Pak. J. Agri. Sci., Vol. 53(2), 439-448; 2016
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI: 10.21162/PAKJAS/16.3347
http://www.pakjas.com.pk

GENOME-WIDE ANALYSIS OF TRIHELIX TRANSCRIPTION FACTOR GENE FAMILY IN *Arabidopsis thaliana*

Erum Yasmeen[1,2], Muhammad Riaz[1,2], Shaiq Sultan[1], Furrukh Azeem[1], Amjad Abbas[3], Kashif Riaz[3] and Muhammad Amjad Ali[3,4,*]

[1]Department of Bioinformatics & Biotechnology, Government College University, Faisalabad-38000, Pakistan; [2]National Centre for Bioinformatics (NCB), Quaid-i-Azam University, Islamabad, Pakistan; [3]Department of Plant Pathology, University of Agriculture, Faisalabad-38040, Pakistan; [4]Centre of Agricultural Biochemistry & Biotechnology, University of Agriculture-38040, Faisalabad, Pakistan.
[*]Corresponding author's e-mail: amjad.ali@uaf.edu.pk

Trihelix proteins are the members of gene family encoding transcriptional factors in plants that take part in plant responses to various cellular activities and stresses. The DNA-binding domain of these proteins is a tryptophan enrich tandem repeat forming helix-loop-helix-loop-helix. We retrieved the protein sequence of 28 candidates of trihelix gene family of *Arabidopsis thaliana*. These 28 proteins are grouped in five subfamilies according to their structural properties. These trihelix members were located on all 5 chromosomes of *Arabidopsis* with uneven distribution. We characterized diversity in amino acid residues in trihelix domain and found conserved motif in trihelix protein. Further, the gene structure analysis showed the distribution of introns and exons on each gene. The promoter analysis was done and 5 *cis*-regulatory elements were located on 1 kb of the promoter sequence. Synteny analysis showed the relationship among the trihelix genes. This study will be helpful in providing the *in silico* genomic information about the trihelix transcriptional factor in *Arabidopsis thaliana.* Moreover, these findings will be helpful in understanding trihelix family for their diverse role in plant stress and development
Keywords: Transcription factor, Trihelix domain, GT-1, *Arabidopsis*, conserved motifs.

INTRODUCTION

Transcriptional factors are the proteins that play an important role in the gene regulation by activating or repressing transcription of downstream target genes and consequently controlling many cellular activities during plant growth and development (Gao *et al.,* 2013; Ling *et al.,* 2011). This is achieved when DNA binding domain of transcriptional factors bind with specific sequences called *cis*-acting elements present in the promoters of the genes being regulated (Nuruzzaman *et al.,* 2012). The plants have more than 60 families of transcription factors with different functional activities (Kaplan-Levy *et al.,* 2012; Qin *et al.,* 2014). In plants, the transcriptional factors are involved in many processes i.e. plant growth and development as well as regulation of abiotic and biotic stress responses of the plants (Osorio *et al.,* 2012; Xie *et al.,* 2009). This paper discusses about analysis of trihelix transcription factor gene family in Arabidopsis.

The DNA-binding proteins characterized by the trihelix motif are solely present in plants. They were discovered in 1990s and are one of the first transcription factor gene family found in plants known as GT factors because of their binding properties with GT elements (Kaplan-Levy *et al.,* 2012; Qin *et al.,* 2014). Trihelix domain attracted the scientists because it is the only class of three spiro spin structure; it contains three tandem repeats helix - loop - helix - loop - helix (Luo *et al.,* 2012). The history of trihelix revealed that they were first identified in pea (*Pisum sativum)* nucleus later in soybean *(Glycine max* Merr.*),* tobacco *(Nicotiana tabacum* L.*),* rice *(oryza sativa),* and *Arabidopsis* (Luo *et al.* 2012). Although the trihelix family is confined to terrestrial plants but their existence in humans, animals and Drosophila needs investigation (Riaño-Pachón *et al.,* 2008). They are not present in the green algae (Chlorophyta) and have undergone massive expansion in the lineage of the common ascendant of terrestrial plants (Lang *et al.,* 2010). In the initial studies, the member of the gene family were thought to be involved in light-responsive gene regulation but the later studies highlighted their involvement in growth, development of tissues, embryo development, petal loss and plant organs, in biotic and abiotic stresses. Various genes of this family show variety of functions in plants like AT5G03680 (*PTL)* (Petal loss) gene is involved in morphological activities of flower organs (Kaplan-Levy *et al.,* 2012), AT1G54060 (ASIL1) and AT3G14180 (ASIL2) are reported to be involved in chlorophyll accumulation during embryo development and GT-1 is involved in light responses (Qin *et al.,* 2014). Trihelix factors bind to motifs called GT elements on the promoter DNA but the trihelix motif is only confined to the GT-1 and GT-2 DNA-binding proteins. Their genomic studies, functional studies and the structural pattern

uncovered the fact that they fall in 5 different clads; GTγ, SH4, SIP1, GT-1 and GT-2 (Kaplan-Levy *et al.*, 2012; Luo *et al.,* 2012).

The recent study has postulated that the trihelix DNA-binding domain is distantly associated to the Myb DNA-binding domain and also classified with myb/SANT-like domains in pfam as they all form a-helix-turn-a-helix structure (Nagano *et al.,* 2001; Qin *et al.,* 2014). Several studies have been carried out on the functional studies of trihelix family but not enough work is done on *in silico* genome wide analysis of trihelix family in *Arabidopsis thaliana.*

The objective of this study was to analyze trihelix gene family from Arabidopsis. Gene structure analysis, phylogenetic analysis, mapping of various trihelix genes on the chromosomes, conserved domain analysis, Synteny analysis, and analysis of *cis*- regulatory elements, have been carried out using various bioinformatics approaches.

MATERIALS AND METHODS

***Database search and sequence retrieval*:** The accession numbers of *Arabidopsis thaliana* trihelix family genes were retrieved from plant transcriptional factor database PlantTFDB http://planttfdb.cbi.pku.edu.cn (Jin *et al.*, 2013). PlantTFBD is an integrative database that gives complete list of transcriptional factors in plant species. The protein sequence of 28 trihelix genes were retrieved from the database excluding different variants. The promoter sequence 1Kb upstream to the start codon was retrieved from Phytozome for the analysis of *cis*-regulatory elements (http://www.phytozome.net/) (Ramirez and Basu, 2009).

***Chromosome mapping of trihelix genes*:** The physical location of each trihelix gene was demonstrated. The chromosomal mapping of all 28 trihelix genes on 5 chromosomes of *Arabidopsis thaliana* was done using chromosome map tool at TAIR (http://www.arabidopsis.org/ jsp/ ChromosomeMap/tool.jsp) (Nuruzzaman *et al.,* 2012)

***Identification of conserved motifs in proteins domain*:** In order to identify protein conserved domain, we used the MEME Suite (motif based sequence analysis tool) (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) (Bailey, 2009) to elucidate motifs and conserved regions shared among 28 trihelix proteins. MEME analysis was executed with following parameters: number of repetitions-any number; maximum number of moifs-20; the optimum motif widths were constrained to between 20 and 90 residues. Furthermore, to verify the conserved region, the multiple sequence alignment was done using CLUSTALW at UNIPRO UGENE software (Okonechnikov, 2012).

***Phylogenetic analysis*:** To explore the evolutionary relationships among the trihelix genes, a Neighbour Joining method-based phylogenetic tree was constructed using molecular evolution genetic analysis (MEGA) package. For this purpose, the multiple sequence alignment of identified trihelix amino acid sequence was performed by CLUSTALW program with default parameters. The unrooted tree was generated using Neighbour joining (NJ) strategy which summarizes the evolutionary distances between 28 members of trihelix gene family. The tree was visualized and presented in circular form and the closely related sequences were grouped in five clads according to Kalpan-Levy *et al*. (2012).

***Gene structure analysis*:** TAIR, (www.arabidopsis.org/), was used to extract and illustrate exon, intron and UTR organization for individual trihelix gene. These gene features were then mapped at gene structure display server GSDS (Hu, 2015).

***Promoter analysis*:** For promoter analysis, we retrieved 1Kb nucleotide sequence upstream to the start codon for all 28 genes using Phytozome database (http://www.phytozome.net/). These were then subjected to PLACE cis-regulatory elements database for the identification of already experimentally defined motifs (Higo *et al.,* 1999). Five cis-regulatory elements were obtained, they were CAATBOX, DOFCOREZEM, W-BOX, GT1-BOX and ARR1AT, and their conserved sequences were CAAT, AAAG, TGAC, GRWAAW and NGATT respectively. Their location was than mapped manually.

***Synteny analysis*:** Circoletto (http://tools.bat.infspire.org/ circoletto/), was used to perform Synteny analysis and visualize sequence identity among the trihelix family genes to study the sequence similarity patterns (Darzentas, 2010).

RESULTS

***Identification of trihelix gene family members and their chromosomal mapping*:** In this study, a total of 28 genes from Arabidopsis genome were identified as the members of trihelix transcription factor family. The accession number, chromosomal location, genomic and CDS length, peptide length and other related information of the identified genes are shown in Table 1. The chromosomal localization studies revealed the uneven distribution of these 28 candidates on all the 5 chromosomes of *A. thaliana* (Fig. 1). The first chromosome contains 7 genes; 4 on the short arm while 3 on the long. All the 4 trihelix genes of chromosome number 2 are located on the long arm. Chromosome 3 has 10 genes; 8 on the short arm and 2 on the long. The forth chromosome harbors only a single gene. The last chromosome hosts 6 genes, 4 on the short arm and 2 on the long arm.

***Phylogenetic analysis and identification of conserved motifs*:** The phylogenetic relationship of trihelix transcriptional factor family was examined by multiple sequence alignment of their amino acid sequences using clustalW and a tree was generated using MEGA 6 software with Neighbour joining method (Tamura *et al.,* 2013). There are no evidences that which sequence evolved first. The alignment showed that tryptophan (W) is highly conserved in all the sequences while phenylalanine (F), leucine (L) and

Table 1. Trihelix transcription factor genes of *Arabidopsis thaliana*, details e.g. chromosome number, number of exons in each gene, protein length, trihelix domain and genomic length, CDS sequence, start and ending position of genes are indicated in the table. These gene are distributed in five clads.

Accession number	Chr#	Exons	Protein length (AA)	Protein Mol. Weight	Isoelectric Point (pI)	Domain size	Start & End Position of Each Gene on genome	gDNA length	CDS length	Clad
AT1G13450	1	5	406	46675.9	6.8667	87-168	4612731..4615205	2475	1221	GT1
AT1G21200	1	1	443	50933.3	6.287	118-212	7421217..7423143	1927	1332	GTY
AT1G31310	1	2	383	42589.1	9.4992	19-106	11198353..1120014	1789	1152	SH4
AT1G33240	1	3	669	74212.7	5.8532	62-146, 435-519	12051471..12054546	3076	2010	GT2
AT1G54060	1	1	383	41731.6	9.064	93-179	20180679..20182324	1646	1152	SIP1
AT1G76870	1	1	385	44957.2	7.0675	87-178	28857250..28858407	1158	1158	GTY
AT1G76880	1	3	603	67879	6.6967	61-145, 408-493	28865500..28868225	2726	1812	GT2
AT2G33550	2	3	314	34861.3	6.0421	38-131	14210032..14211588	1557	950	SH4
AT2G35640	2	2	340	38305.2	9.0497	22-104	14982835..14984182	1348	1023	SH4
AT2G38250	2	2	289	34307.4	7.0249	41-123	16018357..16019500	1144	870	GT1
AT2G44730	2	1	372	40781.8	9.4426	63-154	18437333..18438565	1233	1119	SIP1
AT3G10000	3	2	481	55502.6	7.9162	88-174	3076874..3078907	2034	1446	GT2
AT3G10030	3	7	542	59247.8	6.558	160-246	3092023..3094945	2923	1629	SIP1
AT3G10040	3	1	431	48870.3	8.063	105-212	3096415..3098071	1657	1296	GTY
AT3G11100	3	2	249	28385.8	5.1728	22-102	3476187..3477405	1219	750	SIP1
AT3G14180	3	1	443	48314.1	10.2708	83-170	4707113..4708848	1736	1332	SIP1
AT3G24490	3	1	333	39083	4.3696	90-174	8910770..8912196	1427	1002	SIP1
AT3G24860	3	1	310	35191.4	9.7586	65-145	9073623..9074682	1060	933	SIP1
AT3G25990	3	5	372	42782.9	5.7598	55-139	9504677..9506787	1942	1119	GT1
AT3G54390	3	2	296	33280.8	10.0601	37-120	20137741..20139142	1402	891	SIP1
AT3G58630	3	2	321	36046.4	9.9284	25-124	21683568..21685941	2374	966	SIP1
AT4G31270	4	2	294	33369.4	4.591	19-97	15183188..15184961	1774	885	SH4
AT5G01380	5	2	323	38272.7	6.523	51-133	155639..157601	1963	972	GT1
AT5G03680	5	2	591	66638.8	7.0831	119-204, 421-506	957744..961032	3289	1776	GT2
AT5G05550	5	2	246	28191.8	9.0214	24-106	1639032..1640606	1575	750	SIP1
AT5G28300	5	2	619	71280.1	7.2368	103-172, 459-553	10292651..10295283	2633	1860	GT2
AT5G47660	5	2	398	45575.1	5.9406	303-384	19313008..19314636	1629	1197	GTY
AT5G63420	5	17	911	100554	8.3897	820-900	25400386..25405968	5583	2736	GT1

cysteine (C) are also conserved in many sequences (Fig. 2). The tandem repeat domain of hydrophobic core of trihelix contains two tryptophan residues while the third residue involved in core formation is either tryptophan or phenylalanine. According to Luo *et al.* (2012), the trihelix genes having phenylalanine as a third residue, belong to the GT-2 and GTγ subfamily. While those that contain isoleucine as a third residue are the members of SIP1 subfamily.

Two motifs were predicted as a part of trihelix domain using MEME Suite (motif based sequence analysis tool) (http://meme.nbcr.net/meme/cgi-bin/meme.cgi). Both of them were found to be highly conserved as shown in Fig. 3. It was predicted that two tryptophan residues are highly conserved in most of the sequences. According to Kaplan-levy *et al*. (2012), these tryptophan residues are involved in the formation of helixes with the existence of one tryptophan per each helix. In third helix formation; tryptophan, phenylalanine or isoleucine are involved as indicated with the help of star in motif 1 (Fig. 2B). The motif 1 also shows that cysteine is another conserved residue (Fig. 2B).

The five clads of trihelix genes show little variation in their residues and domains (Fig. 2A, Fig. 4). The members of the clad GT-2 have two domains of trihelix in genes except AT3G10000 (EDA31) gene. In GT-2 and GTγ subfamily, the phenylalanine is responsible in the formation of third helix (Luo *et al.*, 2012). The isoleucine residue forms third helix in the members of SIP1clad while SH4 candidates have lysine (K) and asparagine (N) residues in between the two tryptophan residues of first and second helix respectively (Kaplan-Levy *et al.*, 2012). The analysis showed that motif 14 was conserved in all the members of SH4 clad, motif 12 in GTɣ clad and motif 3 was present in most of the members of SIP1 clad.

The trihelix family with 28 genes is divided into five clads/subfamilies in Arabidopsis; GT1, GT2, GT-ɣ, SH4 and SIP1 (Fig. 4). GT2 clad contains two conserved trihelix domains while all other subfamilies contain a single trihelix domain and the first domain ends with third helix of phenylalanine instead of tryptophan. The GTɣ clad has domain having phenylalanine involved in the third helix formation, while in SIP1, isoleucine is a third residue that is involved in helix formation. SH4 has lysine (K) and asparagine (N) residues in between the two tryptophan of first helix structure. The GT1 ends immediately after the formation of fourth helix.

***Gene structure analysis*:** Gene structure analysis was done to obtain the ratio of intron exon structures in all trihelix genes of *Arabidopsis thaliana* (Fig. 5). Based on these structures, it was observed that the average exon and intron number was 2 in the trihelix genes.

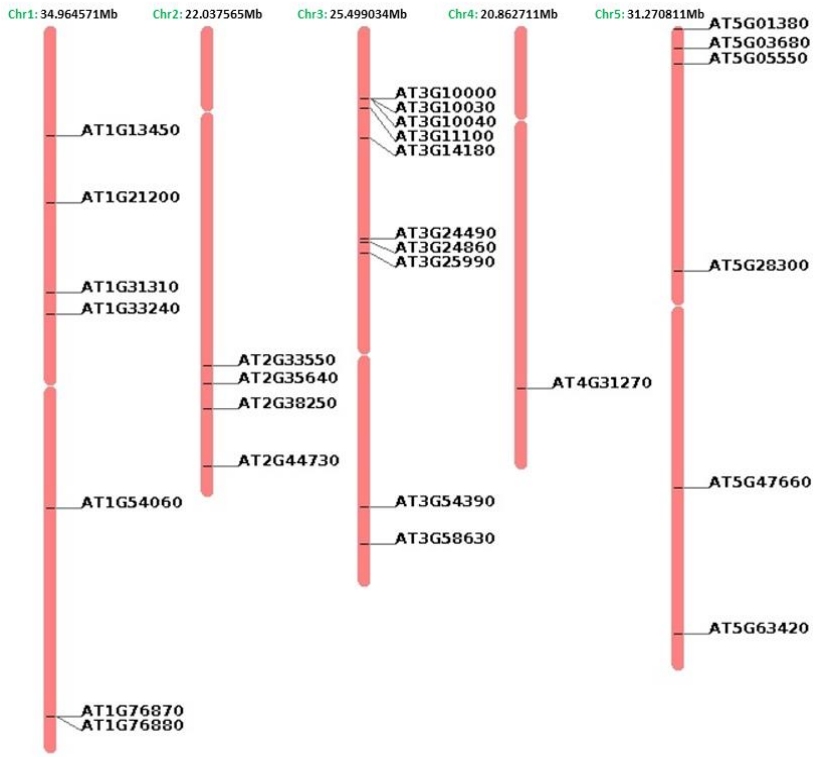

Figure 1. The mapping of 28 trihelix gene of *Arabidopsis thaliana* chromosomes. The chromosome number and size (Kaul *et al.*, 2000) is indicated at the top of each chromosome. The presented genes are indicating their absolute location on respective chromosome using accession numbers.

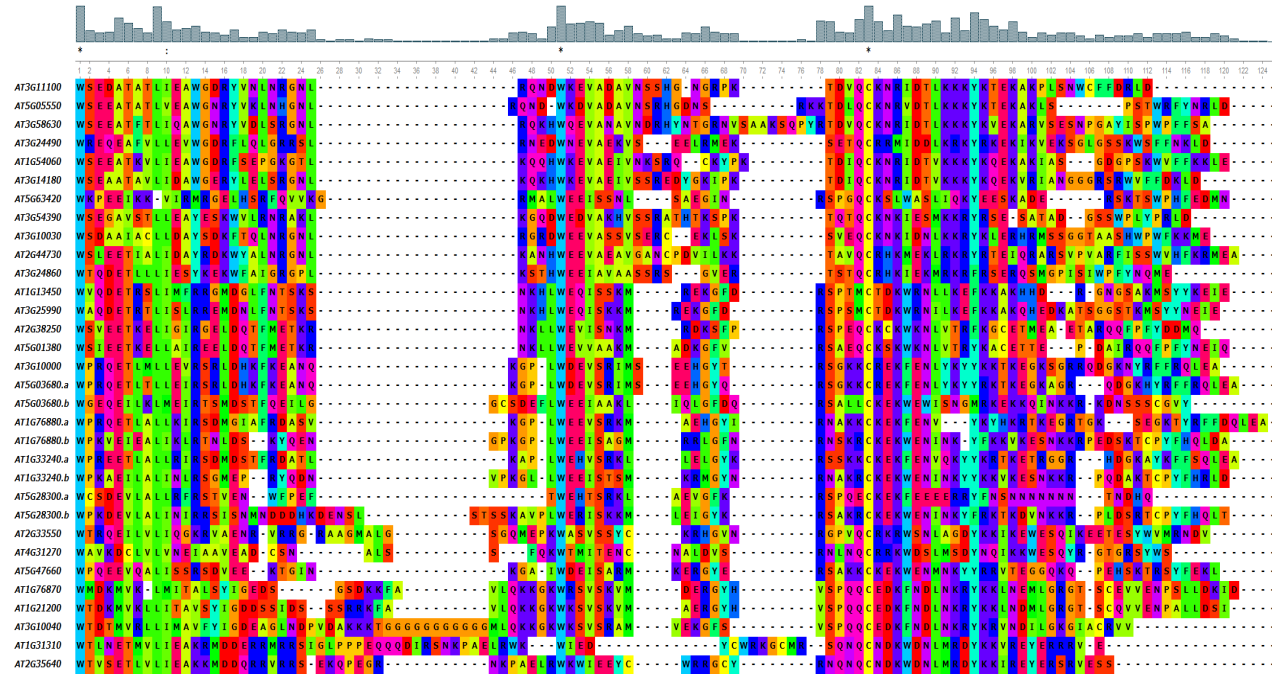

Figure 2A. UGENE alignment. The asterisk (*) shows conservation and identical residues, colon (:) shows maximum residues are similar. Genes e.g. *AT5G03680, AT1G76880, AT1G33240, AT5G28300* have 2 trihelix domains and also represents 2 trihelix motifs (see motif analysis).

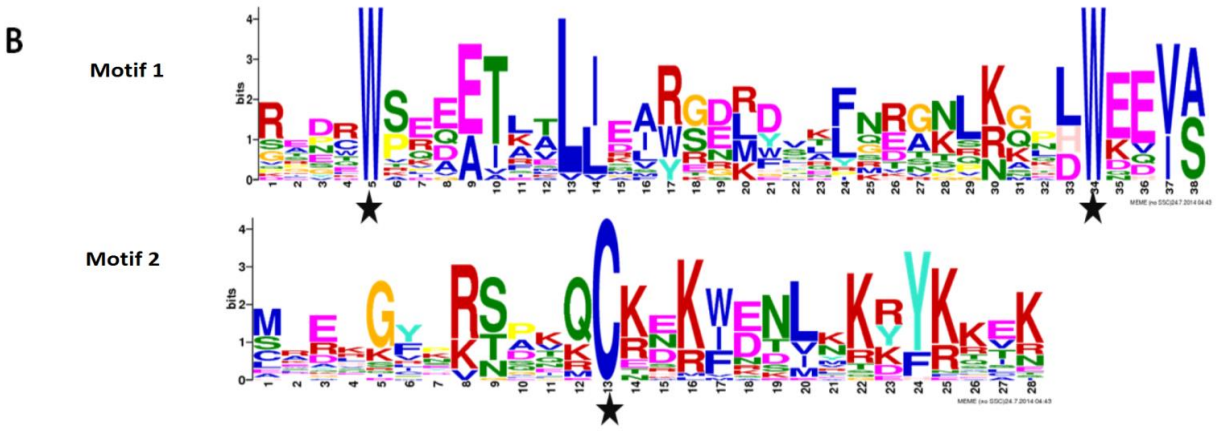

Figure 2B. The two motifs are showing the conserved residues in the trihelix genes of *Arabidopsis thaliana.* The first second and third helix are formed by tryptophan (W) Cysteine (C) is also highly conserved. Phenylalanine (F) and isoleucine (I) have maximum similarity ratio trihelix transcription factor gene family of *Arabidopsis thaliana*.

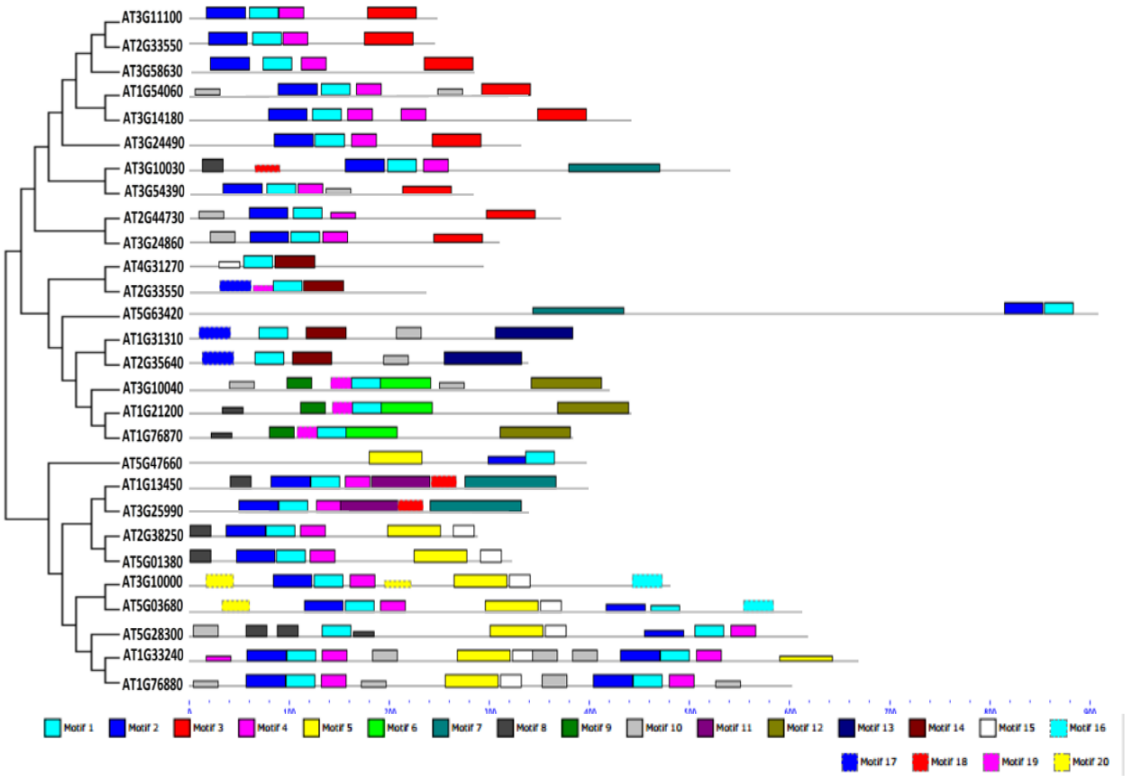

Figure 3. Phylogenetic tree and motif analysis of trihelix genes of Arabidopsis thaliana. ClustalW was used for multiple alignment and the neighbor joining strategy was used to construct unrooted phylogenetic tree which summarizes the evolutionary distances between of 28 members of trihelix of Arabidopsis thaliana using MEGA 6.06 package. Next to the tree are conserved motifs in trihelix proteins of *Arabidopsis thaliana* were found using MEME server. The proteins are represented according to their original accession numbers. Motif1 (colour coding: (TURQUOISE) part of domain and fundamental to trihelix family in Arabidopsis thaliana. The motif represented two times in; AT5G28300, AT1G76880, AT1G33240, AT3G54390 these accessioned sequences. The presence of conserved domain and motif two time further highlighted in multiple sequence alignment (MSA).

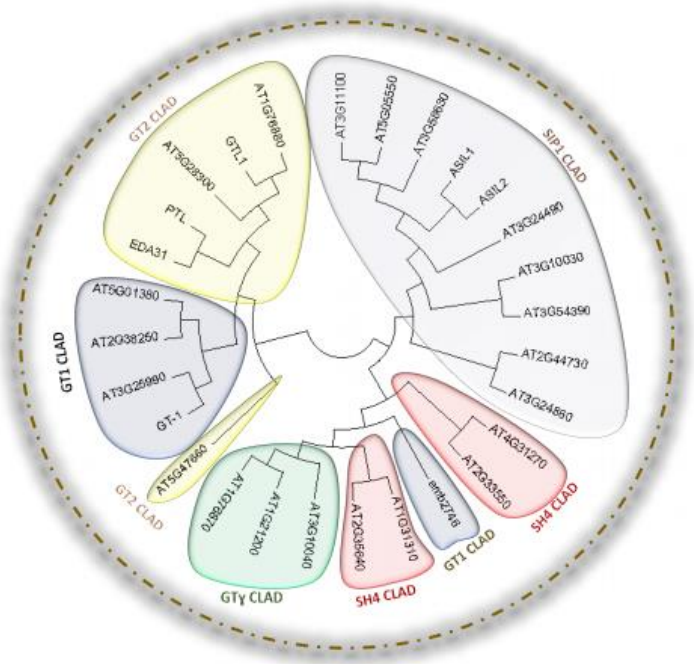

Figure 4. The amino acid sequences were multiply aligned using clustalW and the circular phylogenetic tree was generated using Neighbor joining method using MEGA 6.06 of trihelix genes. The genes are divided in five different clads; GT1, GT2, GTγ, SH4 and SIP1 represented using different colours according to Kaplan-Levy *et al*. (2012).

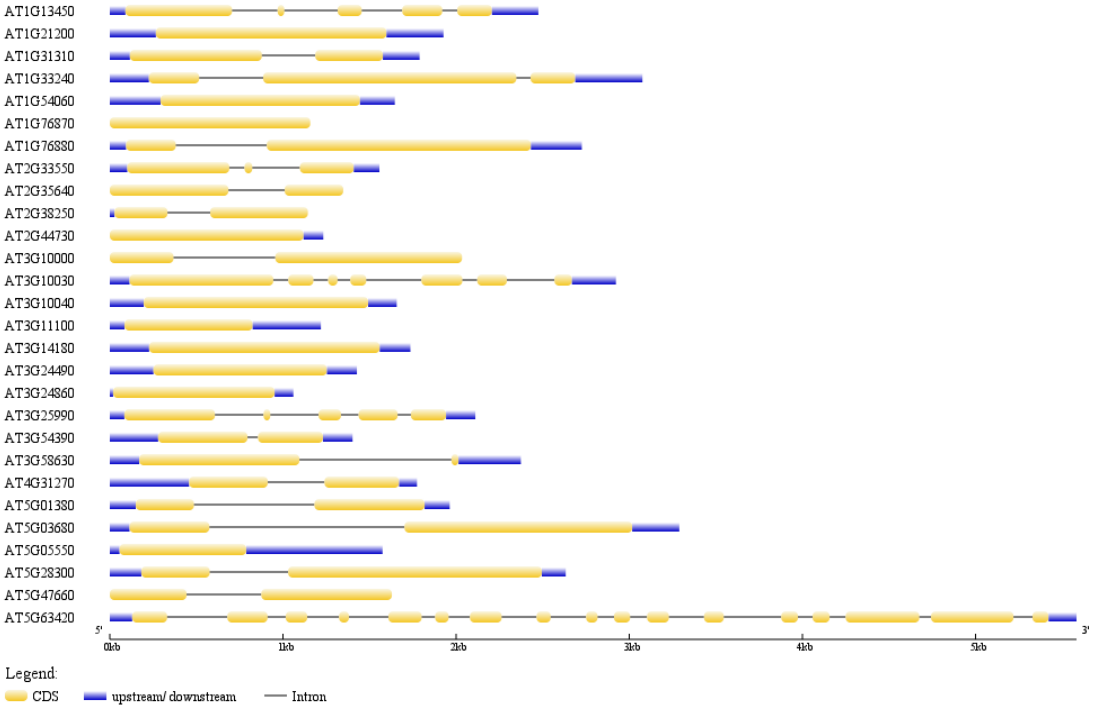

Figure 5. Gene structure of trihelix proteins in *Arabidopsis thaliana* designed in Genes Display Server GSDS. The intron, exon and UTR structures are shown. The light blue area represents exons, lines show the introns while pink region indicates UTR. The above scale can be used to predict the size of introns and exons.

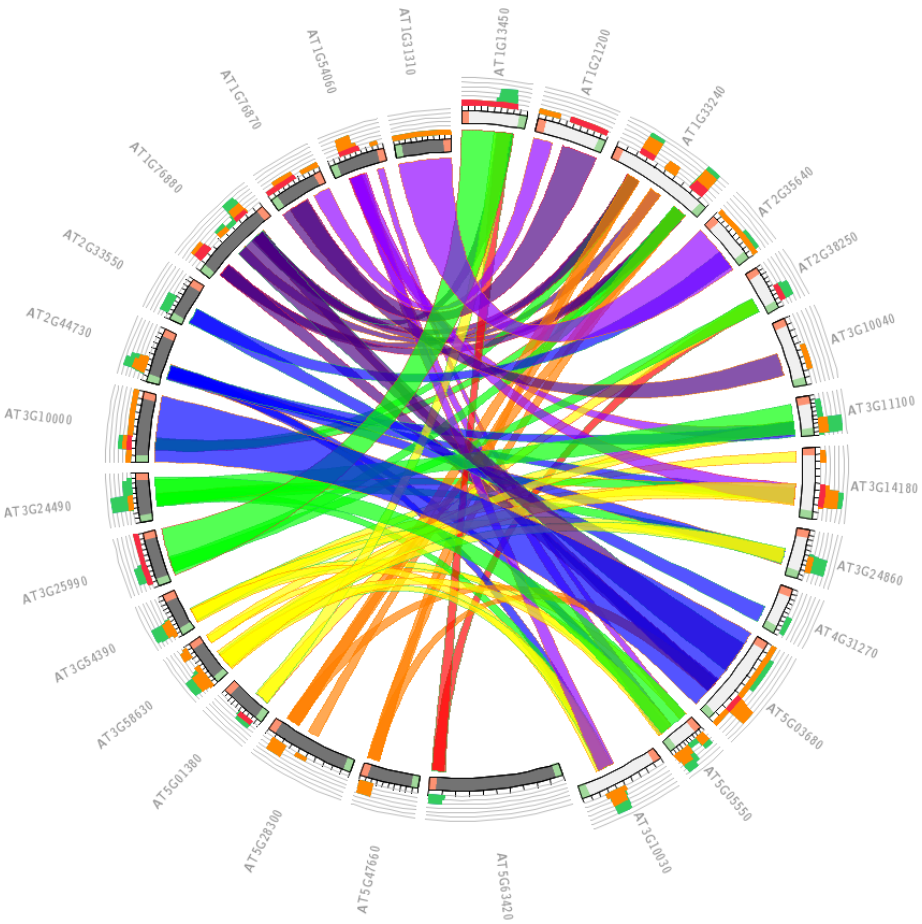

Figure 6. The Syntenic region among all the 28 genes of trihelix transcription factor family generated using circoletto (Darzentas, 2010). The colour variations represents the extent of similarity and homology among the genes. The red colour represents maximum matching portion among the trihelix transcription factor family of *Arabidopsis thaliana.*

There were 8 genes having only single exon. Thirteen genes had two exons, 3 genes had 3 exons and 2 genes had 5 exons while one gene had 7 exons. The largest gene had 17 exons (Table 1).

The genes belonging to SIP1 clad had either one or two exons except for *At3g10030* that had 7 exons. The members of SH4 clad had 2 or 3 exons in their structure. GT1 clads had 2 or five 5, this clad too hosts AT5G63420 (*emb2746)*, the largest gene with 17 exons. The candidates of GT2 clad had 2 or 3 exons in their genes. The members of GTɣ clad had one or two exons.

***Synteny analysis*:** The synteny analysis was done using circoletto (Darzentas, 2010) that performed local alignment and gave a circular output with colorful arcs (Fig. 6). The colour represent the extend of similarity, blue for the lowest similarity, then green, orange and finally red shows the increasing extent of similarity with increasing bit score. Most of genes are orthologs of each other, AT1G54060 (ASIL1) and AT3G14180 (ASIL2) are orthologs and are the result of gene duplication (Barr *et al*., 2012); AT3G10000 (EDA31) and AT5G03680 *(PTL*) are close relatives of each other and show similarity at N-terminal trihelix with 94% (Brewer *et al.*, 2004), GT-1 and AT3G25990, AT1G31310 and AT2G35640 are the orthologs (Jin *et al.*, 2013). The green ends show the N-terminal while the red ends show the C-terminal. The thin blue ribbons represent domains and show the presence at N- terminal or at C-terminal.

***Promoter analysis*:** Gene regulation and expression is controlled by regulatory elements present in the promoter sequence (Rushton *et al.*, 2012). The regulatory elements are known as *cis*-acting regulatory DNA elements. We selected five elements and mapped them on 1 KB promoter sequence upstream to the start codon. They were GT1-BOX, W-BOX, ARR1AT, CAATBOX and DOFCOREZM and their signature sequences are GRWAAW, TGAC, NGATT, CAAT and AAAG respectively (Fig. 7).

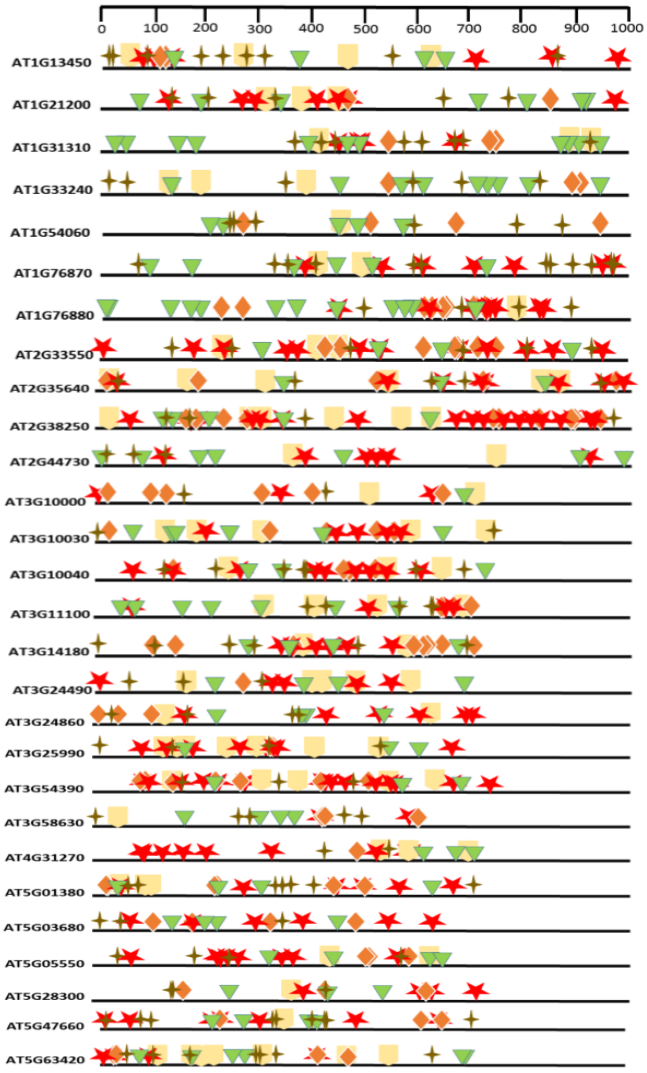

Figure 7. The Schematic representation of various PLACE-based motifs in 1KB promoter region of trihelix genes generated manually. The presence of motifs on +strand is shown in the figure. The symbols are used to represent GT1-BOX, W-BOX, ARR1AT, CAATBOX and DOFCOREZM with the sequences GRWAAW, TGAC, NGATT, CAAT and AAAG, respectively.

On average, the GT1 clad had maximum number of cis-regulatory elements, followed by SH4, GTγ, SIP1 and GT2 respectively (Table 2). The genes belonging to three clads contained GT1-BOX. The two genes- AT1G76870 and AT2G44730 do not have GT1-BOX. AT1G54060 (ASIL1), AT1G31310 and AT1G33240 (GTL1) do not have DOFCOREZM while the gene AT5G03680 (PTL) has no W-box.

Table 2. Frequency distribution of various *cis*-regulatory elements in the promoter region of trihelix genes.

Accession number	ARR1AT	GT1-BOX	W-BOX	CAATBOX	DOFCOREZM
AT1G54060	6	4	1	5	0
AT1G31310	8	3	3	11	4
AT1G13450	9	2	4	4	6
AT1G76880	4	10	1	12	9
AT1G33240	6	3	3	9	0
AT1G21200	5	2	3	7	7
AT1G76870	12	0	2	6	7
AT2G44730	3	0	2	7	6
AT2G33550	6	7	3	4	13
AT2G35640	5	5	6	3	8
AT2G38250	4	10	6	5	18
AT3G11100	5	4	4	7	4
AT3G58630	5	3	1	4	2
AT3G24490	3	1	5	4	5
AT3G10030	2	5	5	5	5
AT3G54390	3	7	5	3	15
AT3G24860	4	3	2	4	6
AT3G10040	7	4	3	3	9
AT3G10000	2	6	2	1	3
AT3G14180	6	8	2	4	6
AT3G25990	4	2	7	3	8
AT4G31270	2	1	3	3	9
AT5G01380	7	6	3	4	6
AT5G63420	7	4	7	6	4
AT5G03680	3	4	0	3	6
AT5G47660	9	3	1	4	7
AT5G28300	3	3	1	3	4
AT5G05550	4	3	2	4	8

DISCUSSION

Trihelix proteins are the member of transcriptional factor family in terrestrial plants that play diverse role in plant responses to various stimuli and help to activate or repress transcription of downstream target genes (Ling *et al.*, 2011). Functional studies of some of the trihelix genes have been of great interest, like AT5G03680 (*PTL*) (petal loss) is involved in morphological activities of flower organs (Kaplan-Levy *et al.*, 2012), the AT1G54060 (*ASIL1*) and AT3G14180 (*ASIL2*) are involved in chlorophyll accumulation during embryo development and GT-1 has the functions in light responsive activities (Qin *et al.*, 2014).

It was the subject of our interest that how many genes of this family exist in Arabidopsis, as well as the characteristics of gene structure, chromosomal locations, phylogenetic relationship, conserved motifs and expression pattern. The tandem repeat 'Spiro spin trihelix domain' (helix-loop-helix-loop-helix) is mostly found in plants. On the basis of differences in amino acid residues involved in helix formation, the members of trihelix family fall in five clads known as GT-1, GT-2, GTγ, SH4 and SIP1. The GT-2 members have two trihelix domains in genes except for AT3G10000 (EDA31) gene that has only one trihelix domain.

As stated by Luo *et al.* (2012), the trihelix genes belonging to GT-2 and GTγ subfamilies have phenylalanine residue in the third helix formation. In SIP1 members, isoleucine forms the third helix. SH4 candidates have lysine (K) and asparagine (N) residues in between the two tryptophan residues of first and second helix respectively (Kaplan-Levy *et al.*, 2012).

Based on the alignment of amino acid sequences, we found trihelix domain and determined the two highly conserved motif of trihelix domain using MEME Suite. The trihelix domain consists of 70 to 85 amino acid residues. The conserved residues in Spiro spin domain were mostly hydrophobic (tryptophan, isoleucine and phenylalanine) whereas tryptophan is involved in the formation of helix structure. The hydrophobicity promotes protein-protein interactions and is responsible for DNA binding, consequently playing a vital role in regulatory functions (Osorio *et al.*, 2012).

Gene localization of trihelix gene family showed that the genes are unevenly distributed throughout all five Arabidopsis chromosomes. Most of them were distributed on chromosome number 3. However, the gene distribution appeared to be uneven on different chromosomes.

Gene structure analysis revealed on an average almost equal number of coding and non-coding sequences. The non-coding genic region affects the gene expression and has strong positive relationship between gene coding and non-coding region (Colinas *et al.*, 2008). The lengthy non-coding genic region could also prevent the re-localization of the gene consequently ensuring the transcribed region attached with the matrix, decreasing the chances of variability of expression.

Five *cis*- acting regulatory elements, CAATBOX, DOFCOREZEM, W-BOX, GT1-BOX and ARR1AT, were located and mapped on 1 kb upstream of initiation codon. They all are involved in gene regulation under stress conditions. Their corresponding sequences are CAAT, AAAG, TGAC, GRWAAW and NGATT, respectively. The extent of the specificity of gene expression depends on *cis*-regulatory elements and their binding and interaction with the transcription factor (Zhou, 1999). The GT1-BOX interacts with GATA element and gives light-induced expression; however, its binding with other light responsive elements is necessary for expression (Zhou, 1999). ARR1AT controls the regulatory system operating in response to environmental stimulus (Sakai *et al.*, 2000). W-BOX is DNA binding place for WRKY transcription factors which are involved in gene regulation of many processes like plant growth and development, leaf senescence, cell signaling and in response to several biotic and abiotic stress responses (Ali *et al.*, 2014; Chi *et al.*, 2013; Rushton *et al.*, 2012; Rushton *et al.*, 2010).

The homology among the genes of trihelix family was predicted using Circoletto (Darzentas, 2010). There were three pairs with maximum homology >75% (red arcs) *AT1G21200* and *AT1G76870*, AT3G10000 (*EDA31*) and AT5G03680 (*PTL*), *GT*-1 and *AT3G25990*. Out of these two pairs are reported to be the orthologues i.e. AT3G10000 (*EDA31)*, AT5G03680 (*PTL*), *GT*-1, *AT3G25990*, AT5G03680 (*PTL)* and AT3G10000 (*EDA31*) gene are closely related and show identity of 94% at N-terminal trihelix (Brewer *et al.*, 2004). Orange arc, >50% homology, between *AT1G31310* and *AT2G35640* is another pair of orthologues (Jin *et al.*, 2013). AT1G54060 (*ASIL1)* is emerged from AT3G14180 (*ASIL2)* by gene duplication in Brassicaceae are the closest homologs of each other and at some extent show functional similarity in maturation (Barr *et al.*, 2012). The homology among the genes could be beneficial to find the functional relatedness among the similar gene.

***Conclusions*:** The present study concluded that the computational analysis of trihelix transcriptional factor family in *Arabidopsis* provides the fundamental information regarding phylogeny, chromosomal mapping, gene structure analysis, conserved motifs, and promoter analysis. The analysis of protein domains showed that tryptophan (W), isoleucine (I) and phenylalanine (F) residues play a key role in the formation of helical structures. The regular expression of trihelix domain was predicted to be RxDR**W**[SP]EE[EA]TL[TA]L[IL]E[AI][RW][GS][DE][LRM]Dxx[FL]N[RE][GA][NK]L[KR][GQ]P[LDH]**WEE**[VI][AS]MxExG[YF]x[RK][ST]PKQC[KR][EDN]K[**WFI**][EDNL[KN]K[RYK]Y[KR]KEK. The *cis*-regulatory elements bind with transcriptional factors but trihelix genes are not restricted to GT1-BOX as few like *At1g76870* and *At2g44730* genes lacked it.

Introns are the non-coding genic regions that play role in gene expression. Most of the genes belonging to trihelix family had introns except a few. The sequences of trihelix genes showed similarity and some genes were orthologues of others, they showed >75% matching. The thin blue ribbons are the regions of trihelix domain in the genes and their presence at N- or C-terminal could be deduced. We conclude that these findings will be helpful resource in understanding trihelix family of Arabidopsis which play diverse role in plant stress conditions and development.

REFERENCES

Ali, M.A., K. Wieczorek, D.P. Kreil and H. Bohlmann. 2014. The beet cyst nematode *Heterodera schachtii* modulates the expression of WRKY transcription factors in syncytia to favour its development in Arabidopsis roots. PLoS One 9:e102-360.

Bailey, T.L., M. Bodén, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li and W.S. Noble. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:202-208.

Barr, M.S., M.R. Willmann and P.D. Jenik. 2012. Is there a role for trihelix transcription factors in embryo maturation? Plant Signal. Behaviour 7:205-209.

Brewer, P.B., P.A. Howles, K. Dorian, M.E. Griffith, T. Ishida, R.N. Kaplan-Levy, A. Kilinc, and D.R. Smyth. 2004. PETAL LOSS, a trihelix transcription factor gene, regulates perianth architecture in the Arabidopsis flower. Development 131:4035-4045.

Chi, Y., Y. Yang, Y. Zhou, J. Zhou, B. Fan, J.-Q. Yu and Z. Chen. 2013. Protein–protein interactions in the regulation of WRKY transcription factors. Mol. Plant 6:287-300.

Colinas, J., S.C. Schmidler, G. Bohrer, B. Iordanov and P.N. Benfey. 2008. Intergenic and genic sequence lengths have opposite relationships with respect to gene expression. PLoS One 3:e3670.

Darzentas, N. 2010. Circoletto: visualizing sequence similarity with Circos. Bioinformatics 26:2620-2621.

Gao, J., H. Peng, X. He, M. Luo, Z. Chen, H. Lin, H. Ding, G. Pan and Z. Zhang. 2013. Molecular phylogenetic characterization and analysis of the WRKY transcription factor family responsive to *Rhizoctonia solani* in maize. Maydica 59:32-41.

Higo, K., Y. Ugawa, M. Iwamoto and T. Korenaga. 1999. Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res. 27:297-300.

Hu, B., J. Jin, A. Guo, H. Zhang, J. Luo and G. Gao. 2015. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296-1297.

Jin, J., H. Zhang, L. Kong, G. Gao and J. Luo. 2013. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. 42:D1182-7.

Kaplan-Levy, R.N., P.B. Brewer, T. Quon and D.R. Smyth. 2012. The trihelix family of transcription factors–light, stress and development. Trends Plant Sci. 17:163-171.

Kaul, S,, H.L. Koo, J. Jenkins, M. Rizzo, T. Rooney and L.J. Tallon *et al*. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815.

Lang, D., B. Weiche, G. Timmerhaus, S. Richardt, D.M. Riaño-Pachón, L.G. Corrêa, R. Reski, B. Mueller-Roeber and S.A. Rensing. 2010. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol. Evol. 2:488-503.

Ling, J., W. Jiang, Y. Zhang, H. Yu, Z. Mao, X. Gu, S. Huang and B. Xie. 2011. Genome-wide analysis of WRKY gene family in *Cucumis sativus*. BMC Genomics 12:471.

Luo, J.-L., N. Zhao and C.-M. Lu. 2012. Plant Trihelix transcription factors family. Hereditas 34:1551-1560.

Nagano, Y., T. Inaba, H. Furukashi and Y. Sasaki. 2001. GENES: STRUCTURE AND REGULATION-Trihelix DNA-binding protein with specificities for two distinct cis-elements. Both important for light down-regulated and dark-inducible gene expression in higher plants. J. Biol. Chem. 276:22238-22243.

Nuruzzaman, M., A.M. Sharoni, K. Satoh, H. Kondoh, A. Hosaka and S. Kikuchi. 2012. A genome-wide survey of the NAC transcription factor family in monocots and eudicots. In: Nuruzzaman *et al*. (eds.), Introduction to Genetics–DNA Methylation, Histone Modification and Gene Regulation. Hong Kong: iConcept Press, ISBN 978-14775549-4-4.

Okonechnikov, K., O. Golosova, M. Fursov and the UGENE team. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166-1167.

Osorio, M.B., L. Bücker-Neto, G. Castilhos, A.C. Turchetto-Zolet, B. Wiebke-Strohm, M.H. Bodanese-Zanettini and M. Margis-Pinheiro. 2012. Identification and in silico characterization of soybean trihelix-GT and bHLH transcription factors involved in stress responses. Genet. Mol. Biol. 35:233-246.

Qin, Y., X. Ma, G. Yu, Q. Wang, L. Wang, L. Kong, W. Kim and H.W. Wang. 2014. Evolutionary history of trihelix family and their functional diversification. DNA Res. 21:499-510.

Ramirez, S.R. and C. Basu. 2009. Comparative analyses of plant transcription factor databases. Curr. Genomics 10:10-17.

Riaño-Pachón, D.M., L.G.G. Corrêa, R. Trejos-Espinosa and B. Mueller-Roeber. 2008. Green transcription factors: a Chlamydomonas overview. Genetics 179:31-39.

Rushton, D.L., P. Tripathi, R.C. Rabara, J. Lin, P. Ringler, A.K. Boken, T.J. Langum, L. Smidt, D.D. Boomsma, N.J. Emme, X. Chen, J.J. Finer, Q.J. Shen and P.J. Rushton. 2012. WRKY transcription factors: key components in abscisic acid signalling. Plant Biotechnol. J. 10:2-11.

Rushton, P.J., I.E. Somssich, P. Ringler and Q.J. Shen. 2010. WRKY transcription factors. Trends Plant Sci. 15:247-258.

Sakai, H., T. Aoyama and A. Oka. 2000. Arabidopsis ARR1 and ARR2 response regulators operate as transcriptional activators. Plant J. 24:703-711.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30:2725-2729.

Xie, Z.M., H.F. Zou, G. Lei, W. Wei, Q.Y. Zhou, C.F. Niu, Y. Liao, A.G. Tian, B. Ma and W.K. Zhang. 2009. Soybean Trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic Arabidopsis. PLoS One 4:e6898.

Zhou, D.X. 1999. Regulatory mechanism of plant gene transcription by GT-elements and GT-factors. Trends Plant Sci. 4:210-214.