Genome-wide identification and analysis of the *trihelix* transcription factors in sunflower

J. SONG1, W.Y. SHEN1,2, S. SHAHEEN1, Y.Y. LI1, Z.R. LIU1, Z. WANG1, H.B. PANG1,2,* and Z. AHMED3,4,5,*

1 College of Life Sciences, Shenyang Normal University, Shenyang 110034, P.R. China
2 Institute of Grassland Science, Northeast Normal University, Key Laboratory of Vegetation Ecology, Ministry of Education, Changchun, 130024, P.R. China
3 Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Xinjiang, 830011, P.R. China
4 Cele National Station of Observation and Research for Desert-Grassland Ecosystem, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Xinjiang, 848300, P.R. China
5 Xinjiang Desert Plant Roots Ecology and Vegetation Restoration Laboratory, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, 830011, P.R. China

*Corresponding authors: E-mail: panghongbo800206@163.com, zeeshanagronomist@yahoo.com

Abstract

The *trihelix* genes encode plant-specific transcription factors, which play a vital role in plant morphological and developmental processes. However, information about the presence of *trihelix* genes in sunflower (*Helianthus annuus* L.) is scarce. Sunflower belongs to composite family and possesses strong drought and salt-alkali tolerance. In this study based on *H. annuus* genome data, we have identified and analyzed the *trihelix* genes with a complete description of their physical and chemical properties, phylogenetic relationships, motif composition, chromosome distribution, exon-intron structure, cis-acting elements, and chromosome collinearity. In *H. annuus*, 31 full-length *trihelix* genes were identified and categorized into six subgroups (SIP, GT1, SH4, Gδ, GT-γ, and GT2). Multiple Em for motif elicitation (MEME), used for conservative motif analysis, identified 10 distinct motifs unevenly distributed on 31 *trihelix* genes. In addition to that, chromosome localization analysis showed the number and distribution of these *trihelix* genes on 17 chromosomes of *H. annuus*. Transcriptional structure analysis revealed the structure of introns and exons of different gene members. Furthermore, *cis*-element analysis identified 19 different types of *cis*-elements mainly related to abiotic stress, hormones, and growth and development of plant. Results of this study manifested novel insights into phylogenetic relationships and possible functions of *H. annuus trihelix* genes. Moreover, these findings can assist in future studies regarding specific physiological effects of *H. annuus trihelix* transcription factors.

Keywords: *cis*-acting elements, chromosome distribution, *Helianthus annuus*, MEME, motif composition, phylogenetic relationships.

Introduction

The *trihelix* transcription factor is a unique gene family, which was firstly discovered and isolated in peas. Due to its specific binding with GT element in the light-induced *rbcS-3A* gene promoter, it was initially named as GT factor (Qin et al. 2014). Later, a typical *trihelix* structure was indicated on the DNA-binding domain (helix-loop-helix-loop-helix) so it was renamed as *trihelix* transcription factor (Nagano 2000). The *trihelix* gene family has been identified in many plants with different number of genes. For example, there were 30 *trihelix* genes identified in soybean (Soja japonica) genome (Nagano et al. 2000). In *Helianthus annuus*, 31 full-length *trihelix* genes were identified and categorized into six subgroups (SIP, GT1, SH4, Gδ, GT-γ, and GT2). Multiple Em for motif elicitation (MEME), used for conservative motif analysis, identified 10 distinct motifs unevenly distributed on 31 *trihelix* genes. In addition to that, chromosome localization analysis showed the number and distribution of these *trihelix* genes on 17 chromosomes of *H. annuus*. Transcriptional structure analysis revealed the structure of introns and exons of different gene members. Furthermore, *cis*-element analysis identified 19 different types of *cis*-elements mainly related to abiotic stress, hormones, and growth and development of plant. Results of this study manifested novel insights into phylogenetic relationships and possible functions of *H. annuus trihelix* genes. Moreover, these findings can assist in future studies regarding specific physiological effects of *H. annuus trihelix* transcription factors.

Received 8 December 2020, last revision 25 January 2021, accepted 2 February 2021.

Abbreviations: ABRE - abscisic acid response element; CDS - coding sequence length; ERE - ethylene reaction element; GSDS - gene structure display server; HMM - Hidden Markov model; LTRs - low temperature responsiveness elements; MEME - multiple Em for motif elicitation; Mr - relative molecular mass; pI - isoelectric point.

Acknowledgements: This work was supported by the Shenyang Youth Science and Technology Project (No. RC190223) and the Hundred Talent Program of Shenyang Normal University (SSDBRJH2002012).

Conflict of interest: The authors declare that they have no conflict of interest.
TRIHELIX GENES IN SUNFLOWER

genes in Arabidopsis thaliana, 41 in Oryza sativa, and 36 in Solanum lycopersicum (Gao et al. 2009, Ji et al. 2015, Yu et al. 2015). These transcription factors are usually classified into subgroups. Five subgroups of GT1, GT2, GTγ, SH4, SIP (Kaplan-Levy et al. 2012) are classified in Arabidopsis (Gao et al. 2009). In rice they are also classified into five subfamilies, but unlike Arabidopsis, GT-1 and GT-2 are combined into a GT clade and a GTδ group is added (Ji et al. 2015), whereas in tomato, the trihelix family is classified into six subfamilies (clades GT-1, GT-2, SH4, SIP1, GTγ, and Gδ (Yu et al. 2015).

Recent studies have shown that this gene family plays a crucial role in the regulation of different growth and developmental responses of plants. For instance, the AtGT-1, AtGT-3a, and AtGT-4 Arabidopsis genes in the GT-1 group are responsible for light response functions (Park et al. 2004), while OsSH4 (OsSHAI) rice genes in SH4 group have functions related to seed shattering development (Konishi et al. 2006). Similarly, ASIL1, ASIL2, At3g10030, and FIP2 Arabidopsis genes belonging to SIP1 group are closely associated with the growth and development of plants (Kuromori et al. 2006, Geraldo et al. 2009, Barr et al. 2012, Gao et al. 2014). The ASIL2 gene plays a crucial role in the early embryonic development of plants (Barr et al. 2012), whereas the ASILI gene is involved in the late embryonic development and regulation of growth from vegetative to reproductive phase (Gao et al. 2014). Likewise FIP2 gene has been found to regulate the timing of plant flowering (Geraldo et al. 2009), while At3g10030 gene showed involvement in plant leaf development (Kuromori et al. 2006). Previous studies also revealed that the trihelix family members play significant roles in biotic and abiotic stress responses in some plants (Xie et al. 2009, Wang et al. 2016, Yu et al. 2018). For example AtGT-3b Arabidopsis gene has function in salt stress response (Park et al. 2004), the ShCIGT gene of tomato enhances cold and drought stress tolerance (Yu et al. 2018), and in the same context OsGTγ-1, OsGTγ-2, and OsGTγ-3 genes belonging to GTγ group in rice also function in abiotic stress responses (Fang et al. 2009).

Sunflower (Helianthus annuus L.) is not only a major oilseed crop but it is also utilized as an important tool in plant research. It can adapt to adverse environmental conditions such as drought and stress and ensures stable yield (Kane and Rieseberg 2007). Thus, the trihelix gene family has been assumed to play a major role in sunflower abiotic stress resistance. Sequencing of the sunflower genome has been completed, but studies on the trihelix gene family of sunflower have not yet been published (Neupane et al. 2018). In this study, we will use bioinformatics methods to comprehensively analyze the phylogeny and homology of the trihelix gene family. For this purpose, chromosome distribution, genetic isomorphism, phylogenetic analysis, genetic structure, motif composition, and cis-element analysis techniques have been used. The purpose of this study was to analyze the structure and function of the sunflower trihelix genes and the phylogenetic relationship between sunflower and other species. Findings of this study can facilitate an insight into the development of sunflower resistance under abiotic stresses, and could provide a theoretical basis for further research on the physiological and biochemical functions of trihelix genes.

Materials and methods

Identification of trihelix genes and physicochemical analysis: Whole genome sequence of sunflower file and gene annotation file were downloaded from the Ensemble database (http://plants.ensembl.org/). The hidden Markov model file (HMM) corresponding to the trihelix gene was downloaded from the Pfam database (http://sfam.org/, PF13837) (Li et al. 2019). We performed two HMM searches to identify the members of our gene family by following the Lozano et al. (2014) (P-value < 0.005). The obtained gene family members were checked by Pfam, NCBI-CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and SMART (http://smart.embl-heidelberg.de/) online websites to identify the conserved protein domain. The gene was retained if it contained a domain, otherwise it was removed.

We used the Expasy (https://web.expasy.org/protparam/) to analyse amino acid number (protein length), relative molecular mass (Mr), isoelectric point (pI), and other physicochemical indexes. Finally, the YLoc website (http://abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi) was used to predict the subcellular localization of each trihelix protein.

Phylogenetic analysis: The protein sequence of the trihelix gene family in sunflower, rice (Ji et al. 2015), and Arabidopsis (Kaplan-Levy et al. 2012) was aligned by using Biodit 7.0, and the unrooted phylogenetic tree was constructed by MEGA7.0 with the MP (maximum-parsimony) method; the bootstrap replications were set to 1 000. Phylogenetic tree results were imported into Figtree software for modification.

Gene structure analysis and motif composition of the trihelix genes: The MOTIF information of the gene family was searched through the online software MEME (http://meme-suite.org/). The parameter settings were as follows: The output motif number was 10, the minimum motif width was set to 6, and the maximum motif width was set to 50. It was allowed to repeat multiple times, and other parameters have used the default values. The MEME generated xml file was imported into TBtools software to export the location information map of MOTIF (Chen et al. 2020). The downloaded sunflower gff3 file was used to obtain the structure information of the trihelix gene, a Perl script (Supplement file 1) was run to generate a txt file containing the transcript information of the sunflower trihelix gene family members, and then the online drawing software Gene Structure Display Server 2.0 (http://gsds.gao-lab.org/) was used to map the gene exon-intron structure.

Chromosome distribution of the trihelix genes: The Perl script was used to obtain the position information of the sunflower trihelix chromosome, then obtained the
gene ID file and the genome length file, and finally placed the gene on the online mapping tool MapGene2Chrom web v2 (http://mg2c.iask.in/mg2c_v2.0/) for the picture of position on chromosome. The generated pictures were modified in the Adobe Illustrator CS6 software.

**Cis-element analysis of trihelix transcription factor family:** We used the Perl script to obtain the 1 500 bp sequence upstream of these genes. Cis-acting elements were predicted at PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) and those for plant growth, plant development, and phytohormone responses were counted. The result was imported into TB-tools (Chen et al. 2020) for drawing heatmap.

**Comparative genomics analysis:** The gene location information of the species and the CDS sequence of the corresponding gene were extracted by Perl script, and then collinear analysis was carried out where the trihelix gene with collinearity between the two species was screened and highlighted.

**Results**

The pI and Mr of 31 identified sunflower trihelix genes were forecasted (Table 1 Suppl.). Sunflower trihelix proteins showed large variations in their length and ranged between 214 and 848 amino acid residues and Mr ranging from 23942.7 to 94437.5 Da. The trihelix proteins in sunflower also revealed large differences in their isoelectric point (pI) values (ranging from 4.41 to 9.44) and mostly concentrated between 6 and 8 (average 6.79). Subcellular localization prediction of sunflower

---

**Fig. 1.** Phylogenetic tree of trihelix proteins from *Helianthus annuus*, *Arabidopsis thaliana*, and *Oryza sativa* L. The phylogenetic tree was clustered into GT-γ, GT1, GT2, SH4, and Gδ. The members of *Helianthus annuus* were shown by the light yellow background.
trihelix proteins showed that trihelix proteins were mostly located in the nucleus, but OTG30479 and OTG38290 were distributed in the chloroplast, while OTG38320 was located in the cytoplasm.

An unrooted phylogenetic tree of trihelix protein from sunflower, rice and Arabidopsis was constructed using Maximum parsimony (MP) method to distinguish the evolutionary relationship of sunflower trihelix genes. Similarly as in other plants, sunflower trihelix transcription factors can be categorized into six groups: SIP, GT1, SH4, Gδ, GT-γ, and GT2, and the group GT2 was further classified into two subgroups, including Tc and TN (Fig. 1). Among the six subfamilies, the SH4 and GT2 contained the highest number of members and the Gδ clade only has one member named OTG30479.

To further understand the existing trihelix motifs, we then performed a complete multiple protein sequence alignment of all 31 trihelix transcription factors (Fig. 2A) and identified 10 distinct motifs. Their conservative motifs were displayed in Fig. 1 Suppl. They were unevenly distributed on 31 trihelix genes (Fig. 2A). All the identified trihelix genes contained motif 1 indicating that during the process of gene differentiation, motif 1 was relatively conservative and does not change substantially.

Interestingly, some motifs were specific to only one or two subfamilies. For example, motif 6 was exclusively found in the GT-γ subfamily, and motif 3 was universal except SH4 clade. Motifs 1, 2, and 3 had two duplications in the most members of clade GT2, but among all the members, OTG30479 was the most special with motif 1 - 4; this corresponded to its genetic structure (Fig. 2B), which had the largest number of exons. So its motif might be due to multiple interruptions by introns. Meanwhile, we also found that protein conserved motifs of the same clade of most groups were similar.

To further investigate the structural components of the trihelix genes, the exon and intron structures of trihelix genes were obtained by comparing them with corresponding genomic DNA sequences (Fig. 2B). The trihelix gene structure showed that OTG30479 had the longest gene structure, including 16 exons and 15 introns. Eight trihelix gene members had only one exon, whereas other genes had 2 - 16 exons. Most trihelix gene members had introns, of which OTG26142 had a long intron. The eight sunflower trihelix genes in the GT2 group all had introns. Additionally, both OTG26730 and OTG09673 contained similar gene structure and motif composition. In terms of untranslated regions (UTR), the six gene members did not contain UTR, and the rest contained 1 - 2 UTRs, of which OTG16790 had a longer UTR. The analysis of the gene structure and motif composition provided further support to the phylogenetic classification of trihelix gene family.

According to the available location information, all the trihelix genes were mapped on the chromosomes (Fig. 3). Chromosome 1, 3, 11, 12, and 16 carried three trihelix genes, and two trihelix genes were located on chromosomes 5, 6, 10, 15, and 17, respectively. Only one trihelix gene was found on chromosomes 2, 4, 8, 9, 13, and 14, and no trihelix gene was found on chromosome 7. These results suggested a random distribution of the trihelix genes among different chromosomes of H. annuus. As shown in Fig. 3,
we found that the gene family members on chromosome 1 were clustered in the lower part of the chromosome, and the three genes were located close to each other. Similarly, the two gene family members on chromosome 11 were in clusters. The length and starting position of these genes were also very similar (Table 1 Suppl.). The distance was particularly short and the distribution of these two genes in the chromosome appeared to be almost overlapped. Similarly, the OTG09292 and OTG09256 genes located adjacent on chromosome 11 were also clustered.

The distribution of different cis-acting elements in the promoter of a gene may indicate the difference in its function and regulation. Different cis-elements responsible for abiotic stress, plant growth and development, and hormone-related-response elements were identified during this study (Fig. 4). Statistical results revealed that hormone-related-response elements, such as abscisic acid response element (ABRE) and CGTCA/TGACG-motif elements were overrepresented in the promoter regions of the trihelix genes. For example, OTF90560 gene had 7 ABRE elements, OTG09256 gene had 6 ABRE elements, and OTG01800, OTG22032, OTG26142 and OTG38320 genes contained 5 ABRE elements, respectively. Plant growth and development related cis-elements appeared only in few members of the gene family, whereas abiotic stress-related elements were detected in the promoter regions of some trihelix genes, including 3 low temperature responsiveness elements (LTRs) for OTF91880, OTG06393, and OTG30479, 2 LTRs for OTG18952, OTG22032 and OTG32052, 2 MYB binding site (MBS) elements for OTF85373, respectively. We observed that the OTG09673 gene in the GT2 subfamily contained three cis-elements related to stress responses (1 MBS, 1 LTR, and 1 TC-rich repeats). In addition, we also observed that OTF95801, OTF96213, OTF98282, OTG09292, OTG06033, and OTG11003 only contained hormone-related-response elements, whereas nothing was found on OTG22742 and OTG07556 genes. Promotor analysis indicated that
the OTF90560 gene had the highest number (19) of cis-elements, containing 7 ABRE, 4 CGTCA, 4 TGACG, 1 TGA, 1 TCA, 1 RY-element and 1 MBS.

Trihelix members with similar phylogenetic relationships in different plants have the same motif composition. Based on the previous studies, we performed homology analysis of the trihelix genes in two eudicots (Arabidopsis and tomato) to construct a representative sunflower homology map (Fig. 2 Suppl.), and presumed the evolutionary origin of the trihelix genes. Through the analysis, we found that the sunflower trihelix gene had one homology to the Arabidopsis thaliana trihelix gene, and had two pairs of homologous genes with tomato. This indicates that there was a collinear gene pairs with dicotyledonous plants, and it is speculated that these homologous pairs were formed after the differentiation of monocotyledonous and dicotyledonous plants (Xie et al. 2018).

Discussion

The trihelix gene family has been widely analyzed in plants (Gao et al. 2009, Ji et al. 2015, Yu et al. 2015). However, information about sunflower trihelix genes is scarce and needs exploration. In this study we identified 31 members of the trihelix gene family from the H. annuus genome database, and then constructed a phylogenetic tree containing 100 trihelix proteins from rice, Arabidopsis, and sunflower (Fig. 1). These genes were categorized into six groups of SIP, GT1, SH4, Gδ, GTγ, and GT2 in sunflower. Previous reports mentioned that the members of a gene family clustering together in the same group tend to possess similar functions (Kaplan-Levy et al. 2012, Liu et al. 2020). Thus, the putative orthologues of sunflower trihelixes with known Arabidopsis and rice were identified.

SIP group gene (OTG28164) was the homologous gene of Arabidopsis AT3G24860 (AST1) (Fig. 2). It responds to abiotic stresses (Xu et al. 2018), implying that OTG28164 may also be involved in hormone signaling and stress responses. GT2 group genes (OTG26730, OTG09673 and OTG33983) were the homolog genes of Arabidopsis AT1G33240 (GTL1) (Fig.2). GTL1 regulates ploidy-dependent cell growth in the Arabidopsis trichomes (Breuer et al. 2009), inferring that these three genes of sunflower may be involved in plant growth.

In terms of gene structure and conserved domain analysis, we found that members of the same clade usually had similar motif composition and exon/intron structure. For example, the motif and exon/intron structures of the OTG26730 and OTG09673 genes from same clade in the GT2 subfamily and the OTG16790 and OTF90560 genes of same clade in the SIP subfamily were similar, and we speculated that they could have similar functions (Fig. 2). Then, by plotting the position map of the genes...
on the chromosome, it was observed that the genes were unevenly distributed on the chromosome, and some genes were clustered on the same chromosome. Cannon et al. (2004) have found that different members of the same gene family may have tandem repeat relationships between adjacent members if they are adjacent or close to each other on the chromosome or in the same region of the evolutionary tree. Therefore, we can speculate that the OTG38290 and OTG38319 genes on chromosome 1 and the OTG09292 and OTG09256 genes on chromosome 11 are tandem repeats (Fig. 3). Gene replication is an important mechanism for gene family evolution (Yang et al. 2008), and tandem replication is one of the major evolutionary patterns (Mao et al. 2016). Therefore, we can use it to verify the existence of tandem replication between genes. Tandem repeat genes may play the same or similar roles in plants and their study may help to understand the importance of genetic evolution.

Promoter sequences play a significant role in the differentiation of gene functions. Analysis of cis-elements in this study indicated that most of the sunflower trihelix gene promoter sequences contained these cis-elements which were associated with hormonal and abiotic stresses (Fig. 4). The cis-elements in sunflower mainly contained three types of stress-related elements, plant growth and development, and hormone-related-response elements. Stress-related elements included the abscisic acid response element ABRE (Fujita et al. 2005), the protein metabolism regulatory element O2 binding site (O2-site) (Yin et al. 2016), and LTR (White et al. 1994). These stress-related elements have been shown to play a major role in plant survival under stressful conditions. These results indicated that the sunflower trihelix was widely involved in abiotic stress responses, providing a theoretical basis for further analysis of the mechanism of action of the trihelix gene family in sunflower. In this study, we found that the 19 cis-elements in the promoter region of the trihelix gene (OTF90560) were screened in the SIP subfamily, and their number was higher than in other genes (Fig. 4). Therefore, we speculated that this gene may have a crucial role in plant growth and stress-related responses.

Abscisic acid is a crucial plant hormone regulating gene expression and stomatal closure, and it is related to abiotic stress responses (Wasilewska et al. 2008, Lee and Luan 2012). The identification of genetic modules related to drought response in rice showed that ABRE motifs are conserved among drought response genes, and these drought response genes may be related to ABA-dependent drought response pathways (Zhang et al. 2012). In this research, we observed that 71 % of the genes contained ABRE cis-elements and the number was large, so we presumed that most of the genes are induced by ABA in drought stress response. In addition, the relationship between some uninduced genes and the ABA signaling pathway remains unexplored and needs further study.

In a comparison with Arabidopsis thaliana and tomato, we found that the trihelix gene on sunflower chromosome 11 had a pair of homologous genes with those on chromosome 3 of Arabidopsis thaliana and chromosome 9 of tomato. Moreover, a pair of homologous genes on the chromosome 10 on the trihelix gene and the tomato chromosome 12 has also been noted. It is found that there is a collinear gene pair with dicotyledonous plants, but further research is needed for the determination of homologous genes.

In this study, a comprehensive search for trihelix transcription factors has been conducted throughout the H. annuus genome data, and 31 members have been identified. Additionally, comparative phylogenetic analysis with Arabidopsis and rice trihelix TFs showed that they are categorized into six subfamilies. The comparative, phylogenetic analyses of trihelix members will be useful to establish a comprehensive functional characterization of the trihelix gene family. The structural information of the protein will elucidate their functional analysis. The cis-element analysis will be useful to understand their possible roles in mediating hormone cross-talk in response to abiotic stresses. Overall, the data presented here has provided a baseline and can be used in future studies for genetic improvement of agronomic and/or stress tolerance traits in sunflower and other members of Asteraceae family.

References


Geraldo, N., Baurle, I., Kidou, S., Hu, X.Y., Dean, C.: FRIGIDA delays flowering in Arabidopsis via a cotranscriptional mechanism involving direct interaction with the nuclear cap-
© The authors. This is an open access article distributed under the terms of the Creative Commons BY-NC-ND Licence.