G3, 2022, **12(5)**, jkac057
https://doi.org/10.1093/g3journal/jkac057
Advance Access Publication Date: 18 March 2022
Genomic Prediction
Featured

Genome-wide analysis of AAAG and ACGT *cis*-elements in *Arabidopsis thaliana* reveals their involvement with genes downregulated under jasmonic acid response in an orientation independent manner

Zaiba H. Khan,[1] Siddhant Dang,[2] Mounil B. Memaya,[3] Sneha L. Bhadouriya,[1] Swati Agarwal,[3] Sandhya Mehrotra,[1] Divya Gupta,[4] Rajesh Mehrotra[1,*]

[1]Department of Biological Sciences, Birla Institute of Technology and Science-Pilani, Zuarinagar, Goa 403726, India,
[2]Department of Biological Sciences, Birla Institute of Technology and Science-Pilani, Pilani, Jhunjhunu, Rajasthan 333031, India,
[3]Department of Computer Science and Information Systems, Birla Institute of Technology and Science-Pilani, Zuarinagar, Sancoale, Goa 403726, India,
[4]Faculty of Bioscience, Institute of Biosciences and Technology, Shri Ramswaroop Memorial University, Barabanki, Uttar Pradesh 225003, India

*Corresponding author: Department of Biological Sciences, Birla Institute of Technology and Science-Pilani, Zuarinagar, Goa 403726, India.
Email: rajeshm@goa.bits-pilani.ac.in

Abstract

Cis-regulatory elements are regions of noncoding DNA that regulate the transcription of neighboring genes. The study of *cis*-element architecture that functions in transcription regulation are essential. AAAG and ACGT are a class of *cis*-regulatory elements, known to interact with Dof and bZIP transcription factors respectively, and are known to regulate the expression of auxin response, gibberellin response, floral development, light response, seed storage proteins genes, biotic and abiotic stress genes in plants. Analysis of the frequency of occurrence of AAAG and ACGT motifs from varying spacer lengths (0–30 base pair) between these 2 motifs in both possible orientations—AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG, in the promoters and genome of *Arabidopsis thaliana* which indicated preferred orientation of AAAG $_{(N)}$ ACGT over ACGT $_{(N)}$ AAAG across the genome and in promoters. Further, microarray analysis revealed the involvement of these motifs in the genes downregulated under jasmonic acid response in an orientation-independent manner. These results were further confirmed by the transient expression studies with promoter-reporter cassettes carrying AAAG and ACGT motifs in both orientations. Furthermore, cluster analysis on genes with AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG motifs orientations revealed clusters of genes to be involved in ABA signaling, transcriptional regulation, DNA binding, and metal ion binding. These findings can be utilized in designing synthetic promoters for the development of stress-tolerant transgenic plants and also provides an insight into the roles of these motifs in transcriptional regulation.

Keywords: Dof; bZIP; *cis*-regulatory elements; transcription factor; promoter

Introduction

Cis-regulatory elements are short, functional noncoding DNA sequences, clustered within the promoter region of genes (Wray *et al.* 2003). These motifs play a key role in regulating the spatio-temporal expression of downstream genes at the transcriptional level. By providing binding sites to the sequence-specific cognate transcription factors (TFs), they lead to the activation or repression of the gene (Stormo 2000). The overall binding specificity of TFs to these short sequences depends upon a variety of factors including distance from the transcription start site (TSS), inter-motif distance, spacer sequence, copy number, and orientation related to the gene. Two motifs that we studied in this present study are AAAG and ACGT. It has been shown that the Dof binding proteins bind to AAAG sequences and the flanking sequences have limited effects (Yanagisawa 2002). However, this is not the case with ACGT core sequences, wherein flanking sequences influence the gene expression (Izawa *et al.* 1993). Dof proteins are shown to be involved in a multitude of functions. The role includes stress responses (Kang and Singh 2000; Ma *et al.* 2015a; Wang *et al.* 2017a), light responses (Yanagisawa and Sheen 1998), response to auxin (Baumann *et al.* 1999), gibberellins (Washio 2001), tissue-specific expression in endosperms (Vicente-Carbajosa *et al.* 1997; Mena *et al.* 1998), floral development (Rojas-Gracia *et al.* 2019). bZIP family of TFs is involved in abscisic acid (ABA) response (Zong *et al.* 2016), salicylic acid (SA) response (Li *et al.* 2019), auxin (Weiste *et al.* 2017), anaerobiosis (Sell and Hehl 2004), jasmonic acid (JA) (Liu *et al.* 2019), and UV responsiveness (Binkert *et al.* 2014).

bZIP and Dof interaction has been shown in the promoter of a seed storage protein gene (Zein) from maize. This promoter contains P-box (Yanagisawa 2000; Mehrotra *et al.* 2014) a highly

Received: December 13, 2021. **Accepted:** February 24, 2022

conserved 7-bp sequence *cis*-element with TGTAAAG conserved sequence and OCS element (ACGT) which are binding sites for Dof (P-box binding factors) and bZIP (O2) TFs, respectively (Chen *et al.* 1996). Members of the bHLH family of TFs such as MYC2, MYC3, and MYC 4 are also known to interact with Dof, These interactions are crucial for mediating JA transcriptional responses in *Arabidopsis thaliana* (Fernández-Calvo *et al.* 2011; Zhuo *et al.* 2020). The motivation for this study is to decipher the interaction between Dof and bZIP TFs if any in stress gene regulation through *cis*-element binding. In this study, we performed a genome-wide and a promoter-wide analysis by employing various in silico approaches to determine the frequency of occurrence and to identify patterns of occurrence of these motifs in tandem gradually increasing spacers from 0 to 30 base pair (bp), in both possible orientations-AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG, in the genome of model plant *A. thaliana*. Optimum inter-motif distance has great importance as promoter activation by ACGT is differentially regulated by the spacing between the 2 motifs (Mehrotra and Mehrotra 2010). Furthermore, a TF binding site (TFBS) analysis was performed to determine the binding sites for several TFs using ConSite: The functional software (Sandelin *et al.* 2004). Following this, microarray data analysis was performed using the genes upregulated/downregulated by hormones (ABA, auxin, ethylene, gibberellin, JA, and SA), environmental conditions (at baseline growth temperature, disease, drought, low water potential, at optimum photosynthetic temperature, salt, 20% inhibition from optimum photosynthetic temperature, 30% inhibition from optimum photosynthetic temperature) with the view to uncover the pattern of AAAG and ACGT distribution in the promoter of stress-induced genes (Mehrotra *et al.* 2013). The data suggested the involvement of AAAG and ACGT motifs in genes downregulated during JA responses in either orientation. Further, transient expression studies have been carried out to confirm the in silico findings. The gene ontology (GO) and cluster analysis revealed clusters of genes to be involved in ABA signaling, transcriptional regulation, DNA binding, and metal ion binding. The data obtained in this study highlights the importance of spacer lengths and orientation which can be used for designing inducible promoters. It also provides an understanding of the involvement of ACGT and AAAG sequences in the transcriptional regulation of stress-induced genes.

Materials and methods

DNA sequence retrieval

Genome-wide and promoter analysis was performed in *A. thaliana* genes (27,416) and promoters. For genome-wide analysis, the DNA sequence for all chromosomes was retrieved from TAIR (The Arabidopsis Information Resource, version 10; Huala *et al.* 2001; Lamesch *et al.* 2012). Further, Python code was used to extract 1 kb upstream promoter region of all genes across 5 chromosomes of *A. thaliana*.

The spacer frequency analysis was aimed at searching for co-occurring AAAG and ACGT elements of varying length increasing from 0 to 30 bp in the entire genome and 1 kb regions of promoters. We calculated the enrichment of these motifs by dividing frequency with the genomic coverage in genome and promoter regions respectively from 0 to 30 bp spacer in both orientations. As it has been hitherto reported that binding of corresponding TFs to the DNA sequence are usually spaced within 25–30 bp, we restricted our analysis to 30 bp spacer distance between AAAG and ACGT motifs (Mehrotra *et al.* 2013). The nucleotide sequence of each spacer between AAAG and ACGT motifs was extracted and the total number of occurrences for each spacer length was determined.

Spacer sequence analysis

Further, another Python code was executed on the extracted sequences to determine the frequency of occurrence of AAAG and ACGT motifs in tandem, for varying spacer lengths from 0 to 30 bp, in both possible orientations, i.e. AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG. Further, exact nucleotide sequences, for all spacer lengths from 0 to 30 bp of AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG, were also determined. Since flanking sites of the core sequence is essential for binding of TFs so we performed an analysis on flanking sites of the motifs: TAAAG and GACGTC, TGTAAAG and GACGTC in both the possible orientation. A comparison of core motifs to a longer motif was done in order to identify whether flanking sequences are predominating or not. There are many reports which show that any flanking sites can alter the binding site of TFs (Guiltinan *et al.* 1990; Morin *et al.* 2006). Further, to test the significance of AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG sequences, we used 12 random combinations of the 2 sequence sets: AAGA, AGAA, GAAA and CATG, GCAT, GTAC, TCAG as controls generated by shuffling AAAG and ACGT in all possible ways. Using the PLACE database, we ensured that each of these tetramer sequences is not conserved *cis*-element themselves (Higo *et al.* 1999). By performing a similar analysis on each control sequence, we compared the frequency of the AAAG $_{(N)}$ ACGT motif with the corresponding frequency of control sequences for the same $N = 0$–30 bp.

Frequency of occurrence of exact sequences for all spacers 0–30 bp were extracted, and consensus sequences for all spacer lengths were generated by calculating percentage A, C, G, T content at each position in all spacers. After calculating the percentage content of the 4 nucleotides, thresholds on percentage content are set for considering a nucleotide at a particular position to be conserved. Since the % G/C content of *A. thaliana* is about 36% (Mishra *et al.* 2009), we have taken the thresholds as 25% for G/C and 40% for A/T. If the frequency of G or C was more than 25% of the GC-content, we assumed that G or C is the preferred base. Similarly, if A or T had a frequency of occurrence of more than 40% at that particular position, then A or T is the preferred base.

Statistical analysis

A paired Student's *t*-test was conducted using the standard protocol to test the statistical significance of the results (McDonald 2014). For this, the frequency of occurrence from 0 to 30 bp between AAAG$_{(N)}$ACGT and ACGT$_{(N)}$AAAG were compared with various combinations of sequences AAGA, AGAA, GAAA, CATG, GCAT, GTAC, TCAG, and paired student t-test was applied to test the statistical significance of the sequences.

TFBS analysis

We wanted to decipher the biological significance of the *cis*-elements along with its spacer contributing to the high level of expression at the transcriptional level. For this, TF binding site analysis was performed using ConSite, a tool used to predict the TF binding site (TFBS) on regulatory elements (Sandelin *et al.* 2004). Prediction of TF binding sites on highly occurring spacer sequences between AAAG and ACGT motifs obtained through our spacer sequence analysis was carried out by prefixing a 139 nucleotide-long minimal promoter sequence (MPS) before those sequences (Kiran *et al.* 2006). Generally, MPS is a minimal promoter sequence that is merely required to achieve a basal level of gene expression in vivo. Other sequences along with the MPS are required to achieve a high level of gene expression. The sequence of MPS is as follows:

TCACTATATATAGGAAGTTCATTTCATTTGGAATGGACACG TGTTGTCATTTCTCAACAATTACCAACAACAACAAACAACAAA CAACATTATACAATTACTATTTACAATTACATCTAGATAAACA ATGGCTTCCTCC. TFBSs present on the highly occurring spacer sequences between AAAG $_{\mathbf{(0–30\,bp)}}$ ACGT and ACGT$_{\mathbf{(0–30\,bp)}}$AAAG motifs and MPS were compared and contrasted to check for binding suitability for TFs. The cut-off score for binding specificity was set to 80%.

Functional analysis

Microarray data of genes upregulated and downregulated under various environmental conditions (at baseline growth temperature, disease, drought, low water potential, at optimum photosynthetic temperature, salt, 20% inhibition from optimum photosynthetic temperature, 30% inhibition from optimum photosynthetic temperature) or by hormones (ABA, auxin, ethylene, gibberellin, JA, and SA) were retrieved from EBI expression atlas (Kapushesky et al. 2010) to analyze whether genes containing multiple core AAAG and ACGT *cis*-elements were upregulated or downregulated during these conditions. By comparing genes regulated by a condition with genes containing multiple AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG elements, we calculated the likelihood of occurrence by the following formula:

$$\text{Likelihood of occurrence} = X/Y \text{ for a particular spacer length}(N = 0\text{–}30\,\text{bp})$$

where $X = (A \cap B)/B$ and $Y = P(A)$;

A = event that a given gene is regulated (up/down) by a particular condition B = event that a given gene contains multiple AAAG-ACGT elements separated by *N* bps. Further, we calculated the overall likelihood of occurrence for each condition (for $N = 0$–30 bp). We set the threshold value as > 1.25 for the likelihood of occurrence for all conditions. This was followed by our previous work (Mehrotra et al. 2013).

GO and cluster analysis on the basis of functional characteristics of genes

We intended to perform GO analysis on the genes downregulated under the JA condition with ACGT$_{(N)}$AAAG and AAAG$_{(N)}$ACGT motifs within their promoters using DAVID (The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery, v6.8) tool (Huang et al. 2009b). We further refined the data by doing functional clustering on the data. The functional annotation clustering on GO term analysis results was conducted using the DAVID (Huang et al. 2009a). Each gene was represented using gene set enrichment analysis (GSEA) to identify GO: biological processes, molecular functions, and cellular components. For each orientation of motifs, clusters were created based on the similarity in functional analysis. Within each cluster, the genes were further sub-clustered based on the similarity between their category and term functions. An enrichment score was fetched for each cluster. Since within sub-clusters several genes were repeated, we took only unique gene identifiers and visualized the clusters. Similarly, cluster analysis of genes based on spacer sequence analysis was also conducted. As discussed above, spacer sequence analysis was conducted on genes for varying spacer lengths from 0 to 30 bp in both AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG orientations. We further performed multidimensional clustering on 74 genes downregulating under JA in both orientations based on their spacer lengths. We found 34 and 33 such genes in AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG orientations respectively. We also discard the spacers that have 0 occurrences across all genes making them insignificant for the clustering. We found 10 and 4 such spacers in AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG orientations, respectively. We used the K-means clustering algorithm to fit or train our model (Kanungo et al. 2002). Since the number of K clusters in K-means is user-generated and sometimes can lead to poor clustering if not selected based on the data. Therefore, we use Silhouette Score (Gat-Viks et al. 2003) to identify the number of optimal clusters. The silhouette score is a well-known and commonly practiced metric to identify the value of K. It uses the concept of mean inter-similarity and mean intra-cluster similarity and ranges from −1 to 1. It iteratively computes these distances/similarities for different values of K. The aim is to increase the intracluster similarity (cohesion) and decrease the inter-similarity (representing the separation from genes in other clusters).

Preparation of promoter-reporter cassette

A minimal promoter *Pmec* sequence containing a TATA-box, a TSS, and the reporter gene *gusA* cloned in the plasmid pUC19 was used. A random sequence of 50 nucleotides (GGATCCGGCTATGGCGGAGCAAGATTCACTCTGCGAGGCCAAAG CTTACCCCGGAAGGATCC), was cloned at the *BamH1* site of the *Pmec*. Further, this promoter-reporter cassette was cloned in the pBluescript SK (±) Phagemid (Stratagene, USA). Upstream of the 50 random nucleotide sequence, different combinations of the AAAG and ACGT motifs: AAAG$_{(N)}$AAAG; $N = 0$, 5, 25 bp, ACGT$_{(N)}$ACGT; $N = 0$, 5, 25 bp, AAAG$_{(N)}$ ACGT; $N = 0$, 5, 25 bp ACGT$_{(N)}$AAAG; $N = 0$, 5, 25 bp were inserted at the *Xba*I site (Supplementary Table 1 and Supplementary Fig. 1). The AAAG/ACGT–Pmec-*gusA* cassettes were coated on gold microparticles and bombarded onto tobacco leaves at 1100 psi, using a biolistic gun (Bio-Rad PDS-1000/He).

Transient expression studies under JA

To study the expression of the reporter gene *gusA* using different minimal promoter cassettes (MPSs) under JA, the bombarded leaves were kept in the Petri dishes with Hoagland solution supplemented with 50 μM MeJA (methyl jasmonate). After treatment, the plates were kept in the plant growth chamber at a temperature of 25°C, 16 h light/8 h dark period for 48 h. The transient expression studies using a biolistic system were performed as described by Mehrotra et al. (2005). In brief, treated leaves were incubated at 25°C and 16 h light/8 h dark photoperiod for 48 h. Subsequently, the leaves were immediately frozen, grounded in liquid nitrogen, and treated with GUS extraction buffer (50 mM Na_2HPO_4 pH 7.0, 1 mM EDTA, 0.1% v/v Triton X-100, 1.0 mM DTT and 0.1% SLS). The glucuronidase activity was assayed in cell-free extracts using 4-methyl umbelliferyl glucuronide (Jefferson et al. 1987). Relative fluorescence of 4-methylumbelliferone (MU) was determined using the Perkin Elmer Spectrofluorometer with excitation at 365 nm and emission at 455 nm. The expression data were analyzed statistically using a *t*-test.

Cloning of full-length promoter and introduction of mutation in ACGT region

We cloned the full-length promoter of PP2C-like gene (Gene id: AT5G59220) (Supplementary Fig. 2) using the Forward primer (5′-CGTCTAGAAAGTATTCACGCACCAAGGT-3′) and reverse primer (5′-GCTCTAGAACAAACACACTCCATCAC-3′). The mutated construct (bold in Supplementary Table 2 and Supplementary

Fig. 2) of it was cloned using site-directed primer: 5′-CCGGATCCATGAAAGTGATGACC*TAATTAGTTGTATTTATAG-3′. We carried out transient expression studies using the GUS as a reporter gene for the full-length promoter of the gene AT5G59220 and its mutated version where ACGT was mutated.

Results

AAAG $_{(\mathrm{N})}$ ACGT orientation is preferred over ACGT $_{(\mathrm{N})}$ AAAG orientation

We performed an analysis of promoters from *A. thaliana* to determine the frequency of occurrence of AAAG and ACGT motifs in tandem with varying spacer lengths from 0 to 30 bp between them. Results indicated preferred orientation of AAAG $_{(\mathrm{N}\,=\,0\text{–}30\,\mathrm{bp})}$ ACGT over ACGT $_{(\mathrm{N}\,=\,0\text{–}30\,\mathrm{bp})}$ AAAG orientation (Fig. 1a). However, we also got similar results when the analysis was performed across the genome of *A. thaliana* (Fig. 1b).

Enrichment of AAAG and ACGT motifs in promoters of *A. thaliana*

We found an average enrichment of 2 folds for most of the spacer length from 0 to 30 bp. We calculated the normalized frequency using the following formula :

$$= \frac{\text{Frequency of occurrence of the motif AAAG – ACGT or ACGT – AAAG (from 0 to 30 bp)}}{\text{Size of the genome}}$$

On normalizing the frequency of occurrence to the genome coverage we found an enrichment of around 2 folds in the promoter regions in AAAG-ACGT (range: 1.7–2.0) and ACGT-AAAG orientation (range: 1.5–2.0) (Fig. 1c).A Student *t*-test was conducted to test the significance of the results which showed that frequencies of occurrences are statistically significant (P-value ≤ 0.001, $t = 4.832$). At all spacer lengths, except $N = 1$, the frequency of occurrence of AAAG$_{(\mathrm{N})}$ACGT orientation is more than ACGT$_{(\mathrm{N})}$AAAG orientation. Moreover, it can be interpreted from Fig. 1a that at certain spacer lengths, there is a very high degree of correlation between the occurrence of common peaks and dips. Common peaks and dips were observed from spacer lengths 6–11, showing a good correlation between 2 orientations (Pearson correlation coefficient of 0.51). A paired student *t*-test was also performed between 2 spacer sequences from 6 to 11 which also showed the difference was significant (P-value $= 0.0012$). Apart from spacer lengths 6–11, common peaks were observed at $N = 15$ and 24, and a common dip was observed at $N = 25$. For the orientation AAAG $_{(\mathrm{N})}$ ACGT, the highest frequency of occurrence was at spacer lengths of $N = 6$, 10, and lowest for $N = 1$. Similarly, for ACGT $_{(\mathrm{N})}$ AAAG orientation, the highest frequency of occurrence was observed at a spacer length of $N = 10$, and lowest at

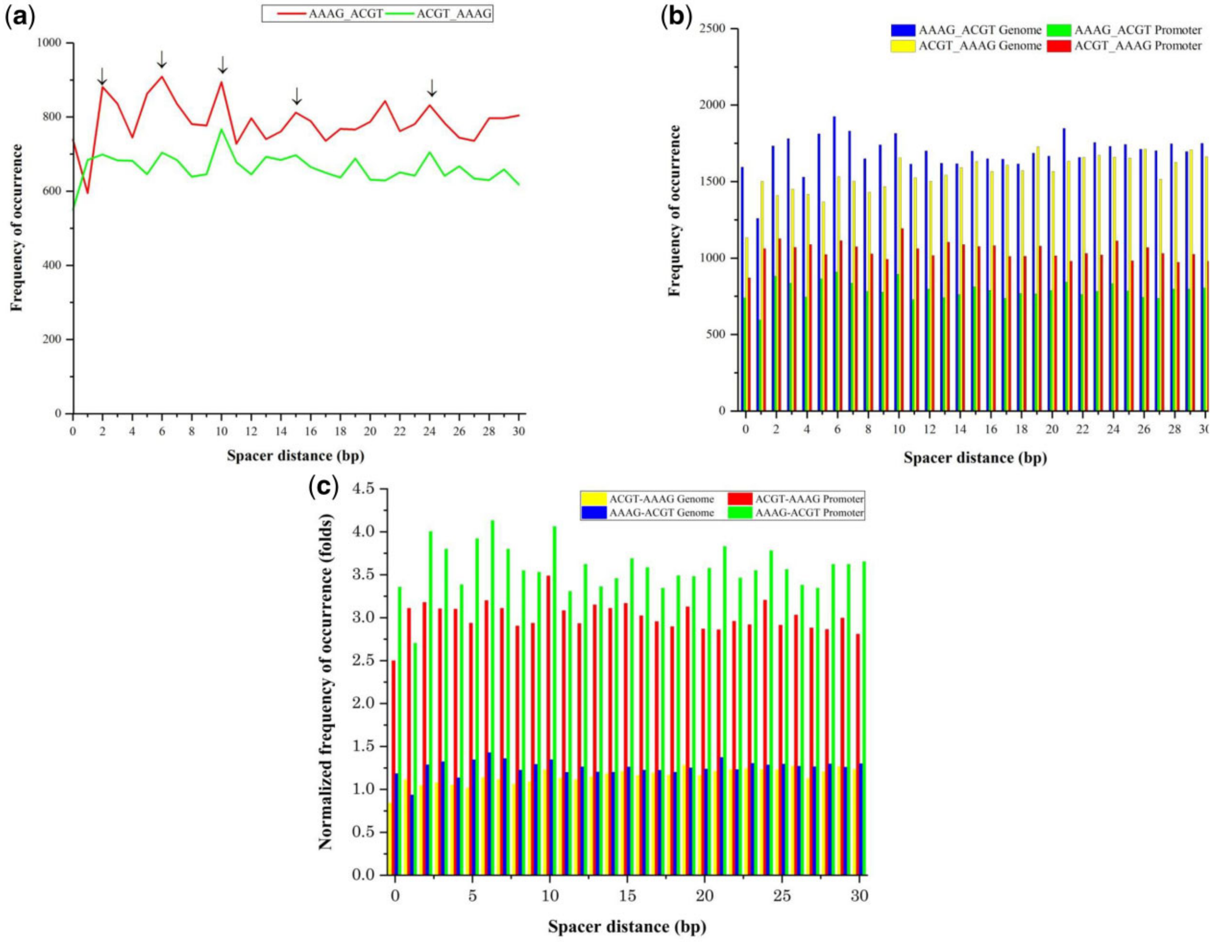

Fig. 1. a) Frequency of occurrence of AAAG $_{(0\text{–}30\,\mathrm{bp})}$ ACGT and ACGT $_{(0\text{–}30\,\mathrm{bp})}$ AAAG motifs in tandem with varying spacer lengths in the promoters of *A. thaliana*. Arrows exhibit common peaks and dips. b) Comparison of frequency of occurrence of AAAG $_{(0\text{–}30\,\mathrm{bp})}$ ACGT and ACGT $_{(0\text{–}30\,\mathrm{bp})}$AAAG in the genome and promoter regions of *A. thaliana*. c) Normalized frequency of occurrence of AAAG $_{(0\text{–}30\,\mathrm{bp})}$ ACGT and ACGT $_{(0\text{–}30\,\mathrm{bp})}$ AAAG in the genome and promoter regions of *A. thaliana* showing the enrichment of motifs in promoters.

$N = 0$ (Fig. 1a). Student 2-tailed t-test indicated the difference between the 2 orientations AAAG $_{(N)}$ ACGT, ACGT $_{(N)}$ AAAG was highly significant. The t-value at 30 df for the test samples was 3.73 while the t critical is 2.04. We also did analysis using flanking sites of AAAG and ACGT motifs with the following sequences TAAAG$_{(N)}$GACGTC and GACGTC$_{(N)}$TAAAG which revealed significant peaks (spacer = 2, 8, 10, 13, 15, 17, 21, 24, 29) (Supplementary Fig. 3). The higher frequencies at various spacer lengths indicate the importance of these spacers in the binding of TFs to these sites which we confirmed through TFBS analysis on 24 spacer lengths.

Low preference for AAAG and ACGT motifs in tandem as compared to other random motifs in tandem

Spacer sequence analysis and frequency of occurrence analysis were also conducted for some control motifs in tandem. Control motifs selected were AAGA, AGAA, and GAAA for AAAG generated by shuffling AAAG sequence in all possible ways likewise ACGT was shuffled to generate: CATG, GCAT, GTAC, and TCAG. It was ensured that control motifs are themselves not *cis*-regulatory elements and are completely random 4-nucleotide long sequences. The following charts (Fig. 2, a and b) show a comparison of frequencies of occurrence for AAAG and ACGT motifs in both orientations with control motifs from spacer lengths 0 to 30 bp. AAAG $_{(N)}$ ACGT (Fig. 2a) and ACGT $_{(N)}$ AAAG (Fig. 2b) both occur moderately less number of times as compared to control motifs, thereby, showing a limited preference for AAAG and ACGT motifs in tandem. Through paired student t-test we found statistically significant differences in these 2 orientations and among the control sequences only these AAAG $_{(N)}$ GCAT, TCAG $_{(N)}$ AGAA, are not statistically significant.

“A” is the most conserved nucleotide in spacer sequences

The spacer sequences were obtained for all 0–30 spacer lengths for both orientations. Further, consensus sequences, based on threshold conditions for percentage A, T, G, C content (Mishra *et al.* 2009) were drawn for all spacer lengths for both possible orientations. It was observed that nucleotide “A” occurred the most number of times; highlighting the fact that “A” is the most conserved nucleotide in the consensus spacer sequences. In the orientation AAAG$_{(N)}$ACGT, “A” nucleotide occurred at the first position in 14 spacer sequences (out of a possible 30), the third position in 4 spacer sequences, and at a few other positions in several other sequences. Whereas, the second most occurring nucleotide in consensus spacer sequences was “G” (Table 1). In the orientation ACGT$_{(N)}$AAAG, nucleotide “G” occurred at the first position in a total of 5 spacer sequences: 1, 3, 6, 25, 27, while, nucleotide “A” occurred the most at 6 places, and nucleotide “T” at 1 instance (Table 2).

A 24-bp spacer sequence *TTGGGCTTTCAAAATTGTTAACTC* between AAAG and ACGT has the maximum number of TF binding sites

TF binding sites on highly occurring spacer sequences, along with MPS were determined using ConSite software as described in the *Materials and Methods* section. 139-nucleotide long MPS was observed to have only 27 plant-specific TF binding sites for 9 TFs: Agamous-like MADS-box protein (AGL3) –2, *A. thaliana* homeobox-leucine zipper protein (Athb-1)–2, *A. thaliana* homeobox-leucine zipper protein (Athb)-5–2, and DOF3–1, Gibberellin- and abscisic acid-regulated MYB (GAMYB)–5, High mobility group box 1 protein (HMG-1)–6, HMG-I/Y–6, Avian myeloblastosis virus MYB TF from Petunia hybrida (MYB.ph3)–1, and SQUAMOSA (SQUA)–2.

A 24-nucleotide long spacer sequence between AAAG and ACGT–AAAG***TTGGGCTTTCAAAATTGTTAACTC***ACGT, was reported to have 43 TF binding sites, a total of 16 additional sites than MPS: Athb-1–1, DOF2–2, DOF3–4, HMG-1–1, and 2 each of Maize nuclear factor binding protein 1A (MNB1A), MYB.ph3, Prolamine-box binding factor (PBF), and SQUA. In the other orientation, a 14-nucleotide long spacer sequence, ACGTGGATGCTATTAT TAAAAG, was reported to have the maximum number of TF binding sites at 39 sites, a total of 12 more

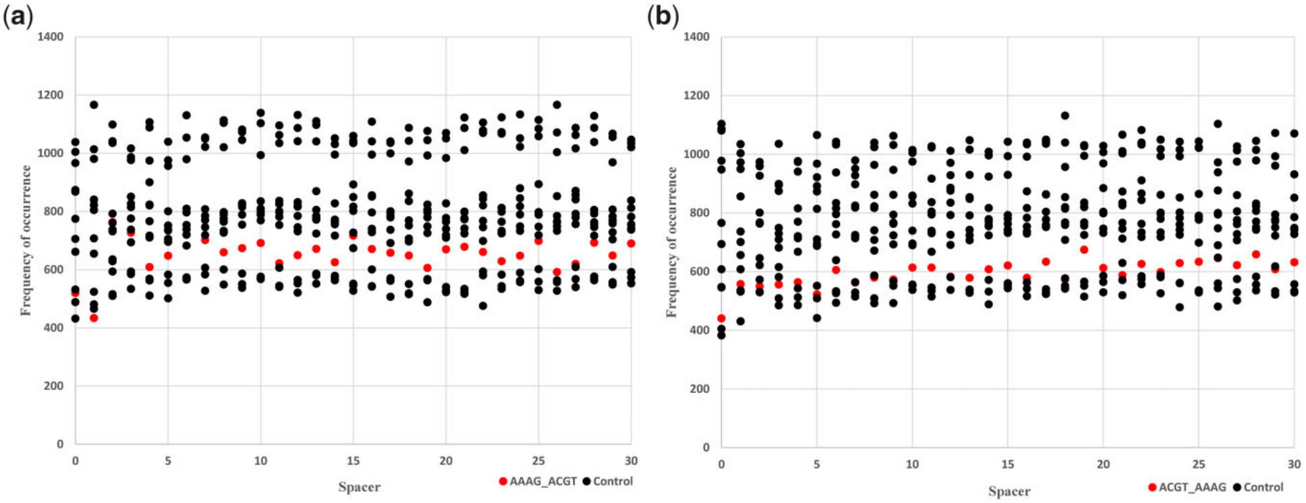

Fig. 2. Comparison of frequency of occurrences of AAAG and ACGT motifs to random 4 bp control sequences. a) Comparison of frequency of occurrences of AAAG $_{(N)}$ ACGT to random 4 bp control sequences. Frequency of occurrence of AAAG$_{(N)}$ACGT and several control motifs shows limited preference for AAAG$_{(N)}$ACGT motif as compared to random motifs. Red dots exhibit AAAG $_{(N)}$ ACGT while black dots represent various control sequences. b) Comparison of frequency of occurrences of ACGT $_{(N)}$ AAAG to random 4 bp control sequences. Frequency of occurrence of ACGT$_{(N)}$AAAG and several control motifs shows limited preference for ACGT$_{(N)}$AAAG motif as compared to random motifs. Red dots exhibit ACGT $_{(N)}$AAAG, while black dots represent various control sequences.

Table 1. Conserved sites in consensus spacer sequences for the orientation AAAG $_{(N)}$ ACGT.

Spacer length (bp)	Position in consensus spacer sequence																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	N																													
2	**G/C**	N																												
3	**G**	N	N																											
4	**G**	N	N	N																										
5	N	N	**G**	N	**C**																									
6	N	N	N	N	N	N																								
7	N	N	N	**A**	N	N	N																							
8	**A**	N	N	N	N	N	N	N																						
9	**A**	N	N	N	N	N	N	N	N																					
10	**A**	N	N	N	N	N	N	N	**A**	N																				
11	**A**	N	**A**	N	**A**	N	N	N	N	N	N																			
12	N	N	N	N	N	N	N	N	N	N	N	N																		
13	**A**	N	N	N	N	N	N	N	N	N	N	N	N																	
14	**A**	N	**A**	N	**A**	N	N	N	N	N	N	N	N	N																
15	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N															
16	N	N	**A**	N	N	N	**A**	N	N	N	N	N	N	N	N	N														
17	N	N	N	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N													
18	**A**	N	N	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N												
19	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N											
20	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N										
21	N	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N									
22	N	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N								
23	N	N	N	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N							
24	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N						
25	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N					
26	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N				
27	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			
28	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N		
29	N	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	**A**	N	**A**	N	
30	**A**	N	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Nucleotides A and T are marked in bold, while nucleotides G and C are marked in bold italics.

than MPS-1 sites of SQUA, and 1 site each for AGL3, Athb-1, Athb-5, bZIP 910, DOF2, DOF3, MNB1A, MYB.ph3, and PBF. Many other sequences between spacer 2, 3, 4, 5, 6, 7, 10, 21, 23, 24, 29, 30 in AAAG$_{(N)}$ACGT orientation while 0, 1, 2, 10, 14, 16, 17, 19, 26, 27, 29 in ACGT$_{(N)}$AAAG orientation had more number of TF binding sites than MPS (Supplementary data sheet 1). Spacer 3, 4, 5, 7, 23, 24, 29 in AAAG $_{(N)}$ ACGT orientation and spacer 0, 1, 10, 14, 16, 19, 26, 27, 29 had statistically significance than other sequences (Supplementary data sheet 1). This means that along with the distance between binding motifs there has been a selection for the sequence of the spacer in transcriptional regulation. We performed a similar TFBS analysis on *Glycine max* and found these TFs binding: AGL3, Athb-1, Athb-5, bZIP 910, bZIP 910, DOF2, DOF3, MNB1A, MYB.ph3, PBF, SQUA, Agamous, AGL3, GAMYB, HMG-IY, HMG-1, suggesting that the process of transcription regulation has essentially being conserved in eukaryotes (Supplementary data sheet 1).

JA response is mediated by AAAG and ACGT repeat elements preferentially in an orientation independent manner

We identified the genes upregulated, downregulated by specific environmental conditions, or by hormones containing an upstream AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG elements. By comparing genes regulated by a condition with genes containing multiple ACGT $_{(N)}$ AAAG and AAAG $_{(N)}$ ACGT elements, we calculated the likelihood of occurrence for each condition (Supplementary data sheets S2 and S3). We calculated the overall likelihood of occurrence for each condition, with a likelihood of 1 being that of random chance. Accordingly, conditions greater than 1.25 or above were selected. Results unveil that likelihood of occurrence value for AAAG$_{(N)}$ACGT and ACGT$_{(N)}$AAAG motif in promoters downregulated in response to JA is 1.29 and 1.35, respectively (Fig. 3, a and b). This points out that AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG being high in occurrence might play a crucial role in the promoters of genes downregulated during the JA response in both the orientation as their values are above the threshold by a significant margin. These observations suggest that AAAG and ACGT motifs are involved in the regulation of genes participating in abiotic and biotic stress conditions independent of their orientation. It was observed that the value of the likelihood of occurrence of none of the other conditions displayed any significant deviation.

GO analysis reveals the presence of AAAG and ACGT in genes involved in ABA signaling, metal-ion binding, and transcription regulation

GO term analysis revealed that among 34 genes with AAAG$_{(N)}$ACGT orientation, 33 overlapping genes (97.1%) were of cellular components, 28 overlapping genes (82.4%) were found to be involved in biological processes and 24 overlapping genes (70.6%) were associated with molecular function in the cell. In the case of 33 genes with ACGT$_{(N)}$AAAG orientation, all 33 genes (100%) were represented in cellular component, 24 overlapping genes (72.7%) were found to be associated with biological processes and 22 overlapping genes (66.7%) were found to be involved with molecular functions. We further refined the data by doing functional clustering on the data. Figure 4 represents 4 clusters for ACGT $_{(N)}$ AAAG orientation. Terms attributed to ABA signaling, transcription regulation, extracellular region, and

Table 2. Conserved sites in consensus spacer sequences for the orientation ACGT$_{(N)}$AAAG.

Spacer length (bp)	Position in consensus spacer sequence																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	**G**																													
2	N	N																												
3	**G**	N	N																											
4	N	N	N	N																										
5	N	N	N	N	N																									
6	**G**	N	N	N	N	N																								
7	N	N	N	N	N	N	N																							
8	N	N	N	N	N	N	N	N																						
9	N	N	N	N	N	N	N	N	**A**																					
10	N	N	**A**	N	N	N	N	N	N	N																				
11	N	N	N	N	N	N	N	N	N	N	**A**																			
12	N	N	N	N	N	N	N	N	N	N	N	N																		
13	N	N	N	N	N	N	N	N	N	N	N	N	N																	
14	N	N	N	N	N	N	N	N	N	N	N	N	N	**A**																
15	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N															
16	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N														
17	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N													
18	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N												
19	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N											
20	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N										
21	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N									
22	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	**A**								
23	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N							
24	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N						
25	**G**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N					
26	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	**A**	N				
27	**G**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N			
28	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N		
29	N	N	N	N	N	N	N	N	N	N	N	**T**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
30	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Most conserved nucleotides A and T are marked in bold, while most conserved nucleotides G and C are marked in bold italics.

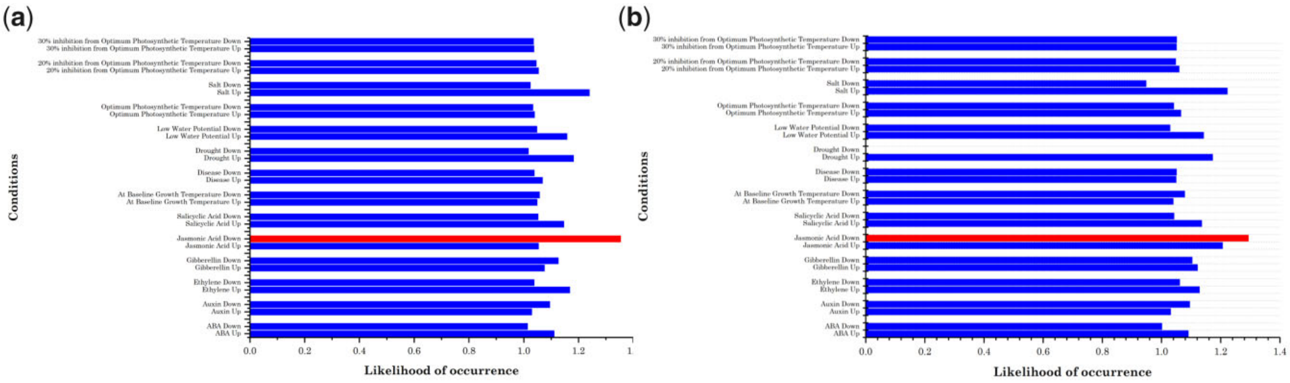

Fig. 3. AAAG and ACGT containing promoters and their regulation under specific conditions. Using microarray-based expression data in *A. thaliana*, an analysis of promoters upregulated/downregulated/up-down regulated under 14 different conditions: environmental conditions (at baseline growth temperature, disease, drought, low water potential, at optimum photosynthetic temperature, salt, 20% inhibition from optimum photosynthetic temperature, 30% inhibition from optimum photosynthetic temperature) and hormones (ABA, auxin, ethylene, gibberellin, JA, SA) revealed downregulation of AAAG$_{(N)}$ACGT containing promoters by JA. a) AAAG $_{(N)}$ ACGT containing promoters and their regulation under specific conditions showing downregulation in response to JA. b) ACGT $_{(N)}$ AAAG containing promoters and their regulation under specific conditions showing downregulation in response to JA.

metal-ion binding were highly represented in the genes downregulated under JA with the ACGT-AAAG orientation with statistical significance. This reveals that all clusters are balanced since they have evenly distributed genes. The graph illustrates the respective gene in the clusters. Furthermore, each cluster was assessed using an Enrichment score (Huang *et al.* 2009a) (Supplementary Table 3). Overlapping gene ids among all clusters are shown in a 4-set Venn diagram (Supplementary Fig. 4). The cluster diagram shows 2 common genes (AT2G38390, AT4G19230) among clusters C1, C3, and C4, 3 common genes

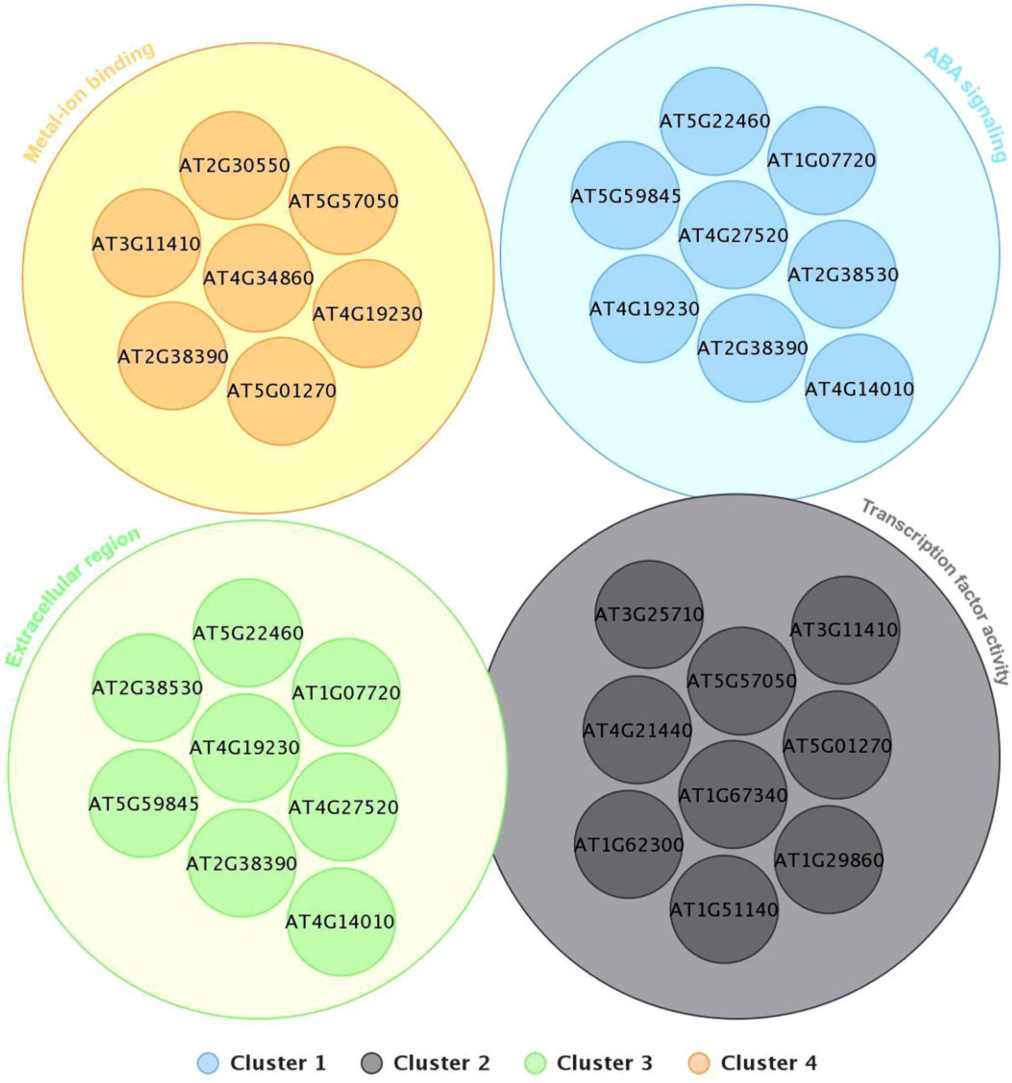

Fig. 4. Cluster representation of JA downregulated genes with ACGT $_{(N)}$ AAAG orientation based on functional characteristics. Cluster 1 represents genes involved with ABA signaling, cluster 2: TF activity, cluster 3: extracellular region, cluster 4: metal-ion binding.

AT5G57050, AT5G01270, AT3G11410 between cluster 2 and 4 and C1 and C3 have 6 common genes AT2G38390, AT5G22460, AT4G19230, AT4G14010, AT4G27520, AT5G59845, AT2G38530, AT1G07720 while C2 and C4 have a total 6 and 2 unique genes respectively.

Figure 5 represents the cluster diagram for the genes downregulated under JA with AAAG $_{(N)}$ ACGT orientation. As the figure reveals, in AAAG $_{(N)}$ ACGT orientation, out of 35 genes, no genes were identified as outliers and all of them were clustered into 6 groups. While clusters C1, C2, C3, and C4 are well-balanced (relatively evenly distributed), clusters C5 and C6 are significantly imbalanced consisting of 3 and 4 genes, respectively. In both the orientations of AAAG and ACGT [AAAG$_{(N)}$ACGT and ACGT$_{(N)}$AAAG], only 1 cluster was found to be significant with an enrichment score >2 (Supplementary Tables 3 and 4). This happened due to the overlapping properties among genes resulting in a soft clustering. A 6-set Venn diagram of the overlapping genes within clusters represents that every cluster has at least one overlapping gene with other clusters (Supplementary Fig. 5). However, no common gene has been found out among all the clusters. Cluster pairs of (C1, C2) have a total of 6 overlapping genes (AT5G01270, AT3G11410, AT5G59220, AT1G18100, AT5G57050, and AT5G52300) among which AT5G01270, AT5G59220 are also present in C4. In cluster pairs of C2, C3 there are 2 genes in common: AT2G38390, AT2G38530. Furthermore, other than the cluster triplets of (C1, C2, and C4) which have 2 genes in common (i.e. AT5G01270 and AT5G59220), no other triplets or larger intersection sets have a common gene. In both orientations, some genes were identified with 0 The individual clusters C1, C2, C3, C4, C5, and C6 have 1, 3, 3, 5, 1, and 0 unique genes.

A larger Silhouette value indicates that clusters are significant. Silhouette score elbow curve for AAAG $_{(N)}$ ACGT and ACGT $_{(N)}$ AAAG orientations are shown in Supplementary Figs. 6 and 7 respectively. The optimal value of K for genes in AAAG $_{(N)}$ ACGT orientation is 5 (Supplementary Fig. 6). Even though $K = 7$, 8, and 9, have higher Silhouette scores than $K = 5$ but those are bad picks for the optimal value of K due to wide fluctuation in the score. The fit time represents the time taken (in seconds) to fit the unsupervised model on the given data. Figure 6 shows the

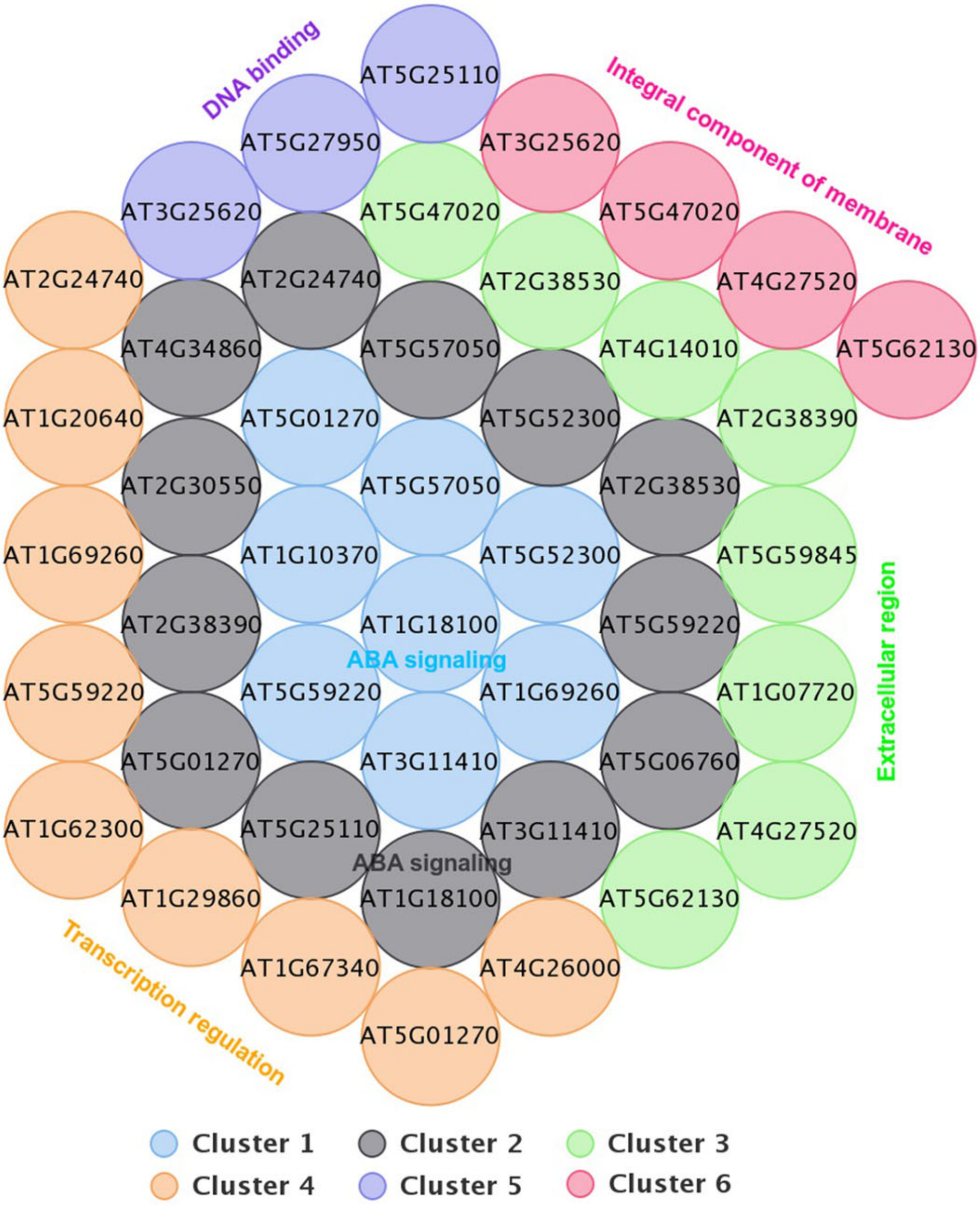

Fig. 5. Cluster representation of JA downregulated genes with AAAG $_{(N)}$ ACGT orientation based on function characteristics. Cluster 1 represents genes involved with ABA signaling; cluster 2: extracellular region; 3: transcription regulation; cluster 4: metal-ion binding; cluster 5: DNA binding; cluster 6: integral component of membrane.

cluster representation of JA genes downregulated in AAAG $_{(N)}$ ACGT orientation. Among 34 genes: 18, 4, 3, 5, and 5 genes were grouped in clusters 0, 1, 2, 3, and 4 respectively with a mean Silhouette score of 0.242. Figure 6 reveals that there are no outlier genes with any similarity from other genes. Each cluster has a minimum of 3 genes. Furthermore, except for cluster 0, all remaining clusters are well-shaped and equally distributed.

We found 7 optimal clusters in ACGT $_{(N)}$ AAAG orientation (Supplementary Fig. 7). The Silhouette score is small due to the small number of resultant genes and sparse data with respect to 27 spacer values. The time graph shows that the algorithm took minimum time to fit the model for $K = 7$. Figure 7 shows the cluster representation of genes with respect to ACGT $_{(N)}$ AAAG orientation. The graph reveals that the clusters C1 and C3 have only one gene (AT3G11410 and AT4G21440, respectively) making them an outlier. AT3G11410 has occurrences on a total of 5 spacers, i.e. 14, 16, 18, 20, and 22 whereas; AT4G21440 has occurrences on 13 and 22 spacer values. The graph reveals that cluster 0 has a maximum number of genes (18) having similar behavior evident for higher values of spacers. Similar to AAAG$_{(N)}$ACGT orientation, 3 clusters are well-shaped and equally distributed identifying one gene as an outlier and the remaining genes as highly similar forming a dense cluster.

In order to gain insight into the regulatory functions of spacer sequences (0–30 bp) between the motifs, we identified the location of AAAG and ACGT motifs in both the orientation: AAAG $_{(N = 0–30)}$ ACGT (Supplementary Table 5) and ACGT $_{(N = 0–30)}$ AAAG (Supplementary Table 6) respectively on the promoters of the

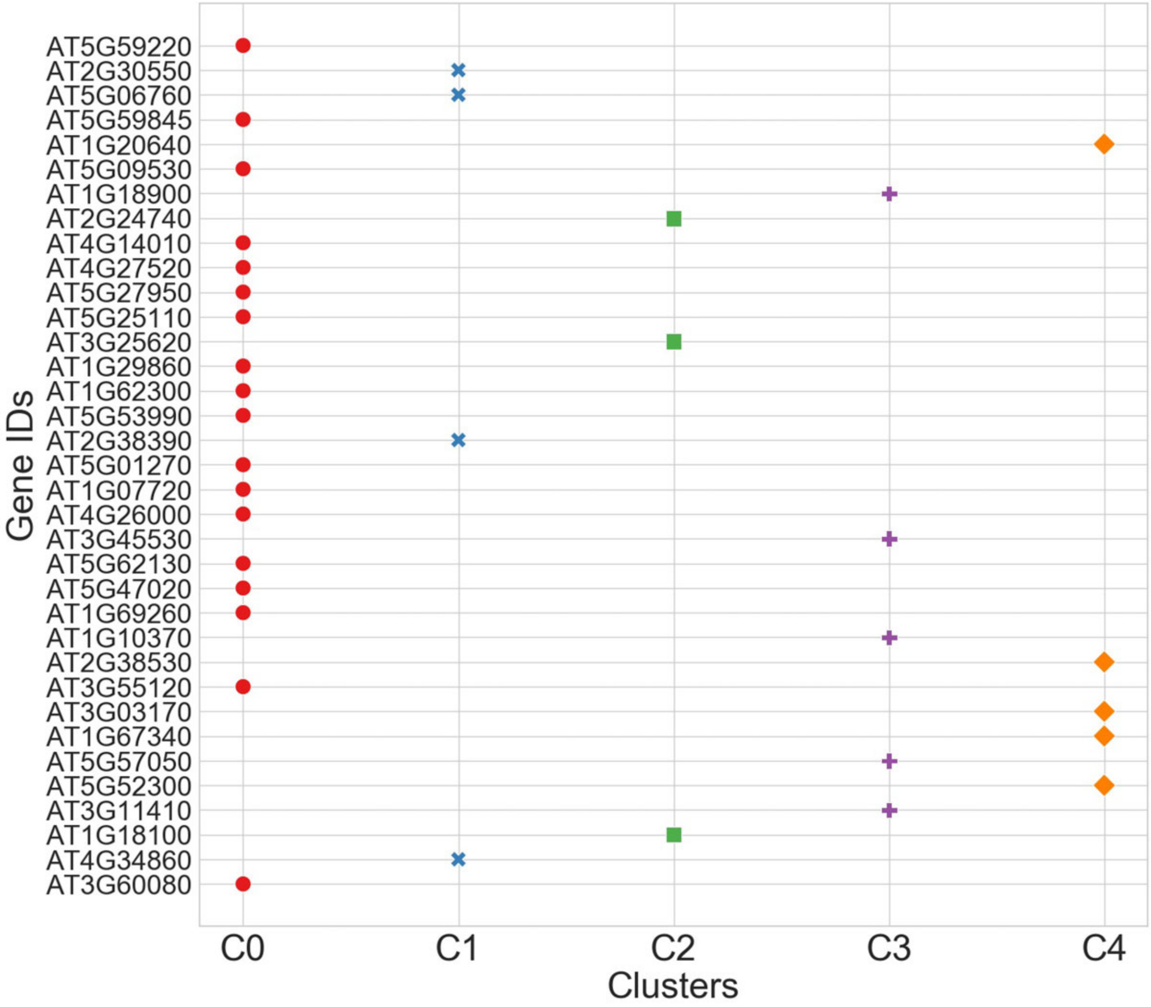

Fig. 6. Scatter plot illustrating clustering of genes downregulated under JA response with AAAG $_{(N)}$ ACGT orientation.

genes downregulated under JA response followed by their categorization according to the spacer lengths. For AAAG-ACGT orientation, 4 genes each were identified for the spacer 7 and 10. Group of the genes at spacer 7, are majorly cellular components. At spacer 10, 3 out of 4 genes are involved in the biological process with enriched term metabolic process.

For ACGT-AAAG orientation, we also got the highest frequency of occurrences i.e. 4 at the spacer length 20, 22, 23. GO term analysis reveals that these 4 genes at spacer length 20 are cellular components, enriched with the term metal-ion binding. Genes at spacer 22 are involved in biological processes, enriched with the term responses to osmotic stress. All 4 genes at spacer 23 are cellular components, enriched with the term nucleus.

Transient expression analysis of AAAG and ACGT reporter cassette

The expression of the *gusA* gene in tobacco leaves bombarded with promoter-reporter cassette carrying a single AAAG and ACGT motif, in tandem and separated by a spacer of 5 and 25 nucleotides under JA was analyzed. The expression of the reporter gene driven by AAAG with ACGT motif separated by 5 and 25 nucleotides in both orientations was also studied. The leaves bombarded with a 50 + Pmec reporter cassette without any AAAG or ACGT motif were regarded as a control.

The data (as shown in Table 3) clearly shows a gradual increase in the reporter gene expression, expressed under the effect of a minimal promoter (50 + Pmec) carrying AAAG or ACGT activator motif as a single copy or 2, in tandem separated by a spacer length of 5, 25 in uninduced conditions. Under JA conditions, the constructs carrying (AAAG) (ACGT) in tandem or separated by a spacer length of 5, 25 decreased the reporter gene expression by 2.58, 2.82, 3.05 folds respectively. A similar trend of reduction was also observed for (ACGT) (AAAG) with the increasing spacer length of 0, 5, 25 bp by 2.34, 3.02, 3.48 folds, respectively, as compared to the control construct.

GUS activity of the protein phosphatase 2C (PP2C)-like promoter (AT5G59220) under JA response in tobacco leaves

Fluorometric assay was performed for the measurement of GUS activity. The expression of GUS in the full-length PP2C-like promoter (AT5G59220) was found to be high in uninduced conditions while in response to MeJA, approximately ∼6.5-fold decrease in GUS expression was observed (4048 ± 286) (Table 4). However, in the mutated construct, the readings are comparable in both the uninduced (18,000 ± 768) and induced (16,072 ± 207) condition revealing a similar yet not significant expression.

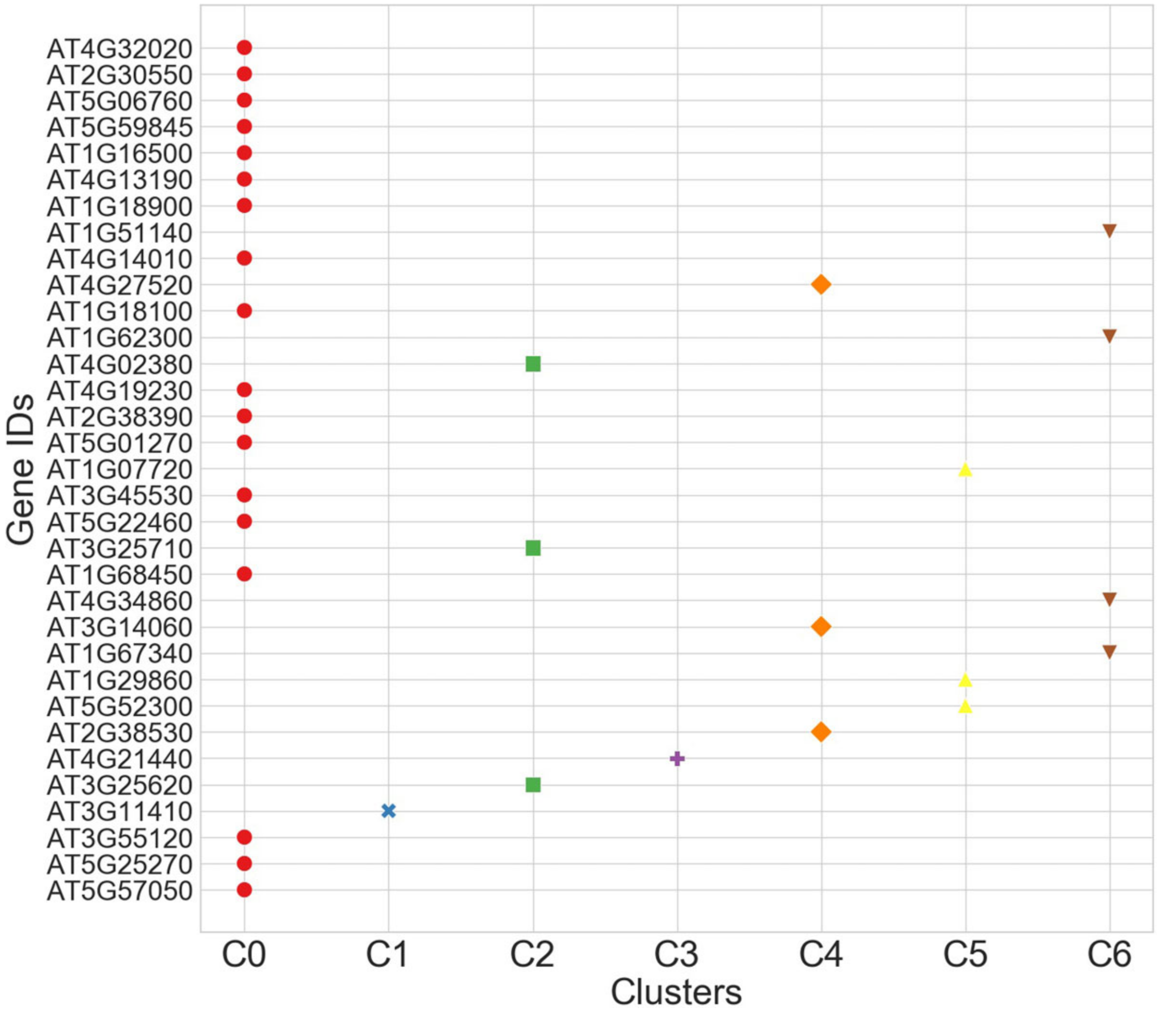

Fig. 7. Scatter plot illustrating clustering of genes downregulated under JA response with ACGT $_{(N)}$ AAAG orientation.

Table 3. Transient expression data showing promoter constructs carrying AAAG and ACGT motifs.

S.No.	Promoter cassette	Uninduced (pmoles/min/mg protein ± SD)	Fold activity as compared to 50 + Pmec	JA (pmoles/min/mg protein ± SD)	Fold induction	P-value
1	50 + Pmec	1,797 ± 26.8	1	1,805 ± 140	1.00	P = 0.477
2	(ACGT) + 50 + Pmec	2,683 ± 107	1.49	2,988 ± 122	1.11	P = 0.11
3	(AAAG) + 50 + Pmec	2,530 ± 95	1.40	3,013 ± 89	1.19	P = 0.06
4	$(ACGT)_2$ + 50 + Pmec	3,208 ± 130	1.78	8,280 ± 137.2	2.58	**P = 0.0008**[b]
5	$(AAAG)_2$ + 50 + Pmec	2,980 ± 99.8	1.65	7,623 ± 182	2.55	P = 0.01[a]
6	(ACGT)(AAAG) + 50 + Pmec	4,282 ± 102	2.38	1,829 ± 112	0.42	P = 0.013[a]
7	(AAAG)(ACGT) + 50 + Pmec	4,462 ± 119	2.48	1,726 ± 112	0.38	P = 0.0134[a]
8	$(ACGT)_{N5}$ (AAAG) + 50 + Pmec	5,999 ± 180.22	3.33	1,983 ± 169	0.33	P = 0.013[a]
9	$(AAAG)_{N5}$ (ACGT) + 50 + Pmec	5,860 ± 129	3.26	2,072 ± 113	0.35	P = 0.010[a]
10	$(ACGT)_{N25}$ (AAAG) + 50 + Pmec	6,300 ± 228	3.50	1,807 ± 89	0.28	P = 0.012[a]
11	$(AAAG)_{N25}$ (ACGT) + 50 + Pmec	5,900 ± 193	3.28	1,930 ± 118	0.32	P = 0.012[a]

[a] Indicates P-value < 0.05 and [b]indicates P-value < 0.001.

Table 4. Transient expression data of the promoter-reporter constructs.

S. No.	Promoter cassette	Uninduced (pmoles/min/mg protein ± SD)	JA (pmoles/min/mg protein ± SD)	Fold reduction	P-value
1	Full-length promoter	26,000 ± 1038	4,048 ± 286	6.4	0.001[a]
2	Mutated construct	18,000 ± 768	16,072 ± 207	5.85	0.07

[a] Indicates P-value < 0.001.

Discussion

Cis-regulatory elements, AAAG and ACGT are known to play a vital role in regulating several processes including floral development (Rojas-Gracia *et al.* 2019), gibberellin response (Washio 2001), light response (Yanagisawa and Sheen 1998), tissue-specific expression (Vicente-Carbajosa *et al.* 1997), biotic and abiotic stress responses (Le Hir and Bellini 2013; Mehrotra *et al.* 2013; Ma *et al.* 2015; Wang *et al.* 2017). Previously we have done analysis on ACGT (Mehrotra *et al.* 2013) and AAAG (Mehrotra *et al.* 2014) *cis*-elements separately. In this present analysis, we have taken both the elements together with an attempt to decipher their interaction if any in stress response as these 2 *cis*-regulatory elements are known to interact with each other synergistically to regulate the expression of certain genes in specific stress conditions and also in the expression of seed storage protein genes (Singh 1998). Analyses performed on the extracted data revealed AAAG $_{(N)}$ ACGT as the preferred orientation of these motifs for TF binding. An interesting observation was that the spacer lengths 6 and 10 occur at a maximum frequency in AAAG$_{(N)}$ACGT orientation, while spacer length 10 also occurs maximally in ACGT$_{(N)}$AAAG orientation. These sequences occur more than the other spacer sequences suggests that these lengths might serve as optimal lengths for TF binding (Kim and Maniatis 1997) also observed from TF binding site analysis (Supplementary data sheet 1). This indicates that along with the distance between motifs there has been a selection for the sequence of the spacer in transcriptional regulation. Further, 2 orientations show a correlation in peaks and dips for certain consecutive spacer lengths, as mentioned above, also point to the reversibility of TF binding or coevolution because of the high synergistic functionality of the 2 motifs (Yamamoto *et al.* 2006). Also, a probable reason for the dips could be that certain lengths might cause steric hindrance in the binding of TFs explaining the reason for the lower preference for certain specific sequences despite an optimal length (Yamamoto *et al.* 2006; Spitz and Furlong 2012). We found 1.5–2.0 folds enrichment of AAAG and ACGT motifs in promoters of *A. thaliana* as compared to its genome. On comparing the occurrence of 2 motifs—AAAG and ACGT in tandem with the other control sequences in the promoter region, it was revealed that these 2 motifs in tandem are less preferred than most of the control cases (Supplementary data sheet 1). From spacer sequence analysis, consensus sequences were drawn for each spacer length in both possible orientations. Consensus sequences showed a high degree of conservation of nucleotide “A.” In addition to this, “A” also occurred at the first position in 14 spacer sequences in the orientation AAAG $_{(N)}$ ACGT. Also, “A” was found at the last position in the spacer sequence in 4 spacer lengths. These observations are in synchronization with the other inferences as AAAGA and AAAAG act as binding sites for several Dof family of TFs (Lijavetzky *et al.* 2003). Further, it was revealed that nucleotide “G” is preferred at the first position in 4 spacer sequences in the orientation ACGT $_{(N)}$ AAAG.

Microarray analysis was carried out to determine the function of these motifs in upregulated or downregulated genes under environmental conditions and in response to phytohormones. Since AAAG and ACGT motifs are enriched in the genes getting downregulated in response to JA it could be deduced that these motifs are involved in the downregulation of genes taking part in JA responses in an orientation independent manner. The data suggest that the binding of the repressor seems to be distance-dependent rather than sequence-dependent so as to bring downregulation under JA response in an orientation-independent manner (Teif and Rippe 2011). GO analysis followed by cluster analysis revealed that the genes downregulated under JA conditions with AAAG and ACGT motifs in both the orientations are associated mainly with biological functions (ABA signaling, transcriptional regulation) cellular component (extracellular region) and molecular functions (metal-ion binding).

Cluster analysis on genes downregulated under JA with AAAG $_{(N)}$ ACGT and ACGT$_{(N)}$ AAAG motifs were conducted for all spacer lengths. In ACGT$_{(N)}$ AAAG orientation, cluster 0 has the maximum number of genes. These genes are mostly present on one spacer only. For example, AT5G22460, AT3G45530, AT5G01270, AT2G38390, AT4G19230, AT1G18100, AT4G14010, AT1G18900, AT4G13190, AT1G16500, AT5G59845, AT5G06760, AT2G30550, and AT4G32020 are present only on 7, 30, 2, 25, 27, 24, 12, 18, 4, 10, 26, 22, 21, and 11 respectively. Cluster 0 also has 4 genes (AT5G57050, AT5G25270, AT3G55120, and AT1G68450) which are present in more than 1 spacer but have similar spacers as other genes in their cluster. Surprisingly, there is no cluster that has genes all present on only 1 spacer length. We found an interesting insight that all the IDs present in cluster 2 (AT3G25620, AT3G25710, and AT4G02380) are present on spacer lengths 9 and 29 with 1 overlapping gene. AT3G11410 and AT4G21440 are identified as outliers in clusters 1 and 3 since these are the only genes with unique behavior. AT3G11410 appears on 5 spacer lengths while AT4G21440 appears twice on 1 spacer itself. In AAAG $_{(N)}$ ACGT orientation, cluster 1 has the maximum number of genes grouped together i.e. 18. All the genes are present on at least 1 spacer while only 2 IDs are present on 2 spacers i.e. AT3G60080 (spacer 3 and 19), AT3G55120 (spacer 3 and 18). In cluster 2, all genes (AT1G18100, AT3G25620, and AT2G24740) are present on exactly 1 spacer length, i.e. 15 and do not appear anywhere else. In cluster 4, all the genes except AT1G20640 are present on 2 spacers and share a common spacer too. For example, while AT1G67340, AT3G03170, AT1G20640 are present on spacer length 2, AT3G03170 if further present with AT5G52300 and AT2G38530 at spacer 28. Except for AT5G57050 AT4G34860, all IDs in cluster 3: AT3G11410, AT1G10370, AT3G45530, AT1G18900 are present on spacer 7 while AT5G57050 is present on spacer 29 and 30. Similarly, except AT4G34860, all gene ids in cluster 1: AT2G38390, AT5G06760, AT2G30550 are present at 10 spacers while AT4G34860 is also present at spacer 20. This analysis reveals that most of the genes in the clusters belong to a particular spacer length suggesting the importance of spacer length between AAAG and ACGT in the promoters of the genes involved in the biological processes in plants or function as cellular components.

The transient expression studies strengthened our in silico findings that the AAAG and ACGT motifs downregulate the gene expression under JA responses irrespective of their orientation. The data further suggests that the AAAG and ACGT motifs might be acting as a negative regulator leading to reduced reporter gene expression in response to JA in (AAAG)$_{N5}$(ACGT), (AAAG)$_{N25}$(ACGT), (ACGT)$_{N5}$(AAAG), (ACGT)$_{N5}$(AAAG) constructs. Our transient expression studies on full-length PP2C- like promoter and its mutated version indicate that the region between AAAG (TGATG) ACGT of this promoter is responsible for downregulation under JA treatment. As the expression of any gene, largely relies upon the *cis*-regulatory elements arranged within the promoter regions, so in order to modulate the expression of the transgene, the role of AAAG and ACGT motifs as the negative regulator in JA responses could be taken into consideration in the designing of synthetic promoters. The results of this study can assist in the promoter designing, as spacer lengths and specific sequences between AAAG and ACGT motifs for binding are required for triggering a stress response (Mehrotra *et al.* 2017). Conservation of spacer lengths of sequences between motifs and their copy number is vital for promoter designing (Mehrotra *et al.* 2011, 2017; Dhatterwal *et al.* 2019).

Conclusion

This study was aimed at analyzing patterns of ACGT and AAAG *cis*-regulatory elements in the *A. thaliana* genome. We established preferences for specific orientation, specific spacer lengths, and specific conserved nucleotides in the promoter region for the above-mentioned 2 motifs in tandem. Further, the interplay of these 2 motifs was observed to have an effect on downregulation of genes under JA responses in an orientation-independent manner. This information would be crucial for designing stress-inducible synthetic promoters for the development of highly productive and stress-resistant transgenic crops.

Data availability

The data generated/analyzed during the study are available in the supplementary data sheets, supplementary figures, and supplementary tables.

Supplemental material is available at G3 online.

ZHK performed the experiments and drafted the article. SD and MBM helped in in-silico and statistical analysis. SLB helped in editing the manuscript. SA performed clustering on gene data and generated inferences. DG helped in gene expression studies. RM conceived the original idea and SM investigated and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Funding

RM and SM are thankful to the Department of Science and Technology (DST) Government of India under the Science and Engineering Research Board (SERB) scheme for financial support. This work was supported by the DST-SERB project bearing file number EMR/2016/002470 sanctioned by the government of India to SM and RM.

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Acknowledgments

We would like to thank the Department of Biological Sciences, Birla Institute of Technology and Science Pilani, Pilani campus and Goa campus, India for their constant support and the facilities provided. We are also thankful to DST for providing Inspire fellowship to ZHK. We are thankful to Sri Hari Raju Sangaraju, Kanikireddy Muralidhar Reddy, D. Akash Reddy, and Atul Rai for helping in spacer frequency analysis.

Literature cited

Baumann K, De Paolis A, Costantino P, Gualberti G. The DNA binding site of the Dof protein NtBBF1 is essential for tissue-specific and auxin-regulated expression of the rolB oncogene in plants. Plant Cell. 1999;11(3):323–333. doi:10.1105/tpc.11.3.323.

Binkert M, Kozma-Bognár L, Terecskei K, De Veylder L, Nagy F, Ulm R. UV-B-Responsive association of the Arabidopsis bZIP transcription factor ELONGATED HYPOCOTYL5 with target genes, including its own promoter. Plant Cell. 2014;26(10):4200–4213. doi:10.1105/tpc.114.130716.

Chen W, Chao G, Singh KB. The promoter of a H_2O_2-inducible, Arabidopsis glutathione S-transferase gene contains closely linked OBF- and OBP1-binding sites. Plant J. 1996;10(6):955–966. doi:10.1046/j.1365-313x.1996.10060955.x.

Dhatterwal P, Basu S, Mehrotra S, Mehrotra R. Genome wide analysis of W-box element in *Arabidopsis thaliana* reveals TGAC motif with genes down regulated by heat and salinity. Sci Rep. 2019;9(1):1681.doi:10.1038/s41598-019-38757-7.

Fernández-Calvo P, Chini A, Fernández-Barbero G, Chico JM, Gimenez-Ibanez S, Geerinck J, Eeckhout D, Schweizer F, Godoy M, Franco-Zorrilla JM, *et al.* The Arabidopsis bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. Plant Cell. 2011;23(2):701–715. doi:10.1105/tpc.110.080788.

Gat-Viks I, Sharan R, Shamir R. Scoring clustering solutions by their biological relevance. Bioinformatics. 2003;19(18):2381–2389. doi:10.1093/bioinformatics/btg330.

Guiltinan MJ, Marcotte WR, Quatrano RS. A plant leucine zipper protein that recognizes an abscisic acid response element. Science. 1990;250(4978):267–271. doi:10.1126/science.2145628.

Higo K, Ugawa Y, Iwamoto M, Higo H. PLACE: a database of plant cis-acting regulatory DNA elements. Nucleic Acids Res. 1999;27(1):297–300. doi: 10.1093/nar/27.1.297.

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. 2001;29(1):102–105. doi:10.1093/nar/29.1.102.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009a;37(1):1–13. doi:10.1093/nar/gkn923.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009b;4(1):44–57. doi:10.1038/nprot.2008.211.

Izawa T, Foster R, Chua NH. Plant bZIP protein DNA binding specificity. J Mol Biol. 1993;230(4):1131–1144. doi:10.1006/jmbi.1993.1230.

Jefferson RA, Kavanagh TA, Bevan MW. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO J. 1987;6(13):3901–3907. doi:10.1002/j.1460-2075.1987.tb02730.x.

Kang HG, Singh KB. Characterization of salicylic acid-responsive, Arabidopsis Dof domain proteins: overexpression of OBP3 leads to growth defects. Plant J. 2000;21(4):329–339. doi:10.1046/j.1365-313x.2000.00678.x.

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithms: analysis and implementation. IEEE Pami. 2002;24(7):881–892. doi:10.1109/TPAMI.2002.1017616.

Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. Gene expression Atlas at the European Bioinformatics Institute. Nucleic Acids Res. 2010;38(Database issue):D690–D698. doi:10.1093/nar/gkp936.

Kim TK, Maniatis T. The mechanism of transcriptional synergy of an in vitro assembled interferon-β enhanceosome. Mol Cell. 1997;1(1):119–129. doi:10.1016/s1097-2765(00)80013-1.

Kiran K, Ansari SA, Srivastava R, Lodhi N, Chaturvedi CP, Sawant SV, Tuli R. The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. Plant Physiol. 2006;142(1):364–376. doi:10.1104/pp.106.084319.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(Database issue): D1202–D1210. doi:10.1093/nar/gkr1090.

Le Hir R, Bellini C. The plant-specific Dof transcription factors family: new players involved in vascular system development and functioning in Arabidopsis. Front Plant Sci. 2013;4:164. Doi: 10.3389/fpls.2013.00164.

Li Q, Jia R, Dou W, Qi J, Qin X, Fu Y, He Y, Chen S. CsBZIP40, a BZIP transcription factor in sweet orange, plays a positive regulatory role in citrus bacterial canker response and tolerance. PLoS One. 2019;4(10):e0223498.doi:10.1371/journal.pone.0223498.

Lijavetzky D, Carbonero P, Vicente-Carbajosa J. Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. BMC Evol Biol. 2003;3:17.doi:10.1186/1471-2148-3-17.

Liu D, Shi S, Hao Z, Xiong W, Luo M. OsbZIP81, a homologue of Arabidopsis VIP1, may positively regulate JA levels by directly targetting the genes in JA signaling and metabolism pathway in rice. Int J Mol Sci. 2019;20(9):2360. doi:10.3390/ijms20092360.

Ma J, Li MY, Wang F, Tang J, Xiong AS. Genome-wide analysis of Dof family transcription factors and their responses to abiotic stresses in Chinese cabbage. BMC Genomics. 2015;16(1):33.doi: 10.1186/s12864-015-1242-9.

McDonald JH. Introduction. In: Handbook of Biological Statistics (3rd edition). Baltimore (MD): Sparky House Publishing; 2014.

Mehrotra R, Gupta G, Sethi R, Bhalothia P, Kumar N, Mehrotra S. Designer promoter: an artwork of cis engineering. Plant Mol Biol. 2011;75(6):527–536. doi:10.1007/s11103-011-9755-3.

Mehrotra R, Jain V, Shekhar C, Mehrotra S. Genome wide analysis of *Arabidopsis thaliana* reveals high frequency of AAAGN7CTTT motif. Meta Gene. 2014;2:606–615. doi:10.1016/j.mgene.2014.05.003.

Mehrotra R, Kiran K, Prakash Chaturvedi C, Anjum Ansari S, Lodhi N, Sawant S, Tuli R. Effect of copy number and spacing of the ACGT and GT cis elements on transient expression of minimal promoter in plants. J Genet. 2005;84(2):183–187. doi:10.1007/BF02715844.

Mehrotra R, Mehrotra S. Promoter activation by ACGT in response to salicylic and abscisic acids is differentially regulated by the spacing between two copies of the motif. J Plant Physiol. 2010;167(14): 1214–1218. doi:10.1016/j.jplph.2010.04.005.

Mehrotra R, Renganaath K, Kanodia H, Loake GJ, Mehrotra S. Towards combinatorial transcriptional engineering. Biotechnol Adv. 2017;35(3):390–405. doi:10.1016/j.biotechadv.2017.03.006.

Mehrotra R, Sethi S, Zutshi I, Bhalothia P, Mehrotra S. Patterns and evolution of ACGT repeat cis-element landscape across four plant genomes. BMC Genomics. 2013;14:203.doi:10.1186/1471-2164-14-203.

Mena M, Vicente-Carbajosa J, Schmidt RJ, Carbonero P. An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. Plant J. 1998;16(1):53–62. doi:10.1046/j.1365-313x.1998.00275.x.

Mishra AK, Agarwal S, Jain CK, Rani V. High GC content: critical parameter for predicting stress regulated miRNAs in *Arabidopsis thaliana*. Bioinformation. 2009;4(4):151–154. doi:10.6026/97320630004151.

Morin B, Nichols LA, Holland LJ. Flanking sequence composition differentially affects the binding and functional characteristics of glucocorticoid receptor homo- and heterodimers. Biochemistry. 2006;45(23):7299–7306. doi:10.1021/bi060314k.

Rojas-Gracia P, Roque E, Medina M, López-Martín MJ, Cañas LA, Beltrán JP, Gómez-Mena C. The DOF transcription factor Sldof10 regulates vascular tissue formation during ovary development in tomato. Front Plant Sci. 2019;10:216.doi:10.3389/fpls.2019.00216.

Sandelin A, Wasserman WW, Lenhard B. ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res. 2004;32(Web Server issue):W249–W252. doi: 10.1093/nar/gkh372.

Sell S, Hehl R. Functional dissection of a small anaerobically induced bZIP transcription factor from tomato. Eur J Biochem. 2004; 271(22):4534–4544. doi:10.1111/j.1432-1033.2004.04413.x.

Singh KB. Transcriptional regulation in plants: the importance of combinatorial control. Plant Physiol. 1998;118(4):1111–1120. doi: 10.1104/pp.118.4.1111.

Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–626. doi: 10.1038/nrg3207.

Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16–23. doi:10.1093/bioinformatics/16.1.16.

Teif VB, Rippe K. Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers. Phys Biol. 2011;8(4): 044001.doi:10.1088/1478-3975/8/4/044001.

Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ. A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. Proc Natl Acad Sci USA. 1997;94(14):7685–7690. doi: 10.1073/pnas.94.14.7685.

Wang H, Zhao S, Gao Y, Yang J. Characterization of Dof transcription factors and their responses to osmotic stress in poplar (*Populus trichocarpa*). PLoS One. 2017;12(1):e0170210.doi:10.1371/journal.-pone.0170210.

Washio K. Identification of Dof proteins with implication in the gibberellin-regulated expression of a peptidase gene following the germination of rice grains. Biochim Biophys Acta. 2001; 1520(1):54–62. doi:10.1016/s0167-4781(01)00251-2.

Weiste C, Pedrotti L, Selvanayagam J, Muralidhara P, Fröschel C, Novák O, Ljung K, Hanson J, Dröge-Laser W. The Arabidopsis bZIP11 transcription factor links low-energy signalling to auxin-mediated control of primary root growth. PLoS Genet. 2017;13(2): e1006607.doi:10.1371/journal.pgen.1006607.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 2003;20(9):1377–1419. doi:10.1093/molbev/msg140.

Yamamoto MP, Onodera Y, Touno SM, Takaiwa F. Synergism between RPBF Dof and RISBZ1 bZIP activators in the regulation of rice seed expression genes. Plant Physiol. 2006;141(4):1694–1707. doi:10.1104/pp.106.082826.

Yanagisawa S. Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. Plant J. 2000;21(3):281–288. doi:10.1046/j.1365-313x.2000.00685.x.

Yanagisawa S. The Dof family of plant transcription factors. Trends Plant Sci. 2002;7(12):555–560. doi:10.1016/s1360-1385(02)02362-2.

Yanagisawa S, Sheen J. Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression. Plant Cell. 1998;10(1):75–89. doi:10.1105/tpc.10.1.75.

Zhuo M, Sakuraba Y, Yanagisawa S. A Jasmonate-activated MYC2-Dof2.1-MYC2 transcriptional loop promotes leaf senescence in Arabidopsis. Plant Cell. 2020;32(1):242–262. doi:10.1105/tpc.19.00297.

Zong W, Tang N, Yang J, Peng L, Ma S, Xu Y, Li G, Xiong L. Feedback regulation of ABA signaling and biosynthesis by a bZIP transcription factor targets drought-resistance-related genes. Plant Physiol. 2016;171(4):2810–2825. doi:10.1104/pp.16.00469.

Communicating editor: B. Gregory