	Chr3	ATHB1	KKRRLTTEQVHLLEKSFETENKLEPERKTQLAKKLGLQPRQVAVWFQNRRARWKTK
AT1G69780.1	Chr1	ATHB13	KKRRLNMEQVKTLEKNFELGNKLEPERKMQLARALGLQPRQIAIWFQNRRARWKT
AT5G65310.1	Chr5	ATHB5	KKRRLGVEQVKALEKNFEIDNKLEPERKVKLAQELGLQPRQVAIWFQNRRARWKTK
AT2G22430.1	Chr2	ATHB6	KKRRLSINQVKALEKNFELENKLEPERKVKLAQELGLQPRQVAVWFQNRRARWKTK
AT4G40060.1	Chr4	ATHB16	KKRRLKVDQVKALEKNFELENKLEPERKTKLAQELGLQPRQVAVWFQNRRARWKT
AT5G15150.1	Chr5	ATHB3	KKKRLNLEQVRALEKSFELGNKLEPERKMQLAKALGLQPRQIAIWFQNRRARWKTK
AT5G06710.1	Chr5	HAT14	KKLRLSKDQSAFLEDSFKEHSTLNPKQKIALAKQLNLRPRQVEVWFQNRRARTKLK
AT2G46680.1	Chr2	ATHB-7	NQRRFSDEQIKSLEMMFESETRLEPRKKVQLARELGLQPRQVAIWFQNKRARWKSK
AT4G16780.1	Chr4	ATHB2,ATHB-2, HAT4, HB-	KKLRLSKDQSAILEETFKDHSTLNPKQKQALAKQLGLRARQVEVWFQNRRARTKLK
AT4G36740.1	Chr4	ATHB40, HB40, HB-5	RKRKLTDEQVNMLEMSFGDEHKLESERKDRLAAELGLDPRQVAVWFQNRRARWK
AT2G44910.1	Chr2	ATHB4, ATHB-4, HB4	KKLRLSKDQALVLEETFKEHSTLNPKQKLALAKQLNLRARQVEVWFQNRRARTKLK
AT4G17460.1	Chr4	HAT1, JAB	KKLRLSKDQSAVLEDTFKEHNTLNPKQKLALAKKLGLTARQVEVWFQNRRARTKL
AT3G60390.1	Chr3	HAT3	KKLRLSKEQALVLEETFKEHSTLNPKQKMALAKQLNLRTRQVEVWFQNRRARTKLK
AT3G01220.1	Chr3	ATHB20, HB20	KKKRLQLEQVKALEKSFELGNKLEPERKIQLAKALGMQPRQIAIWFQNRRARWKTR
AT1G26960.1	Chr1	AtHB23, HB23	KKRRLNMEQLKALEKDFELGNKLESDRKLELARALGLQPRQIAIWFQNRRARSKTK
AT2G18550.1	Chr2	ATHB21, HB-2, HB21	RKRKLSDEQVRMLEISFEDDHKLESERKDRLASELGLDPRQVAVWFQNRRARWKNK
AT2G22800.1	Chr2	HAT9	KKLRLTKQQSALLEESFKDHSTLNPKQKQVLARQLNLRPRQVEVWFQNRRARTKLK
AT3G61890.1	Chr3	ATHB12, ATHB-12, HB-12	NQKRFSEEQIKSLELIFESETRLEPRKKVQVARELGLQPRQVAIWFQNKRARWKTK
AT4G37790.1	Chr4	HAT22	KKLRLTKQQSALLEDNFKLHSTLNPKQKQALARQLNLRPRQVEVWFQNRRARTKLK
AT5G47370.1	Chr5	HAT2	KKLRLSKDQSAFLEETFKEHNTLNPKQKLALAKKLNLTARQVEVWFQNRRARTKLK
AT5G66700.1	Chr5	ATHB53, HB53, HB-8	RKRKLTDEQVNMLEYSFGNEHKLESGRKEKIAGELGLDPRQVAVWFQNRRARWKN
AT2G01430.1	Chr2	ATHB17, ATHB-17, HB17	KKLRLTREQSRLLEDSFRQNHTLNPKQKEVLAKHLMLRPRQIEVWFQNRRARSKLK
AT5G03790.1	Chr5	ATHB51, HB51, LMI1	KKKRLTSGQLASLERSFQEEIKLDSDRKVKLSRELGLQPRQIAVWFQNRRARWKAK
AT1G27045.1	Chr1	ATHB54, HB54	KKRKLTPIQLRLLEESFEEEKRLEPDRKLWLAEKLGLQPSQVAVWFQNRRARYKTK
AT1G70920.1	Chr1	ATHB18, HB18	KKLRLTKEQSHLLEESFIQNHTLTPKQKKDLATFLKLSQRQVEVWFQNRRARSKLK
AT2G36610.1	Chr2	ATHB22, HB22	QLKFLERSFQEEIKLNPDRKMKLNPDRKMKLSKELGLQPRQIAVWFQNRKARWKNK
AT5G53980.1	Chr5	ATHB52, HB52	KKKRLTQDQVRQLEKCFTMNKKLEPDLKLQLSNQLGLPQRQVAVWFQNKRARFK
AT1G79840.1	Chr1	GL2	KYHRHTTDQIRHMEALFKETPHPDEKQRQQLSKQLGLAPRQVKFWFQNRRTQIK
AT1G05230.1	Chr1	HDG2	RYHRHTQLQIQEMEAFFKECPHPDDKQRKQLSRELNLEPLQVKFWFQNKRTQMK
AT4G00730.1	Chr4	AHDP, ANL2	Absent
AT4G21750.1	Chr4	ATML1	RYHRHTPQQIQELESMFKECPHPDEKQRLELSKRLCLETRQVKFWFQNRRTQMK
AT4G04890.1	Chr4	PDF2	RYHRHTQRQIQELESFFKECPHPDDKQRKELSRDLNLEPLQVKFWFQNKRTQMK
AT4G25530.1	Chr4	FWA, HDG6	RTHRRTAYQTQELENFYMENPHPTEEQRYELGQRLNMGVNQVKNWFQNKRNLEK
AT4G32880.1	Chr4	ATHB8, ATHB-8, HB-8	YTPEQVEALERLYNDCPKPSSMRRQQLIRECPILSNIEPKQIKVWFQNRRCREKQRK
AT2G34710.1	Chr2	ATHB14, ATHB-14, PHB,	YTPEQVEALERVYTECPKPSSLRRQQLIRECPILSNIEPKQIKVWFQNRRCREKQRK
AT3G61150.1	Chr3	HDG1, HD-GL2-1	RYHRHTPKQIQDLESVFKECAHPDEKQRLDLSRRLNLDPRQVKFWFQNRRTQMK
AT1G30490.1	Chr1	ATHB9, PHV	YTPEQVEALERVYAECPKPSSLRRQQLIRECPILCNIEPRQIKVWFQNRRCREK
AT1G52150.1	Chr1	ATHB15, ATHB-15, CNA,	YTPEQVEALERLYHDCPKPSSIRRQQLIRECPILSNIEPKQIKVWFQNRRCREK
AT1G73360.1	Chr1	ATHDG11, EDT1, HDG11	RYHRHTAQQIQRLESSFKECPHPDEKQRNQLSRELGLAPRQIKFWFQNRRTQLK
AT2G32370.1	Chr2	HDG3	KYNRHTQLQISEMEAFFRECPHPDDKQRYDLSAQLGLDPVQIKFWFQNKRTQNK
AT5G52170.1	Chr5	HDG7	KYHRHTSYQIQELESFFKECPHPNEKQRLELGKKLTLESKQIKFWFQNRRTQMK
AT1G17920.1	Chr1	HDG12	RFHRHTPHQIQRLESTFNECQHPDEKQRNQLSRELGLAPRQIKFWFQNRRTQKK
AT1G34650.1	Chr1	HDG10	NRRHHSNHQVQRLEAFFHECPHPDDSQRRQLGNELNLKHKQIKFWFQNRRTQAR
AT3G03260.1	Chr3	HDG8	TCHRHTPQQIQRLEAYFKECPHPDERQRNQLCRELKLEPDQIKFWFQNKRTQSK
AT4G17710.1	Chr4	HDG4	RYHRHTASQIQQMEALFKENAHPDTKTRLRLSKKLGLSPIQVKFWFQNKRTQIK
AT5G60690.1	Chr5	IFL, IFL1, REV	YTAEQVEALERVYAECPKPSSLRRQQLIRECSILANIEPKQIKVWFQNRRCRDK
AT5G46880.1	Chr5	HB-7, HDG5	RYHRHTNRQIQEMEALFKENPHPDDKQRKRLSAELGLKPRQVKFWFQNRRTQMK
AT5G17320.1	Chr5	HDG9	GYHRHTNEQIHRLETYFKECPHPDEFQRRLLGEELNLKPKQIKFWFQNKRTQAK

The signature box named "WFQNrr" was identified by UGNENE in which "WFQN" residues were highly conserved while "rr" residues were semi-conserved. When protein sequences of all the 48 genes of HD-ZIP family were compared to identify the whole domain, the total of 44 residues before the WFQNrr and 6 residues after the WFQNrr were identified as conserved with few alternative residues in some genes. So, the Homeo domain of 56 amino acid residues was identified (Figure 2). Table 1 represents the conserved homeo domain sequence in all the transcription factors of HD-ZIP family of *Arabidopsis*. In this study, we analyzed the HD-ZIP transcription factors family of *Arabidopsis thaliana* in order to determine its conserved functional domain. For that purpose, we retrieved the amino acid sequences of all the 48 genes and then multiply aligned them by Unipro UGENE. Amino acids 'W', 'F', 'Q', 'N' were identified as highly conserved while 'r' and 'r' amino acids were identified as semi-conserved with adjacent position in every gene. These amino acids were not present on the same position in all genes. They were present in different position in every gene but where ever they found, they found as a box which means that these amino acids were present as a conserved box in the domain in all the genes. Beyond this, 44 residues at upstream side and 6 residues at downstream side of this box were identified as conserved in all genes with few alternatives and thus it was investigated that the complete domain was present in distributed form and on different position in every gene. So, in order to draw a consensus sequence of the domain of all the genes as well as to identify the complete domain as a whole, we delete the amino acid residues one by one from both; the upstream and the downstream side of WFQNrr from each gene to bring this box on the same position in every gene. By this, WFQNrr were placed on the same position in every gene when there were only 44 residues before this box and 6 residues after this box. Thus, Homeo domain of 56 residues was identified in all genes having WFQNrr amino acids on position number 45, 46,47,48,49 and 50 respectively. A consensus sequence of domain was identified as in table 1 (Figure 2).

Figure 2. Multiple sequence alignment of 48 transcription factors of *Arabidopsis thaliana* by UGENE software to identify the conserved domain in all transcription factors. The signature sequence WFQNrr is identified by UGENE and these amino acids are shown in figure by " *" and " : " shape. The sequence of domains of all transcription factors consisting of 56 amino acids are shown in table

Phylogenetic analysis: ClustalW was used for the multiple sequence alignment of the protein sequences of all 48 HD-ZIP genes of *Arabidopsis.* After alignment, a phylogenetic tree was constructed with neighbour joining method by using MEGA 6.0 software. A total of 4 sub families were observed in the phylogenetic tree. All the 4 sub families have conserved homeobox domain (HD). HD-ZIP subfamily I, II, III and IV have 17, 10, 5 and 16 genes respectively and each gene in each subfamily was categorized in different clades/classes according to their ancestor and evolutionary relationship as well as functions (Figure 3).

It was concluded from the phylogenetic analysis that these specific subfamilies of HD-ZIP played specific role in specific type of responses and also in the development of *Arabidopsis thaliana*. For phylogenetic analysis, an unrooted tree was generated from all the 48 genes of HD-ZIP transcription factor family of *Arabidopsis thaliana* (Figure 3). The HD-ZIP I subfamily has 17 members and was divided into 7 clades that were named as clade α, β1, β2, Φ, ε, γ and δ (Figure 3). The proteins of HD-ZIP I subfamily were identified by the existence of a Homeo Domain (HD) that was closely linked to the Leucine Zipper (LZ) motif.

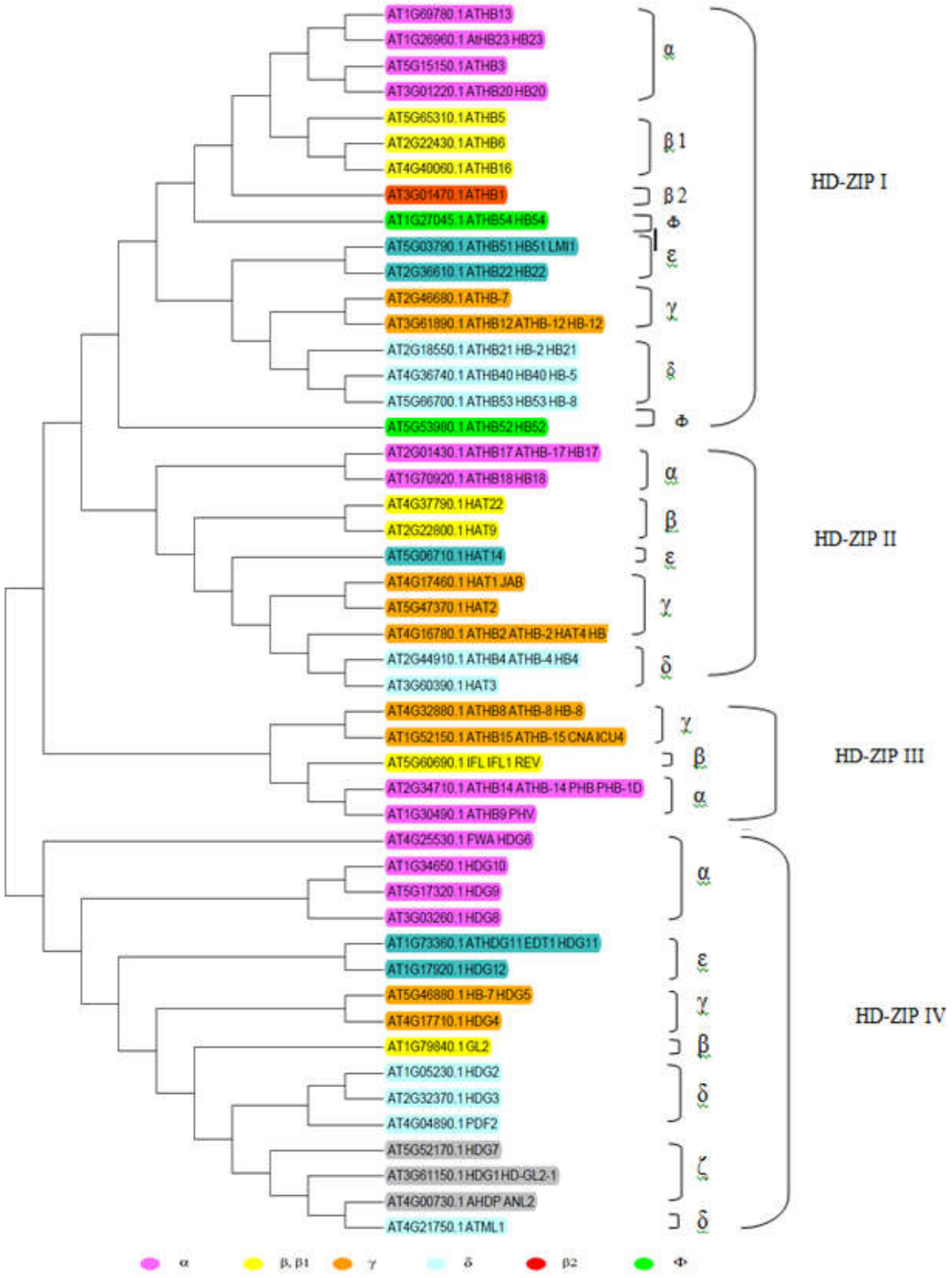

Figure 3. Rectangular phylogenetic tree of HD-ZIP family of *Arabidopsis thaliana* indicating the division of HD-ZIP family into four sub families on the basis of evolution and function and also showing the transcription factors of each sub family that are further categorized into clades

It was reported that some proteins of the HD-ZIP I family were involved in the sucrose signaling and abscisic acid pathways. These proteins also played a critical role in the abiotic stress responses of the plants and cotyledon as well as involved in the leaf development and embryogenesis (Himmelbach *et al*., 2002; Johannesson *et al*., 2003). The HD-ZIP II subfamily has 10 members that were divided into 5 clades and named as clade α, β, ε, γ and δ (Figure 3). Proteins of all the members of this family have a conserved set of cysteine molecules inside and the LZ motifoutside (Tron *et al*., 2002). Most of the genes of HD-ZIP II subfamily was incriminated in phytochrome-mediated organ development e.g. leaf morphogenesis (Ciarbelli *et al*., 2008). Some genes were also responsive to the changes of light quality and auxin and shade avoidance as exhibit by biochemical and genetic analyses (Morelli and Ruberti 2002; Sawa *et al*., 2002). The HD-ZIP III and HD-ZIP IV subfamilies were characterized by the presence of two additional domains; first domain was steroidogenic acute regulatory protein related lipid transfer (START) and second

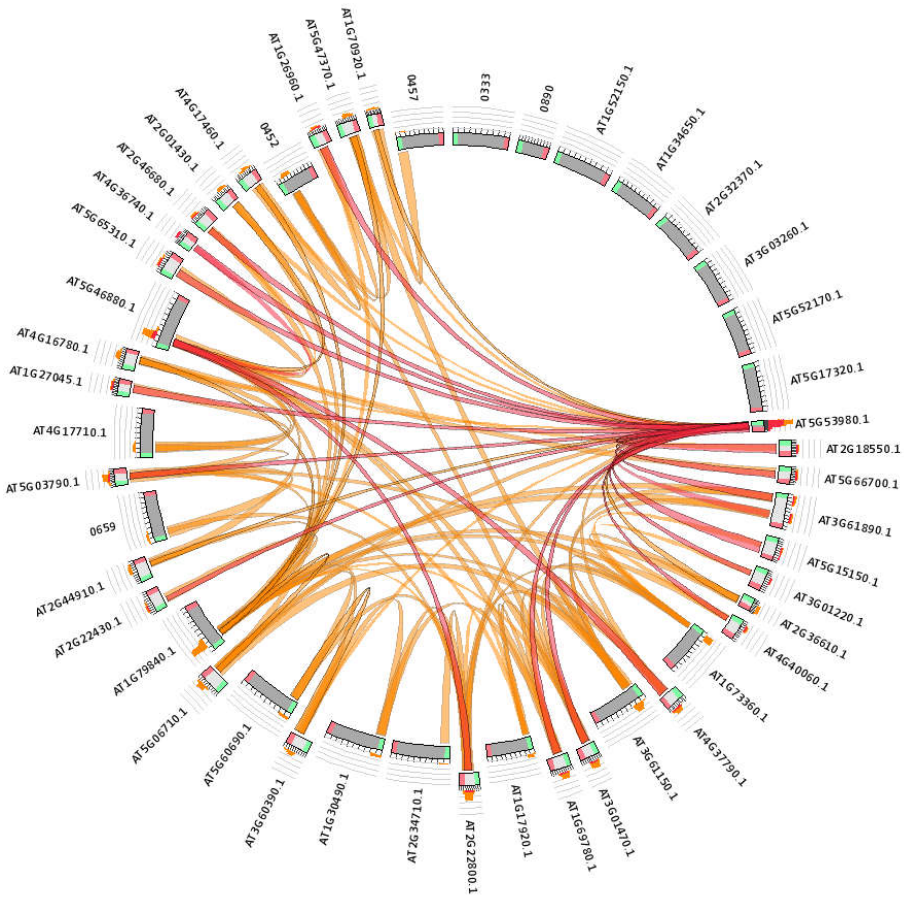

Figure 4. Synteny analysis; synteny relationship among all the 48 transcription factors of HD-ZIP family of *Arabidopsis thaliana*

domain was START-adjacent domain (SAD). These two subfamilies were distinguished by a fifth domain that was the C-terminal MEKHLA motif. This domain was present in HD-ZIP III subfamily but it was absent in HD-ZIP IV subfamily (Mukherjee and Bürglin 2006). The HD-ZIP III subfamily has 5 members that were divided into 3 clades and named as clade α, β and γ (Figure 3). The members of HD-ZIP III subfamily were the key developmental regulators for the *Arabidopsis* shoot radial patterning and apical embryo. They were also involved in the development of lateral organ polarity, formation of shoot meristem, vascular differentiation and transportation of auxin (Baima *et al*., 2001; Emery *et al*., 2003; McConnell *et al*., 2001; Ohashi-Ito and Fukuda 2003; Prigge *et al*., 2005). The HD-ZIP IV subfamily has 16 members that were divided into 6 clades and named as clade α, β, ε, γ, δ and ζ (Figure 3). The proteins of the HD-ZIP IV subfamily were played role in epidermal processes, formation of trichome, accumulation of anthocyanin and development of root (Javelle *et al*., 2011). This provides an overall picture about the composition and classification of all genes of HD-ZIP transcription factors family.It also help to understand the evolutionary relationship among all the genes in more detail as each gene is further classified into different clades. This will help in the further functional and comparative annotation of all the transcription factors of HD-ZIP family.

Synteny analysis

Synteny relationship among all the 48 transcription factors of HD-ZIP family of *Arabidopsis thaliana* was observed in a circular figure as colour bars as in Figure 4. Synteny analysis revealed tandem duplication and segmental duplication events in most of the transcription factors within the HD-ZIP family. It was observed that transcription factors AT5G17320.1, AT5G52170.1, AT3G03260.1, AT2G32370.1, AT1G34650.1, AT1G52150.1, 0890 and 0333 have no synteny relationship (Figure 4). Synteny analysis describe the tanden, whole genome and segmental duplication events that have significance in the evolution of many organisms(Xu *et al*., 2012). Genome duplication events played a significant role in the expansion of family and in *arabidopsis* there are three duplication events (Maere De Bodt *et al*., 2005). Based on the comprehensive analysis of chromosomal mapping, gene structure analysis, phylogenetic analysis and motif analysis, we observed segmental and tandem duplication in most of the transcription factors of all the 4 subfamilies of HD-ZIP transcription factors family of *Arabidopsis thaliana*. It is confirmed by the synteny analysis of HD-ZIP family where duplication events were shown by color ribbons. Transcription factor "AT5G53980.1" that fall in Φ clade of HD-ZIP I subfamily has been observed to play a prominent role in these duplication events. It was also observed that some transcription factor were not involved in segmental and tandem duplication events and thus these factors have their unique sequences and belonged to HD-ZIP III and HD-ZIP IV subfamily (Figure 4).

Analysis of Cisregulatory elements in promoter: Actually the transcriptional factors are keys in a sense that open the lock of expression of genes in stress conditions like biotic and abiotic stresses. The first step for promoter analysis was the retrieval of 1000 bp upstream to the start codon for each gene of HD-ZIP family of *A. thaliana* .then data was put into the PLACE database to find out the transcription factor binding

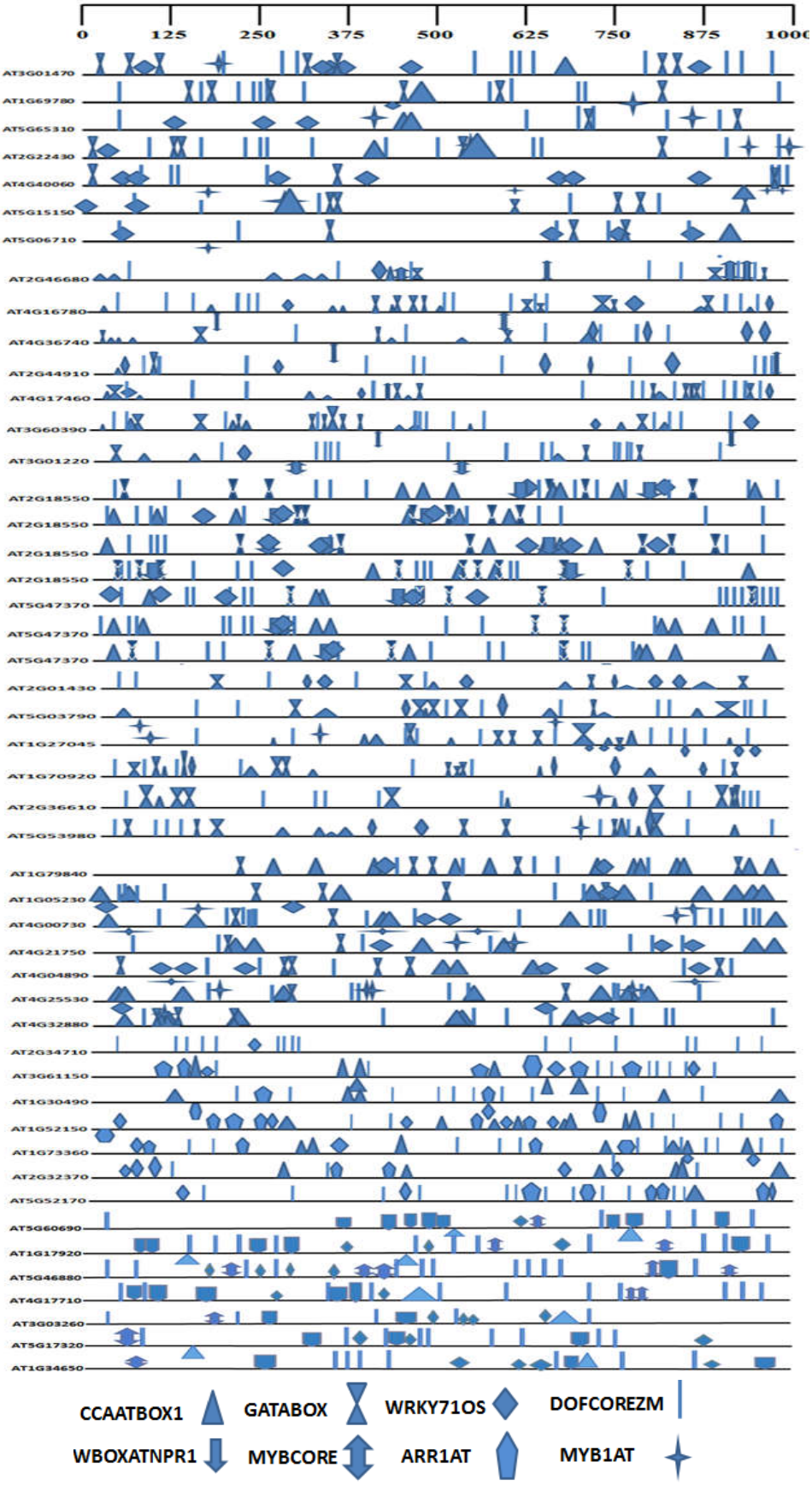

Fig. 5. Promoter analysis; schematic representation of various PLACE-based motifs in 1KB promoter region of HD-ZIP family. Most of motifs present on +strand is shown in the figure. The motif codes and respective sequences are indicated in the figure

sites (motifs) in the promoter region .we selected five most abundant cis regulatory elements for each nucleotide sequence in all members of HD-ZIP. Three out of five cis regulatory elements named as CCAATBOX1, WORKY710S & DOFCOREZM were present in all members of HD-ZIP family, moreover they were present on positive strand of DNA. GATABOX was present almost 85% in promoter region of 48 genes while there was variability at 5th cis regulatory element as in Figure (5), which was mostly present on negative strand of DNA. Table 2 represents the selected regulatory elements, their repetition on promoter, their presence on strand + or – and their positions on strand.

Table 2. Accession number of all the 48 transcription factors of HD-ZIP family of *Arabidopsis thaliana* and the factors name (regulatory elements), sequence of the factors, their presence on strand + or - , their repetition and location on promoter is indicated in this table

Sr. No.	Accession No.	Factor Name	Sequence of factor	Strand	Repetition	Location on Promoter
1	AT3G01470	CCAATBOX1	CCAAT	+	1	656
		WRKY71OS	TGAC	+	6	103,320,352,368,447,861
		DOFCOREZM	AAAG	+	12	179,209,270,309,549,601,605,629,791,897, 919,966
		GATABOX	GATA	+	7	26,56,115,312,361,814,842
		MYB1AT	WAACCA	+	1	187
2	AT1G69780	CCAATBOX1	CCAAT	+	1	445
		WRKY71OS	TGAC	-	1	497
		DOFCOREZM	AAAG	+	12	50,156,219,236,242,262,321,574,605,693,700,985
		GATABOX	GATA	+	6	151,180,252,452,580,805
		MYB1AT	WAACCA	-	1	762
3	AT5G65310	CCAATBOX1	CCAAT	+	2	422,427
		WRKY71OS	TGAC	+	3	125,260,326
		DOFCOREZM	AAAG	+	6	46,623,692,716,828,891
		GATABOX	GATA	+	2	713,931
		MYB1AT	WAACCA	+	2	407,859
4	AT2G22430	CCAATBOX1	CCAAT	+	2	397,542
		WRKY71OS	TGAC	+	1	45
		DOFCOREZM	AAAG	+	13	33,96,166,231,251,255,326,422,499,628,640,904,982
		GATABOX	GATA	+	5	6,129,136,527,818
		MYB1AT	WAACCA	+	3	539,932,990
5	AT4G40060	CCAATBOX1	CCAAT	-	1	914
		WRKY71OS	TGAC	+	7	49,77,281,377,675,687,856
		DOFCOREZM	AAAG	+	7	83,127,140,262,972,976,997
		GATABOX	GATA	+	3	10,268,979
		MYB1AT	WAACCA	-	4	170,618,962,987
6	AT5G15150	CCAATBOX1	CCAAT	+	1	277
		WRKY71OS	TGAC	+	2	2,88
		DOFCOREZM	AAAG	+	5	65,158,339,681,813
		GATABOX	GATA	+	6	357,378,610,753,793,919
		MYB1AT	WAACCA	+	1	274
7	AT5G06710	CCAATBOX1	CCAAT	+	1	912
		WRKY71OS	TGAC	+	4	74,651,752,870
		DOFCOREZM	AAAG	+	5	53,217,667,736,861
		GATABOX	GATA	+	3	339,700,764
		MYB1AT	WAACCA	-	1	173
8	AT2G46680	CAATBOX1	CAAT	+	6	11,19,252,284,298,413
		WRKY71OS	TGAC	+	1	384
		DOFCOREZM	AAAG	+	7	83,319,452,797,808,972,983
		GATABOX	GATA	+	3	476,878,969
		MYBCORE	CNGTTR	+	1	667
9	AT4G16780	CAATBOX1	CAAT	+	10	39,173,320,326,371,440,123,596,855,962
		WRKY71OS	TGAC	+	4	261,558,770,976
		DOFCOREZM	AAAG	+	14	77,126,158,228,234,244,510,534,635,668,678,911,919,9
		GATABOX	GATA	+	9	389,407,486,492,640,691,750,758,870
		MYBCORE	CNGTTR	-	2	178,606
10	AT4G36740	CAATBOX1	CAAT	+	6	129,148,159,426,581,724
		WRKY71OS	TGAC	+	4	727,797,920,955
		DOFCOREZM	AAAG	+	6	280,464,697,737,789,801
		GATABOX	GATA	+	4	22,169,362,633
		MYBCORE	CNGTTR	-	1	342
11	AT2G44910	CAATBOX1	CAAT	+	1	64
		WRKY71OS	TGAC	+	5	97,253,664,738,818
		DOFCOREZM	AAAG	+	11	103,132,235,399,414,456,611,762,954,974,979
		GATABOX	GATA	+	1	119
		MYBCORE	CNGTTR	+	1	989
12	AT4G17460	CAATBOX1	CAAT	+	6	14,118,273,327,462,817
		WRKY71OS	TGAC	+	3	66,352,958
		DOFCOREZM	AAAG	+	15	97,155,209,364,418,704,788,798,835,850,869,875,887,9
		GATABOX	GATA	+	7	38,447,480,791,801,853,906
		MYBCORE				
13	AT3G60390	CAATBOX1	CAAT	+	9	15,149,243,279,423,487,572,756,825
		WRKY71OS	TGAC	+	2	746,968
		DOFCOREZM	AAAG	+	13	38,49,205,263,449,467,482,515,609,783,842,858,905
		GATABOX	GATA	+	9	120,163,229,274,290,313,348,353,776
		MYBCORE	CNGTTR	-	2	431,953
14	AT3G220	CAATBOX1	CAAT	+	3	113,160,673
		WRKY71OS	TGAC	+	1	222
		DOFCOREZM	AAAG	+	13	193,302,320,335,341,548,641,685,736,744,757,763,892
		GATABOX	GATA	+	3	77,721,787
		MYBCORE	CNGTTR	-	2	580,264
15.	AT2G18550	CAATBOX1	CAAT	+	7	436,460,515
		WBOXATNPR1	TTGAC	+	2	625,848
		DOFCOREZM	AAAG	+	12	89,125,332,351
		GATABOX	GATA	+	6	59,206,250
		WRKY71OS	TGAC	+	2	626,849

…….Continue

16.	AT4G37790	CAATBOX1	CAAT	+	6	46,119,209
		WBOXATNPR1	TTGAC	+	2	260,498
		DOFCOREZM	AAAG	+	11	33,75,89,191,
		GATABOX	GATA	+	6	275,289,459
		WRKY71OS	TGAC	+	3	159,261 ,499
17.	AT1G26960	CAATBOX1	CAAT	+	5	23,573,641
		WBOXATNPR1	TTGAC	+	3	252,301,666
		DOFCOREZM	AAAG	+	8	59,85,104,115
		GATABOX	GATA	+	6	210,374,588
		WRKY71OS	TGAC	+	5	253,302,626
18.	AT2G22800	CAATBOX1	CAAT	+	4	382,529,598,939
		WBOXATNPR1	TTGAC	-	1	767
		DOFCOREZM	AAAG	+	12	58,64,132,207
		GATABOX	GATA	+	9	71,76,118,396
		WRKY71OS	TGAC	+	1	261
19.	AT5G47370	CAATBOX1	CAAT	+	3	101 ,351,359
		WBOXATNPR1	TTGAC	+	2	96,449
		DOFCOREZM	AAAG	+	16	71,132,139,154
		GATABOX	GATA	+	5	235,481,509
		WRKY71OS	TGAC	+	5	35 ,97,165
20.	AT5G66700	CAATBOX1	CAAT	+	7	22,79,324,467
		WBOXATNPR1	TTGAC	+	1	252
		DOFCOREZM	AAAG	+	15	7,93,97,164,175
		GATABOX	GATA	+	2	631,650
		WRKY71OS	TGAC	+	1	253
21.	AT3G61890	CAATBOX1	CAAT	+	7	21,279,343,787,
		WBOXATNPR1	TTGAC	+	1	353
		DOFCOREZM	AAAG	+	10	113,151,363,492
		GATABOX	GATA	+	4	61,252,402,680
		WRKY71OS	TGAC	+	1	354
22	AT2G01430	CAATBOX1	CAAT	+	4	386,619,765,884
		WRKY71OS	TGAC	+	6	259,276,473,755,800,849
		DOFCOREZM	AAAG	+	4	39,47,209,295
		GATABOX	GATA	+	5	160,352,409,680
		MYB1AT	WAACCA	+	1	424
23	AT5G03790	CAATBOX1	CAAT	+	6	43,282,363,613,690,868
		WRKY71OS	TGAC	+	2	401,505
		DOFCOREZM	AAAG	+	11	68,100,186,383,461,634,804,810,988,992,996
		GATABOX	GATA	+	5	248,393,429,735,969
		MYB1AT	WAACCA	-	2	55,166
24	AT1G27045	CAATBOX1	CAAT	+	7	234,314,362,475,787,830,971
		WRKY71OS	TGAC	-	6	708,736,785,889,913,942
		DOFCOREZM	AAAG	+	10	145,213,434,455,496,613,796,852,880,940
		GATABOX	GATA	+	6	123,370,509,539,553,656
		MYB1AT	WAACCA	+	2	841,846
25	AT1G70920	CAATBOX1	CAAT	+	5	177,207,252,603,752
		WRKY71OS	TGAC	+	5	122,137,613,704,863
		DOFCOREZM	AAAG	+	6	1,47,90,381,494,874
		GATABOX	GATA	+	9	38,84,106,181,214,367,371,497,936
		MYB1AT	WAACCA	+	1	173
26	AT2G36610	CAATBOX1	CAAT	+	3	49,588,772
		WRKY71OS	TGAC	+	1	768
		DOFCOREZM	AAAG	+	8	219,255,344,876,901,943,972,986
		GATABOX	GATA	+	8	23,101,109,349,763,873,879,964
		MYB1AT	WAACCA	+	2	729,3
27	AT5G53980	CAATBOX1	CAAT	+	7	227,257,262,279,771,791,942,
		WRKY71OS	TGAC	+	3	339,373,788
		DOFCOREZM	AAAG	+	9	5,41,72,84,291,545,704,724,895
		GATABOX	GATA	+	8	14,123,151,454,581,649,667,810
		MYB1AT	WAACCA	+	1	652
28	AT1G79840	CCAATBOX1	CAAT	+	12	268,293,400,516,542,726,787,792,845,853,958,995
		WRKY71OS	TGAC	+	2	395,739
		DOFCOREZM	AAAG	+	5	432,526,633,667,797
		GATABOX	GATA	+	5	219,454,491,612,940
29	AT1G05230	CCAATBOX1	CAAT	+	9	9,47,368,711,754,876,938,963,973
		WRKY71OS	TGAC	+	1	742
		DOFCOREZM	AAAG	+	7	51,58,65,101,666,691,805
		GATABOX	GATA	+	4	241,344,510,738
		MYB1AT	WAACCA	-	2	146,868
30	AT4G00730	CCAATBOX1	CAAT	+	6	27,152,420,424,679,983
		WRKY71OS	TGAC	+	2	486,510
		DOFCOREZM	AAAG	+	19	97,174,192,196,211,219,410,465,614,703,721,725,867,884,906,943,959,964,974
		GATABOX	GATA	+	2	185,351
		MYB1AT	WAACCA	+	1	847
31	AT4G21750	CCAATBOX1	CAAT	+	6	210,435,477
		WRKY71OS	TGAC	+	3	424,824,863
		DOFCOREZM	AAAG	+	7	68,176,392,573,772,809,854
		GATABOX	GATA	+	2	196,368
		MYB1AT	WAACCA	+	2	527,592

……Continue

32	AT4G04890	CCAATBOX1	CAAT	+	3	502,534,649
		WRKY71OS	TGAC	+	6	105,141,219,654,699,873
		DOFCOREZM	AAAG	+	5	176,247,349,849,924
		GATABOX	GATA	+	6	43,289,296,426,455,913
		MYB1AT	WAACCA	-	2	118,867
33	AT4G25530	CCAATBOX1	CAAT	+	8	50,55,137,287,559,730,773,807
		WRKY71OS	TGAC	-	2	48,654
		DOFCOREZM	AAAG	+	10	177,256,374,386,517,554,749,754,800,876
		GATABOX	GATA	+	2	278,685
		MYB1AT	WAACCA	+	4	183,394,400,776
34	AT4G32880	CCAATBOX1	CAAT	+	5	48,113,202,534,686,
		WRKY71OS	TGAC	+	2	718,743
		DOFCOREZM	AAAG	+	10	77,424,548,600.664,751,778,825,837,979
		GATABOX	GATA	+	3	106,125,196
		MYB1AT	WAACCA	+	1	109
35	AT2G34710	CCAATBOX1	CAAT	+	3	240,263,534
		WRKY71OS	TGAC	+	1	234
		DOFCOREZM	AAAG	+	17	41,119,127,133,167,272,277,283,295,647,692,760,767,868, 873,923,951
		ARR1AT	NGATT	+	3	297,710,934
36	AT3G61150	CCAATBOX1	CAAT	+	4	160,272,363,583
		WRKY71OS	TGAC	+	3	171,651,868
		DOFCOREZM	AAAG	+	8	188,395,718,757,761,805,824,846
		ARR1AT	NGATT	+	5	102,130,565,704,772
		AMYBOX1	TAACARA	+	1	623
37	AT1G30490	CCAATBOX1	CAAT	+	5	127,375,379,823,990
		WRKY71OS	TGAC	-	2	142,587
		DOFCOREZM	AAAG	+	11	205,278,428,501,516,565,590,629,717,752,880
		ARR1AT	NGATT	+	2	250,567
38	AT1G52150	CCAATBOX1	CAAT	+	7	289,502,577,603,671,683,782,
		WRKY71OS	TGAC	+	5	39,280,439,582,680
		DOFCOREZM	AAAG	+	6	371,426,800,818,901,939
		ARR1AT	NGATT	+	6	185,202,258,561,613,984
		AMYBOX1	TAACARA	-	1	17
39	AT1G73360	CCAATBOX1	CAAT	+	6	320,451,474,738,836,927
		WRKY71OS	TGAC	+	1	337
		DOFCOREZM	AAAG	+	14	85,143,169,518,594,612,732,778,801,841,886,895,972, 992
		ARR1AT	NGATT	+	3	95,211,637
		AMYBOX1	TAACARA	+	1	749
40	AT2G32370	CCAATBOX1	CAAT	+	5	276,455,725,841,995
		WRKY71OS	TGAC	+	5	57,64,101,686,728
		DOFCOREZM	AAAG	+	3	113,347,866
		ARR1AT	NGATT	+	2	354,420
41	AT5G52170	CCAATBOX1	CAAT	+	1	875
		WRKY71OS	TGAC	+	4	141,479,872,970
		DOFCOREZM	AAAG	+	13	170,292,421,484,601,608,658,692,718,723,773,815,865
		ARR1AT	NGATT	+	4	622,963,804,809
		AMYBOX1	TAACARA	+	1	714
42	AT5G60690	CCAATBOX1	CCAAT	-	2	535,786
		WRKY71OS	TGAC	+	1	606
		DOFCOREZM	AAAG	+	5	127,726,818,839,941
		GATABOX	GATA	+	8	376,430,458,488,505,744,794,
		MYBCORE	CNGTTR	+	1	448
43	AT1G17920	CCAATBOX1	CCAAT	-	2	172,445
		WRKY71OS	TGAC	+	3	372,491,674
		DOFCOREZM	AAAG	+	11	179.191,243,274,482,514.547,735,889.898,950
		GATABOX	GATA	+	5	87.96.226,277,935
		MYBCORE	CNGTTR	+	2	570,814
44	AT5G46880	CCAATBOX1	CCAAT	--	--	--
		WRKY71OS	TGAC	+	3	182,247,271
		DOFCOREZM	AAAG	+	12	43,66,241,464,422,490,495,607,610,631,637,880
		GATABOX	GATA	+	1	822
		MYBCORE	CNGTTR	+	4	221,396,409.915
45	AT4G17710	CCAATBOX1	CCAAT	+	1	486
		WRKY71OS	TGAC	+	2	275,416
		DOFCOREZM	AAAG	+	11	56,87,352,426,499,593,710,750,928,935,951
		GATABOX	GATA	+	5	59,98,146,355,368
		MYBCORE	CNGTTR	+	3	782,790,803
46	AT3G03260	CCAATBOX1	CCAAT	+	1	692
		WRKY71OS	TGAC	+	4	496,536,543,657
		DOFCOREZM	AAAG	+	5	26,220,414,526,706
		GATABOX	GATA	+	2	265,457
		MYBCORE	CNGTTR	+	1	192
47	AT5G17320	CCAATBOX1	CCAAT	-	1	157
		WRKY71OS	TGAC	+	3	383,447,876
		DOFCOREZM	AAAG	+	9	98,362,405,470,491,572,629,733,748
		GATABOX	GATA	+	3	311,440,718
		MYBCORE	CNGTTR	+	1	70
48	AT1G34650	CCAATBOX1	CCAAT	+	1	612
		WRKY71OS	TGAC	+	4	529,603,628,876
		DOFCOREZM	AAAG	+	3	372,376,397
		GATABOX	GATA	+	3	255,682,977
		MYBCORE	CNGTTR	+	1	78

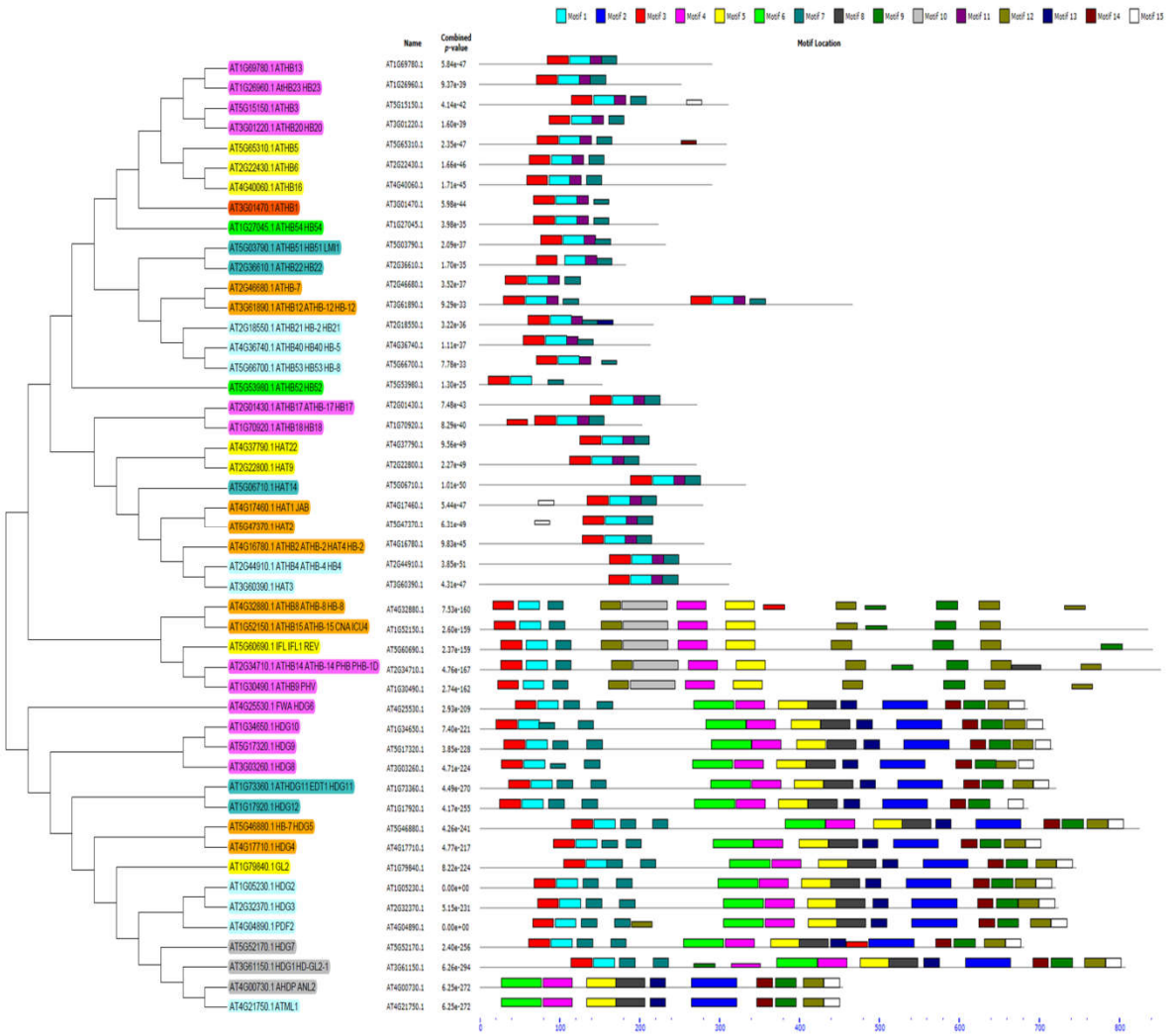

Fig. 6. Phylogenetic relationship and motif composition; Conserved motif analysis was done by using MEME online software and tree was constructed by using MEGA 6 software through multiple sequence alignment

CCAATBOX1, WORKY710S, GATABOX & DOFCOREZM were selected for analysis of cis regulatory elements. CCAATBOX1 cis regulatory element with site CCAAT was present in all members of HD-ZIP family. CCAAT Box regulates the flowering of *A.thaliana* and mostly present in non-coding region of heat shock proteins of eukaryotes. DOFCOREZM with site AAAG involved in tissue specific expression.WORKY710S is a stress responsive cis regulatory element and it was present in all members of HD-ZIP family of *Arabidopsis thaliana*. Promoter analysis of cis regulatory element had revealed that almost every gene contain repetitive binding sites of transcriptional factors on the promoter region (Figure 5).

Identification of conserved motifs in proteins domain: Conserved motif analysis has also revealed the homology between sub-families of HD-ZIP transcription factors. As sub-families I and II have same number of motifs and nearly same number of exons and similiarly sub-families III and IV have same number of motifs and exons. Thus besides the phylogenetic analysis, motif analysis also showed the evolutionary relationship between the HD-ZIP sub-families. Motif 1 links all the sub-families as it is conserved in all the transcription factors of all the 4 sub-families (Fig. 6). On the basis of occurrence of motifs we concluded that members of HD-ZIP family I and II are closely related and members of subfamily III and IV also have homology. Motif 1 was conserved in 47 members. Majority of members of this family almost first 28 members were containing 3-4 exons and few of remaining members containing 7-10 exons while there were 2-3 exceptional genes containing 17-18 exons which is showing their evolutionary relatedness and differences as in Figure (6).

Gene structure analysis: Gene structure analysis was used to map the exons, introns and UTR regions on the gene and thus to find out the location. These regions especially the coding regions that could be further used for functional annotation. The structure analysis of all HD-ZIP genes was done to have deep look into the evolution of HD-ZIP genes in *Arabidopsis thaliana*. The observations revealed that number of exons were 1 to 18 in each gene.AT2G46680.1,AT3G61890.1, AT5G53980.1 had just one exon, while there were three genes that contain 18 exons including: AT1G30490.1,AT1G52150.1 & AT5G60690.1.The conserved number of introns and exons in genes was an indication for evolutionary relatedness as shown in Figure 7. Gene structure analysis revealed similarities and differences between genes of HD-ZIP family of *Arabidopsis thaliana* to explain their evolutionary relationship (Liu *et al*., 2011). According to phylogenetic analysis HD-ZIP family was divided into four subfamilies on the basis of functional and structural similarities so to validate

Fig. 7. Gene structure analysis; Exon/intron structures of HD-ZIP genes from *Arabidopsis thaliana by an* online software "fancy gene". Colored boxes are showing exons while dotted line between boxes representing introns. location of intron/exon can b estimated by using scale

that fact we compared the results of gene structure analysis, phylogenetic analysis and conserved motif analysis. In the result of comparison, majority of Members of HD-ZIP subfamily I and II were rich in few motifs like motif 1,3,15 and 11 while there were few exceptions that contain repeats of these motifs. Members of HD-ZIP family III had 10 out of 15 motifs and members of HD-ZIP family IV were composed of almost all 15 motifs which were selected to perform conserved analysis as shown in Figure (6).

Conclusion

HD-ZIP proteins are a class of transcription factors family that has a conserved homeodomain in all of its factors. Here the present study concluded computational analysis of HD-ZIP transcription factors family in *Arabidopsis thaliana*, providing information about chromosomal mapping to locate the genes on chromosomes, motif analysis and conserved domain analysis. Conserved domain analysis showed that 44 residues were conserved before the signature box WFQNrr and 6 residues were conserved after the WFQNrr with few alternative residues in some genes, a total of 4 sub families were observed in the phylogenetic analysis. All the 4 sub families have conserved homeobox domain (HD). This classification and analysis further categorized the transcription factors of 4 HD-ZIP subfamilies into different clades and revealed a deep evolutionary relationship among them. Thus help to explore further functions of these factors. Gene structure analysis was performed to map the location of exons and introns in gene. Promoter analysis of cis regulatory element had revealed that almost every gene contain repetitive binding sites for transcriptional factors on the promoter region.

REFERENCES

Chen, X., Chen, Z., Zhao, H., Zhao, Y., Cheng, B. and Xiang, Y. 2014. Genome-Wide Analysis of Soybean HD Zip Gene Family and Expression Profiling under Salinity and Drought Treatments. *PloS one*., 9 : 87156.

Hu, R., Chi, X., Chai, G., Kong, Y., He, G., Wang, X., Shi, D., Zhang, D. and Zhou. G. 2012. Genome-wide identification, evolutionary expansion, and expression profile of homeodomain-leucine zipper gene family in poplar (Populus trichocarpa). *PloS one*., 7 : 31149.

Liu, H., Yang, W., Liu, D., Han, Y., Zhang, A. and Li, S. 2011. Ectopic expression of a grapevine transcription factor VvWRKY11 contributes to osmotic stress tolerance in Arabidopsis. *Molecular biology reports.,* 38 : 417-27.

Baima, S., Possenti, M., Matteucci, A., Wisman, E., Altamura, M. M., Ruberti, I. and Morelli, G. 2001. The Arabidopsis ATHB-8 HD-zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. *Plant Physiology.,* 126(2) : 643-655.

Ciarbelli, A. R., Ciolfi, A., Salvucci, S., Ruzza, V., Possenti, M., Carabelli, M., Fruscalzo, A., Sessa, G., Morelli, G. and Ruberti, I. 2008. The Arabidopsis homeodomain-leucine zipper II gene family: diversity and redundancy. *Plant molecular biology.,* 68(4-5) : 465-478.

Emery, J. F., Floyd, S. K., Alvarez, J., Eshed, Y., Hawker, N. P., Izhaki, A., Baum, S. F. and Bowman, J. L. 2003. Radial Patterning of Shoots by Class III HD-ZIP and KANADI Genes. *Current Biology.,* 13(20) : 1768-1774.

Himmelbach, A., Hoffmann, T., Leube, M., Höhener, B. and Grill, E. 2002. Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in Arabidopsis. *The EMBO journal.,* 21 : 3029-38.

Javelle, M., Klein-Cosson, C., Vernoud, V., Boltz, V., Maher, C., Timmermans, M., Depège-Fargeix, N. and Rogowsky, P. M. 2011. Genome-wide characterization of the HD-ZIP IV transcription factor family in maize: preferential expression in the epidermis. *Plant Physiology.,* 157(2) : 790-803.

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 molecular biology.,* 51 : 719-29.

Mcconnell, J.R., Emery, J., Eshed, Y., Bao, N., Bowman, J. and Barton, M.K. 2001. Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature.,* 411 : 709-13.

Morelli, G. and Ruberti, I. 2002. Light and shade in the photocontrol of growth. *Trends in plant science.,* 7 : 399-404.

Mukherjee, K. and Bürglin, T.R. 2006. MEKHLA, a novel domain with similarity to PAS domains, is fused to plant homeodomain-leucine zipper III proteins. *Plant Physiology.,* 140 : 1142-50.

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 and Cell Physiology.,* 44 : 1350-8.

Prigge, M.J., Otsuga, D., Alonso, J.M., Ecker, J.R., Drews, G.N. and Clark, S.E. 2005. Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development. *The Plant Cell Online.,* 17 : 61-76.

Sawa, S., Ohgishi, M., Goda, H., Higuchi, K., Shimada, Y., Yoshida, S. and Koshiba, T. 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. *The Plant Journal.,* 32(6) : 1011-1022.

Shaiq S., Muhammad A., Rana M., Farrukh A., Habibullah N., M.Hussnain S., Ertuğrul F., Khadim H., Amjad A. 2016. Genome-wide analysis of stress responsive WRKY transcription factors in Arabidopsis thaliana. *Turkish Journal of Agriculture-food Sciences and Technology*, 4(4): 279-290.

Tron, A.E., Bertoncini, C.W., Chan, R.L., Gonzalez, D.H. 2002. Redox regulation of plant homeodomain transcription factors. *Journal of Biological Chemistry.,* 277 : 34800-7.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America.,* 102(15) : 5454-5459.

Xu, G., Guo, C., Shan, H., Kong, H. 2012. Divergence of duplicate genes in exon–intron structure. *Proceedings of the National Academy of Sciences.,* 109 : 1187-92.

****