Roohollah Shamloo-Dashtpagerdi, Hooman Razi, Massumeh Aliakbari, Angelica Lindlöf, Mahdi Ebrahimi, Esmaeil Ebrahimie

A novel pairwise comparison method for in silico discovery of statistically significant cis-regulatory elements in eukaryotic promoter regions: application to Arabidopsis

Journal of Theoretical Biology, 2015; 364:364-376

A novel pairwise comparison method for in silico discovery of statistically significant cis-regulatory elements in eukaryotic promoter regions: application to Arabidopsis

[1]Roohollah Shamloo-Dashtpagerdi, [1]Hooman Razi[*], [1]Massumeh Aliakbari, [2]Angelica Lindlöf, [3]Mahdi Ebrahimi, [1&4]Esmaeil Ebrahimie[*]

[1]Department of Crop Production and Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran.[2]Systems Biology Research Centre, School of Life Sciences, University of Skövde, Skövde, Sweden. [3]Alumnus from Saarland University, Department of Informatics, Saarbrucken, Germany.[4]Discipline of Genetics, School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

*Co-corresponding authors

e-mail: razi@shirazu.ac.ir

e-mail: esmaeil.ebrahimie@adelaide.edu.au

Abstract

Cis regulatory elements (CREs), located within promoter regions, play a significant role in the blueprint for transcriptional regulation of genes. There is a growing interest to study the combinatorial nature of CREs including presence or absence of CREs, the number of occurrences of each CRE, as well as of their order and location relative to their target genes. Comparative promoter analysis has been shown to be a reliable strategy to test the significance of each component of promoter architecture. However, it remains unclear what level of difference in the number of occurrences of each CRE is of statistical significance in order to explain different expression patterns of two genes. In this study, we present a novel statistical approach for pairwise comparison of promoters of Arabidopsis genes in the context of number of occurrences of each CRE within the promoters. First, using the sample of 1000 Arabidopsis promoters, the results of the goodness of fit test and non-parametric analysis revealed that the number of occurrences of CREs in a promoter sequence is Poisson distributed. As a promoter sequence contained functional and non-functional CREs, we addressed the issue of the statistical distribution of functional CREs by analyzing the ChIP-seq datasets. The results showed that the number of occurrences of functional CREs over the genomic regions was determined as being Poisson distributed. In accordance with the obtained distribution of CREs occurrences, we suggested the Audic and Claverie (AC) test to compare two promoters based on the number of occurrences for the CREs. Superiority of the AC test over Chi-square (2 × 2) and Fisher's exact tests was also shown, as the AC test was able to detect a higher number of significant CREs. The two case studies on the Arabidopsis genes were performed in order to biologically verify the pairwise test for promoter comparison. Consequently, a number of CREs with significantly different occurrences was identified between the promoters. The results of the pairwise comparative analysis together with the expression data for the studied genes revealed the biological significance of the identified CREs.

Keywords: CREs occurrence; Motif enrichment, Transcriptional regulation

1. Introduction

The orchestrated spatial and temporal regulation of gene expression is a complex process that occurs at different checkpoints in the cell. Cis regulatory elements (CREs) are known as an important segment of the blueprint of transcriptional regulation (Qiu, 2003; Wittkopp and Kalay, 2011; Zheng et al., 2003). The interaction of transcription factors with CREs, located within promoter regions, leads to modulate transcription of target genes. Indeed, promoters contain functional DNA sequences which receive and integrate signals from multiple transcription factors by their modular and combinatorial nature (Vedel and Scotti, 2011; Werner, 2001).

Identification of CREs and their organization modules has opened a new vista in understanding gene expression and regulation (Deihimi et al., 2012; Hosseinpour et al., 2013). Recently, we developed a new approach for gene discovery irrespective from gene coding (BLAST), based on identifying distinct organization and combination of CREs (in view of order and distance) on promoter regions and identification of the genes with similar promoter architecture within whole genome (Hosseinpour et al., 2013). Recently, it has been demonstrated that CREs on the promoter regions of genes in wild wheat are more variable and frequent than the cultivated wheat which contributes in fast response and better understanding of environmental conditions for wild genotype (Babgohari et al., 2014). Due to the unique characteristic of transcription factors in binding to CREs on promoter regions of different genes, a small number of transcription factors are enough to regulate a considerable number of genes and play the central role in functional genomics (Mahdi et al., 2013; Mahdi et al., 2014). Interestingly, a small number of transcription factors and microRNAs, as the two main commanders of system biology, can regulate a genomic region involved in a particular phenomenon (hot spots) (Alisoltani et al., 2014). Currently, illustrating transcription factor based regulatory networks in different biological events is of great interest (Bakhtiarizadeh et al., 2014; Bakhtiarizadeh et al., 2013; Ebrahimie et al., 2014; Hosseinpour et al., 2012).

The growing availability of fully sequenced plant genomes and gene expression data together with substantial progress in bioinformatic tools have made it possible to computationally analyze the role of CREs in transcriptional regulation. A range of different computational models has been developed to identify over-represented CREs within promoter regions. One widely established model is to group genes based on their expression profiles and thereafter detect over-

represented CREs within each group (Elemento et al., 2007; Sinha and Tompa, 2003). However, the same motif may be found in the promoters of genes which fall into different groups. Another common approach, referred as phylogenetic footprinting, relies on the assumption that CREs are likely to be conserved across promoters of orthologous genes (Brohée et al., 2011; Kellis et al., 2003). The major disadvantages of phylogenetic footprinting are that species-specific regulatory elements will be missed and non-functional conserved motifs may be supposed as regulatory elements (Gao et al., 2013; Pennacchio and Rubin, 2001). In parallel with the computational methods, chromatin immunoprecipitation followed by sequencing (ChIP-seq) technology has experimentally enabled genome-wide discovery of cis-regulatory elements that act as transcription factor binding sites (Ladunga, 2010; Park, 2009). This high-throughput technology provides invaluable information to study CREs associated with gene regulation. For instance, ChIP-seq method has been reported for the identification of genome-wide targets of different transcription factors in Arabidopsis (Schiessl et al., 2014; Tao et al., 2012; Zhang et al., 2013).

The complex network of transcriptional regulation has led to the establishment of recent models that incorporate the combinatorial nature of CREs. These models take into account the presence or absence of CREs, the number of occurrences of each CRE, as well as of their order and location relative to their target genes (Mikkelsen and Thomashow, 2009; Pilpel et al., 2001; Segal and Widom, 2009; Zou et al., 2011). Comparative promoter analysis is a reliable strategy to test the significance of each component of promoter organization. Organizational similarities and differences of CREs among different promoters may contribute to the specific expression profiles of their corresponding genes. The evolutionary conservation of the relative order and location of CREs in promoters, referred to as a promoter module or framework, indicates their importance in gene regulation (Werner et al., 2003; Cohen et al., 2006). A number of studies have demonstrated the significant role of number of occurrences of each CRE within a promoter on the expression of target gene(Bussemaker et al., 2001; Foat et al., 2005; Mehrotra et al., 2011; Pilpel et al., 2001; Rushton et al., 2002). Although, it is still unclear what level of difference in the number of occurrences of each CRE between two promoters is of statistical significance to explain different expression patterns of two corresponding genes.

In this study, we present a novel statistical method for pairwise comparison of promoters of Arabidopsis genes in the context of number of occurrences of each CRE within the promoters.

This method is able to identify common or distinct CREs with significantly different number of occurrences between the two promoters. First, the statistical distribution of number of CREs within a given promoter is determined. We also exploit the ChIP-seq data of several experiments to determine the statistical distribution of number of occurrences of a given functional CRE across the target genomic regions identified by the corresponding regulatory protein. Then, a relevant pairwise test is employed to compare two promoters in terms of number of occurrences of their CREs. The contribution of the identified CREs on gene expression needs to be verified by studying the expression profiles of the corresponding genes. Using the proposed statistical approach, two case studies are performed. The first case study is to illustrate the ability of our approach in identifying significant CREs in comparison with the widely used approach which is analysis of promoters of groups of co-expressed genes to reveal over-represented CREs within each group. To do this, the promoters of the two groups of the co-expressed Arabidopsis genes differing in their responsiveness to light are analyzed. Then, Arabidopsis *AtHAM1* and *AtHAM2* genes are selected from these two groups of genes to perform the pairwise comparison on their promoter regions to discover significantly different CREs. The second case study is to compare between the promoters of two Arabidopsis key regulatory genes, *AtMYC2* and *AtMYB2*, which encode transcription factors involved in stress response and tolerance (Abe et al., 2003; Kazan and Manners, 2013). Thereafter, the result of pairwise comparative promoter analysis is combined with expression data of these genes in drought and heat stress conditions in order to explain the biological significance of the identified CREs.

2. Materials and methods

The workflow diagram which summarizes the various steps of the proposed method for the discovery of CREs with statistically significantly different number of occurrences between the two promoters is presented in Figure 1.

2.1. Arabidopsis Promoter sequences; collection and sampling

As putative promoter sequences, 1500bp upstream of all Arabidopsis genes (including 5'UTR) were downloaded from Ensemble Plants (plants.ensembl.org). Since the number of promoters was a finite number of values (from 1 to *n* and *n* being the total number of promoter sequences) with an equal probability of observation (1/*n*), we considered the number of promoters as having a Discrete Uniform (D. Uniform) distribution according to the following:

$$F(k; a, b) = ([k]-a+1)/(b-a+1)$$

Where *F* is the D. Uniform cumulative distribution function(CDF), *k* is any subset of promoters, *a*=1 and *b*=total number of promoter sequences (*n*). With these parameters and by the use of EasyFit software version 5.5 (http://www.mathwave.com), we generated 1000 random numbers based on the Mersenne Twister algorithm. This random number generator algorithm has the potential to generate very high quality pseudorandom numbers which is of choice for most statistical simulations (L'Ecuyer, 2012; Xiang and Benkrid, 2009). Using 1000 randomly generated numbers, similarly 1000 promoter sequences were sampled for subsequent analyses.

To confirm the accordance of sampling method from a biological aspect, all Arabidopsis promoter sequences were classified into functional categories according to the Pageman ontology tool (http://mapman.mpimp-golm.mpg.de/), which supports the use of MapMan, KEGG, MIPS, and GO ontologies (Usadel et al., 2006).The classification was based on the accession numbers of the genes for which promoters had been collected. The statistical distribution of sequences among functional categories was determined using the EasyFit software. The same procedure was performed on the sampled sequences (i.e., the subset of 1,000 promoter sequences). Finally, the consistency between the distribution of all and sampled sequences was checked by ranking all the fitted distributions for each instance.

2.2. Identification of CREs

An in-house database containing the previously identified plant CREs was built from a merge of motifs from the plantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/; (Lescot et al., 2002) database and motifs extracted from the literature. The forward and reverse strands of the promoter regions were searched for the input CREs by using an in-house developed Perl script. Only perfect matches to the motif were favored, i.e. the search did not have a scoring function. Additionally, only motifs with >4 IUPAC letters (www.iupac.org) were regarded.

2.3. ChIP-seq data analysis

The data of nine two-sample ChIP-seq experiments were retrieved from EBI (http://www.ebi.ac.uk/) (Table 1).

We used Galaxy web tool (https://usegalaxy.org/) (Goecks et al., 2010) to upload and analyze the data of each experiment. After quality control of the data, the control and chiped samples were mapped separately to Arabidopsis TAIR10 genome by Bowtie package with default parameters. SAM tool on Galaxy was used to exclude unmapped read. MACS algorithm (Zhang et al., 2008) with customized parameters (tag size=26, bandwidth=300bp, p-value cutoff≤1.00e-05 and mfold=20-32) was used to call peaks representing enriched binding sites, and afterward BED and FAST formatted files of the peaks were fetched from Galaxy. All the peaks were subjected to the Regulatory Sequence Analysis Tools (RSAT) Web server (http://rsat.ulb.ac.be/rsat/) (Thomas-Chollier et al., 2012)to discover statistically enriched CREs.

2.4. Statistical analyses of CREs

2.4.1. Goodness of fit test

The goodness of fit (GOF) test measures the compatibility of a random sample with a theoretical probability distribution function. In other words, these tests show how well the selected distribution fits the data(Quinn and Keough, 2002). There are three common GOF tests, namely Chi-square, Anderson-Darling (A-D) and Kolmogorov–Smirnov (K–S) (Grinstead and Snell, 1997; Quinn and Keough, 2002).

The K-S test is based on the empirical cumulative distribution function (ECDF) and when there is a large number of categories and the categories can be ordered in some way, the K-S test is preferred over the others (Quinn and Keough, 2002). The A-D procedure is a general test to compare the fit of an observed CDF to an expected CDF and achieves high statistical power for small samples (Saculinggan and Balase, 2013). The number of occurrences of each CRE in each promoter sequence is a finite number from 0 to *n* (0, 1, 2,…, *n*), whereby it is logical to assume that the distribution of CREs in a promoter is discrete. Based on this assumptions, the K–S and A-D tests were performed for seven main discrete distributions, namely Poisson, D.Uniform, Geometric, Logarithmic, Hyper-geometric, Binomial and Negative Binomial, using the EasyFit software with a significance level of α=0.05 in order to find the best fitted distribution for the number of occurrences of CREs in the promoters.

2.4.2. Rank-based tests

Using the GOF test statistics, EasyFit software ranks the fitted distributions from 1 (with minimum statistics) to *n* (with maximum statistics). A lower rank means a better fitness. In this study, each discrete distribution was assigned as an independent group with the ranked data. For non-normal distributions, rank-based methods might be used to compare groups(Quinn and Keough, 2002; Rumsey, 2011). The Kruskal–Wallis test (sometimes described as a "non-parametric ANOVA"), was performed for the statistical comparison of groups (discrete distributions) using the SAS software (version 9.0). This test is based on ranking the pooled data, determining the rank sums within each group and calculating the statistic that follows a chi-square distribution (Quinn and Keough, 2002). The Mann-Whitney U test was also performed for post hoc comparisons.

2.4.3. Comparative promoter analysis of CREs occurrences

Based on the distribution of number of CREs that was identified in the previous steps, an Audic and Claverie (AC) test was developed to carry out pairwise promoter comparisons. The AC test is a pairwise statistical test commonly used for the detection of differentially expressed genes (Audic and Claverie, 1997; Romualdi et al., 2001; Shamloo-Dashtpagerdi et al., 2013). Audic

and Claverie (1997) developed the following equation which involves the sampling size that is the total number of picked clones from a given cDNA library. This equation is used when one wishes to compare gene profiles that have been calculated from the random picking of different numbers of clones, N_1 and N_2. The mathematical problem is to establish probability for a given cDNA to be picked up x times when the sampling size is N_1 and Y times when the sampling size is N_2. This equation applies to the analysis of counts in experiments differing by the total number of clones (Audic and Claverie, 1997). In practice, the equation is used to analyze experiments performed on two different libraries using different sampling sizes.

$$P(y|x) = \binom{N_2}{N_1}^{y} \frac{(x+\ y)!}{x!\ y!\ (1 + N_2/N_1)^{(x+y+1)}}$$

We employed the AC test in accordance with our concept. The AC test gives the conditional probability of observing *x* number of a given CRE in promoter *A*, if the same CRE has been observed *y* times in promoter *B*. N_1 and N_2 are the total number of CREs in promoters *A* and *B*, respectively. The null hypothesis is that there is no difference in the number of specific CRE between promoters *A* and *B*. The frequently occurring common CREs, mainly TATA box, may largely affect the total number of CREs in each promoter and thereby may mask the importance of some other CREs. We applied the AC test on the two case studies. In each of the case studies, the AC test was performed with and without counting the TATA box motifs in order to test whether the deviation of sampling size caused by the most frequent motif had an impact on the results.

Pairwise comparisons of several CREs between two promoters were facilitated by using the AC test available at IDEG6 web tool (http://telethon.bio.unipd.it/bioinfo/IDEG6/) (Romualdi et al., 2003). The false discovery rate (FDR q-value) method (Benjamini and Hochberg, 1995) was used to adjust p-values derived from the AC test for multiple testing. The q-values were computed using QVALUE software (Storey and Tibshirani, 2003). The statistically differential CREs between two promoters were identified using q-value ≤ 0.01.

We compared the AC test with the Chi-square (2 × 2) and Fisher's exact tests (Bohm and Zech, 2010; Quinn and Keough, 2002; Rumsey, 2011), in order to verify which of these tests is more sensitive and able to detect more significant CREs.

2.5. Case study 1: Comparative promoter analysis of Arabidopsis *AtHAM1* and *AtHAM2* genes

To verify the biological relevance of the proposed pairwise promoter comparison method, a promoter analysis was done using the data generated by (Ouyang et al., 2011). They performed a ChIP-seq experiment along with microarray analysis to identify direct targets of *FHY3*, a key component in phytochrome A signaling and the circadian clock (Li et al., 2011), in darkness (D) and far-red (FR) light conditions in the Arabidopsis genome. Comparison between the ChIP-seq and microarray data indicated that *FHY3* quickly regulates the expression of 197 and 86 genes in D and FR, respectively (Ouyang et al., 2011). Here, we had the two groups of the co-expressed genes; the one included the up-regulated genes expressed in both D and FR conditions and the other group comprised the up-regulated genes expressed only in D condition. Based on the ChIP-seq data, the genes with at least one *FHY3* binding site at their promoters or 5ʾUTR regions were selected from each group. The putative promoter sequences containing 1500bp upstream (including 5ʾUTR) of the genes of each group were obtained from Ensemble Plants (http://plants.ensembl.org/index.html). Each of the promoter sequences was subjected to discovery of CREs using our in house database which was previously described. In order to discover the differentially enriched CREs between the two groups of genes, the group of genes expressed only in D condition was considered as the background and the significant enriched CREs (relative to the background sequences), which is expected to contain some light-responsive CREs, were identified using Fisher's exact test with FDR-adjusted p-values (q-value ≤ 0.05). After that, one gene from each group was chosen to apply the pairwise methods for comparison of their promoters. The promoters of *AtHAM1* and *AtHAM2* genes belonging to the same transcription factor family but with different expressions in response to light were subjected to the pairwise comparison in order to identify the differentially significant CREs. Finally, we investigated the ability of the pairwise test, in comparison with the method of grouping of co-expressed genes, to identify significant CREs associated with the expression profiles of the two examined genes.

2.6. Case study 2: Comparative promoter analysis of Arabidopsis *AtMYC2* and *AtMYB2* genes

The pairwise tests were also applied for a comparative promoter analysis of the Arabidopsis *AtMYC2* and *AtMYB2* genes. The microarray data for *AtMYC2* and *AtMYB2* genes under drought and heat stress conditions were obtained (Kilian et al., 2007) and collected using the "The Bio-Array Resource for Plant Biology" (BAR) (http://bar.utoronto.ca). All the experimental conditions were similar in the two assays. The stress treatments were initiated 18 days after sowing. Plant samples were taken in two biological replicates with the same time points: 0 min, 30 min, 1 h, 3 h, 6 h, 12 h and 24 h after the onset of stress treatment (Kilian et al., 2007). The co-expression profiles for the two genes in drought and heat stresses were depicted using Microsoft Excel 2013. Pearson correlation coefficient (α=0.05) was calculated for the expression profiles of *AtMYC2* and *AtMYB2* genes in each condition using SAS 9.0 software.

3. Results

3.1. Statistical and biological validation of promoter sampling

The promoters from all Arabidopsis genes were obtained, resulting in the total number of 27415 promoter sequences. Thereafter, 1,000 promoters were randomly sampled from the total number of promoters by using random numbers based on the Mersenne Twister algorithm in the EasyFit software. According to the CDF, the total and sampled promoters followed a D. Uniform distribution in accordance with each other, which thereby confirm the reliability of the sampling procedure (Figure 2).

The population of all Arabidopsis promoters (27415) and the subset of 1000 Arabidopsis promoters were classified into functional categories based on the accession numbers of the relevant genes using the Pageman ontology tool (Usadel et al., 2006). Of the 27415 and 1000 promoter sequences, 36.67 and37.69% fell into "Not assigned" and "Not assigned-unknown" categories, respectively. The remaining sequences of both sets of promoters shared 30 common categories with a relatively similar percentage of sequences assigned to each category (Figure 3).In both all and sampled promoters, the categories namely Protein (21.65 vs 20.38%), RNA (16.71 vs 15.32%), Signaling (7.57 vs 7.58%), Stress (6.65 vs 7.74%) and Transport (5.86 vs 7.1%) contained the highest percentage of the sequences. These results revealed that the sample of promoter sequences was biologically consistent with the population of all promoter sequences. Moreover, a goodness of fit (GOF) test was applied on the distribution of sequences among functional categories for each of all and sampled promoter sequences, to test if the same statistical distribution is fitted to both sets of promoters. The Kolmogorov–Smirnov GOF tested for seven main discrete distributions (Poisson, D. Uniform, Geometric, Logarithmic, Hyper-geometric, Binomial and Negative Binomial; using the EasyFit software with α=0.05) showed that the logarithmic distribution was ranked first for both all and sampled promoters (Table 2).

This clearly indicated that the obtained sample of promoter sequences is an adequate representative of Arabidopsis promoters from a biological perspective. Since the sampling method was verified to be statistically and biologically sound, the sample population was judged to be sufficient to reflect the total population.

3.2. CREs identification

The CREs and their number of occurrences were found in each of the 1000 Arabidopsis promoters. CREs within 1500bp up-stream of 1000 random sampled Arabidopsis genes were identified. 35 different types of CREs, in average, were identified within each promoter. The average number of CREs was 108 within each promoter. The range of occurrences of CREs was 1 to 58 in the promoters.

A summary of analysis of the ChIP-seqexperiments aimed at identifying motifs acting as transcription factor binding sites has been shown in table 3. Totally 114 functional CREs were found through analysis of the ChIP-seq datasets. The number of occurrences of each CREs across the Arabidopsis genome was determined in each experiment.

3.3. Determining the best fitted statistical distribution for the CREs occurrences

In order to determine the relevant statistical distributions for CREs numbers within each promoter, the goodness of fit test was performed for seven main discrete distributions including Poisson, D. Uniform, geometric, logarithmic, hyper-geometric, binomial and negative binomial using EasyFit Software (α=0.05). For each promoter, the distributions fitted to CREs were ranked in which the most relevant distribution was ranked first. The GOF test was also performed to determine the statistical distribution of the occurrences of each functional CREs identified by each of the ChIP-seq experiment and then the fitted distributions were ranked based on GOF statistics. The ranking of the fitted distributions for both the datasets derived from the sampled promoters and the ChIP-seq experiments are shown in table 4.

The CREs had a Poisson distribution, as the best-fitted distribution, in 765 out of 1000 the promoters. The best-fitted distributions for the CREs of 174 and 55 promoters were D. Uniform and geometric, respectively. The results also showed that Poisson distribution was among the distributions fitted to CREs in all of the promoters. We found that the Poisson distribution was the best fitted distribution for the occurrences of 91% of the functional CREs (104 out of 114) across the genomic regions.

To reinforce the results of the GOF tests and to identify a statistically significant distribution of the CREs, the non-parametric ANOVA Kruskal–Wallis test was carried out on the ranks of the fitted distributions. In both the datasets derived from the 1000 promoters sample and the ChIP-

seq experiments, the results of the Kruskal–Wallis test showed a strong statistical significant difference among the distributions (Table 5). Moreover, since the Poisson distribution was the relevant distribution fitted to the CREs in most of the promoters, we applied the Mann-Whitney U test for a pairwise comparison between this distribution and the remaining ones (i.e., Poisson versus D. Uniform, Poisson versus Geometric and so on) (Table 6). The Mann-Whitney U test revealed significant difference between the Poisson distribution and the other distributions for both the datasets.

As a result, the Poisson distribution is assumed to adequately represent the statistical distribution associated with CREs occurrences within a given Arabidopsis promoter and also with the occurrences of the functional CREs over the corresponding genomic regions. This finding led us to use a Poisson-based pairwise test (the Audic and Claverie test) to compare CREs between two promoters in terms of their number of occurrences. The ability of Audic and Claverie test to find significantly different CREs between the two promoters is presented in the form of two case studies.

3.4. Case study 1: Comparative promoter analysis of *AtHAM1* and *AtHAM2*

The 68 up-regulated genes in FR conditions and the 57 up-regulated genes in D conditions formed the two groups of the co-expressed genes subjected to promoter analysis in order to identify differentially enriched motifs. Based on our database of CREs, we found 60 distinct CREs which had more occurrences in the promoters of the group of FR co-expressed genes. Of those, irrespective of TATA box motif, four CREs including C_2C_2-DOF (DNA binding with One Finger) binding site, MYB binding site, MYB15 and I-box were statistically significantly enriched in the FR co-expressed genes relative to the other group of genes. Some of the significant CREs may be associated with the differential expression of the two groups of genes in response to light. DOF proteins are plant specific transcription factors involved in seed development as well as signaling and response to light and phytohormone (Mahdi et al., 2014). MYB binding sites are abundant in promoters of stress responsive genes, however there are some evidence that they work together with other proteins to confer light responsiveness (Babgohari et al., 2014). MYB15 is R2R3 type MYB transcription factor involved in cold regulation of number of genes (Mahdi et al., 2013). I-box is a CRE available at the light and circadian clock responsive plant promoters (Agarwal et al., 2006; Borello et al., 1993). It is noteworthy that the

differentially enriched CREs were not detected in some of the promoter sequences. Moreover, there were some CREs that were not identified as the differentially enriched CREs but they had significantly different abundance in promoters of two genes of interest, each of which assigned to its respective group. These disadvantages of promoter analysis of co-regulated groups led us to use the AC pairwise promoter comparison between two genes selected from the studied groups of the co-expressed genes.

Two members of GRAS transcription factor family, *AtHAM1* and *AtHAM2* were selected from the groups of FR and D up-regulated genes, respectively for further analysis. *AtHAM1*and *AtHAM2* function in different processes such as meristem maintenance, shoot and root indeterminacy, shoot branching, chlorophyll biosynthesis and root growth (Engstrom, 2012; Engstrom et al., 2011; Schulze et al., 2010; Stuurman et al., 2002). However, the microarray data showed that their expression pattern is different in FR and D conditions(Ouyang et al., 2011). *AtHAM1* up-regulated 1.6 fold in both FR and D conditions whereas *AtHAM2* up-regulated 1.4 fold only in D conditions(Ouyang et al., 2011). Only one of the differentially enriched CREs (MYB binding site), identified by promoter analysis of the two groups of the co-expressed genes, was present in the promoters of *AtHAM1* and *AtHAM2* genes and there was no copy of the other enriched CREs neither in *AtHAM1* nor in *AtHAM2* promoters.

Based on our database of CREs, the total number of CREs in the *AtHAM1* and *AtHAM2* promoter sequences were 112 and 146, respectively, when TATA box motifs were accounted. By excluding TATA box motifs, the total number of CREs in the two promoters reduced to 76 CREs. The AC pairwise promoter comparison was performed to clarify how the number of occurrences of CREs may contribute to the different expression patterns of these genes in response to light. This test was done under two circumstances, with and without counting the occurrences of TATA box in the promoters, in order to check whether the inclusion of the highly frequent motifs such as TATA box affects the results of AC test. The AC pairwise test detected a number of significantly differential CREs between *AtHAM1* and *AtHAM2* promoters. When the TATA box motifs were taken into account, the AC test identified nine CREs which had statistically different occurrences between the two promoters. The results revealed that the binding site of PEND protein, a DNA-binding protein in the inner envelope membrane of the developing chloroplast (Sato et al., 1998; Terasawa and Sato, 2005), the 5’UTR Py-rich element conferring high transcription levels (Nejad et al., 2013) and heat shock element (HSE) had higher

occurrences in the *AtHAM1* promoter. Furthermore, the three significant CREs (BoxI, GT1 and G-Box) which were more abundant in the promoter of FR responsive gene (*AtHAM1*) are involved in light responsiveness. On the other hand, TATA box, Skn-1, CAAT-box had higher number of occurrences in the *AtHAM2* promoter (Table 7). By excluding the TATA box motifs, the results slightly changed as TATA box and G-Box were no longer significant CREs, while the CREs called TC-rich repeat and AT-rich became significant and added to the same other significant CREs (Table 7).

The comparison between the AC, Chi-square (2 × 2) and Fisher's exact tests, with and without the inclusion of TATA box motifs, showed that the AC test was superior to the other two tests, as it was able to detect a higher number of significant CREs with lower q-values (Table 7).

3.5. Case study 2: Comparative promoter analysis of *AtMYC2* and *AtMYB2* genes

The pairwise tests were also used for a comparative analysis of the promoters of Arabidopsis *AtMYC2* and *AtMYB2* genes. The expression profiles of the two genes under drought and heat stress conditions were obtained (Figure 4) and a Pearson correlation coefficient was worked out between the expressions of the two genes in each conditions.

There was a positive correlation (0.7686; $\alpha = 0.05$) between the expression levels of the two genes in the drought stress conditions and a negative correlation (-0.6332; $\alpha = 0.05$) in the heat stress conditions. It can be inferred from the expression profiles (Figure 4) and the correlation coefficients that: (1) the expression levels of *AtMYC2* was generally higher than *AtMYB2* in both drought and heat stress conditions and (2) unlike the drought condition, the expression trends of the two genes were opposite to each other in the heat stress condition. In the early stages of heat stress, *AtMYC2* expression decreased while the expression levels of *AtMYB2*increased. In the late stages of heat stress, *AtMYC2* and *AtMYB2* were expressed in an inverse manner.

We used the AC pairwise statistical method to compare the promoter sequences of *AtMYC2 and AtMYB2* based on the number of occurrences for the CREs. There were 31 and 28 different types of CREs in *AtMYC2* and *AtMYB2* promoter sequences, respectively, of which 19 were common between the two genes. Regarding the number of occurrences of each CRE, the total number of CREs (including the TATA box) in the *AtMYC2* and *AtMYB2* promoter sequences were 181 and

167, respectively. By excluding the TATA box motifs, the total number of CREs reduced to 91 and 78 in the promoters of *AtMYC2* and *AtMYB2*, respectively.

According to the AC test, nine CREs were significantly different (q-value ≤ 0.01) between the two promoter sequences (Table 8). In this case study, there was no difference in the results obtained from two states of with and without inclusion of TATA box motifs. The five significant CREs involved in several biological processes namely ABRE, G-box, MBS, TA-rich region and unnamed-4 were more abundant in the promoter of *AtMYC2* promoter. On the other hand, Box I, ERE, TATA box and CAAT-box CREs had higher number of occurrences in the *AtMYB2* promoter. In addition, a comparison between the AC test and Chi-square (2 × 2) and Fisher's exact tests showed that the AC test was superior to the other two tests, as it was able to detect more significant CREs (Table 8).

4. Discussion

4.1. A pairwise statistical method for comparative promoter analysis

Comparative promoter analysis is a promising strategy for the detection of common and different CREs within promoters of the genes with similar or different expression profiles (Cohen et al., 2006; Conceição et al., 2010; Deihimi et al., 2013; Dieterich et al., 2005; Gao et al., 2013; Gómez-Porras et al., 2007; Maruyama et al., 2012; Moghadam et al., 2013; Ramezani et al., 2013; Zamani Babgohari et al., 2013). While there are several studies indicating that number of occurrences of CREs has a significant effect on the expression pattern of the target gene (Bussemaker et al., 2001; Foat et al., 2005; Mehrotra et al., 2011; Pilpel et al., 2001; Rushton et al., 2002), the lack of powerful and reliable statistical method to compare CREs occurrences between promoters is a major drawback in this area of computational biology.

The development of statistical inference requires assumptions about the probability distribution of a data set. Knowledge about the distribution of data is essential to select the appropriate statistical method (Bohm and Zech, 2010; Rumsey, 2011).Using the sample of 1000 Arabidopsis promoters, the results of the goodness of fit test and non-parametric analysis revealed that the number of occurrences of CREs in a promoter sequence is Poisson distributed. As a promoter sequence contained functional and non-functional CREs, we addressed the issue of the statistical distribution of functional CREs by analyzing the ChIP-seq datasets. The results showed that the number of occurrences of functional CREs over the genomic regions was determined as being Poisson distributed. Therefore, a Poisson-based pairwise test could be proposed to compare CREs within two promoters in terms of their number of occurrences.

The proposed test is based on the Audic and Claverie test. Audic and Claverie (1997) established a rigorous test for pairwise comparison of transcript profiles, based on Poisson distribution assumption. The pairwise promoter comparison aimed to identify the CREs which may contribute to the different expression profiles of two genes. The results obtained in this study indicated that the Audic and Claverie test is more powerful for identification of CREs with differential number of occurrences in comparison with the Chi-square and Fisher’s exact tests, because it was able to detect more significant differential CREs with lower q-values.

It should be noted that the ability of AC test to identify differentially significant CREs is affected by the size and property of known CREs within a promoter. The total number of CREs

in a promoter may largely be overestimated by the presence of highly frequent short CREs such as TATA box. In addition, some CREs exhibit positional preferences relative to the transcription start site (Hartmann et al., 2005). Although in our case studies the exclusion of TATA box motifs, as the most frequent one, did not make a significant change in the results of the AC tests, we suggest taking these issues into account in order to theoretically obtain more reliable and meaningful results from AC test.

4.2. Biological verification of the significant CREs identified by the pairwise test

4.2.1. Comparative analysis of *AtHAM1* and *AtHAM2* promoters

The expression of *AtHAM1* and *AtHAM2* increased in darkness, whereas only *AtHAM1*responded to light (Ouyang et al., 2011). This indicated that the expressions of these genes in response to light may be controlled by different regulatory elements. *AtHAM1* and *AtHAM2* had been identified as the direct targets of *FHY3*, a key component of phytochromeA signaling and the circadian clock (Ouyang et al, 2011). *AtHAM1*, but not *AtHAM2*, had shown cycling expression circadian conditions supposed to be associated with *FHY3* binding site within its promoter (Li et al., 2011). The identification of *FHY3* binding sites in the promoter of *AtHAM2* implied that there may be some other CREs contributing to light responsiveness of *AtHAM1*.

The pairwise comparison test between the promoters of *AtHAM1* and *AtHAM2* detected the two light responsive CREs (GT-1 and Box-I) which were statistically over-represented in the *AtHAM1* promoter, whilst these CREs were not highlighted by promoter analysis of the two groups of genes (containing *AtHAM1* and *AtHAM2*) with the differential expression in response to light. On the other hand, most of the differentially enriched CREs between the two groups of the co-expressed genes were not present in the promoters of *AtHAM1* and *AtHAM2*. It indicated the necessity of pairwise promoter comparison between two genes of interest for more accurate evaluation. GT-1 and Box-I motifs are essential for light-controlled transcriptional activity (López-Juez, 2007). The presence of different light responsive elements in the *AtHAM1* promoter suggested that a combination of different cis-acting sequences, as light responsive units (LRUs), may be required to confer proper responsiveness to light (Jiao et al., 2007). GT-1 was present only in the *AtHAM1* promoter. The GT-1 element is a binding site of GT-1 transcriptional

activator and is sufficient for light induction (Kaplan-Levy et al., 2012). Interestingly, it has been reported that the GT-1 element participates in phytochrome A signaling and circadian rhythm under light condition (Kaplan-Levy et al., 2012; Zhou, 1999).

The other identified light responsive element, BoxI, is the feature of photoperiod-responsive genes (Mongkolsiriwatana et al., 2009). Promoter analysis of photoperiod-responsive genes revealed that a combination of light responsive elements such as BoxI with CREs involved in other biological processes formed a coordinated gene regulation in response to light (Mongkolsiriwatana et al., 2009).

The 5'-UTR Py-rich stretch- element was another significant CREs which had more occurrences in the *AtHAM1* promoter. This CRE has a fundamental role in high transcription levels of cell cycle genes (Nejad et al., 2013) and there is no report about its possible role in response to light. It may contribute to higher transcription level of *AtHAM1* relative to *AtHAM2* in darkness.

4.2.2. Comparative analysis of *AtMYC2* and *AtMYB2* promoters

Response to drought

The method was applied to the promoter sequences for the Arabidopsis *AtMYC2* and *AtMYB2* genes. A conserved ABA-responsive *cis*-regulating element named ABRE (ABA responsive element; PyACGTGGC) was found in the promoter regions of *AtMYC2* and *AtMYB2*. This supports the previous reports that *AtMYC2* and *AtMYB2* proteins function as transcriptional activators of ABA-inducible genes under drought stress(Abe et al., 2003). Both genes may response to ABA via this element, which may cause coordinated increase in their expression levels; as evidenced by the positive correlation between the expression levels of these genes reported in this study. On the other hand, based on the result of the AC test, there were significantly more ABRE elements in the *AtMYC2* promoter which resulted in a higher expression of *AtMYC2* than *AtMYB2* in the drought condition. In addition, among significant different CREs, there were three MYB Binding Site (MBS) elements in the promoter of *AtMYC2*, suggesting that the expression of *AtMYB2* in drought condition and under control of ABA probably enhance the expression of *AtMYC2*. In fact, *AtMYB2* may be a positive regulator of *AtMYC2* expression.

Heat response

A number of significantly different CREs between the promoter sequences of the *AtMYC2* and *AtMYB2* genes were identified which may be associated with the differential expression profiles of these gens under heat stress. These CREs were jasmonic acid (JA) responsive elements (G-box and TA-rich region) and Ethylene-Responsive Element (ERE). The first two CREs were solely present in the *AtMYC2* promoter. In contrast, ERE motif was only present in the *AtMYB2* promoter. The G-box is a CRE found in a broad range of plant promoters, which is responsible for the light-response(Xu and Johnson, 2001; Yamaguchi-Shinozaki and Shinozaki, 2005). The TA-rich region is known, as an enhancer, to increase the expression of target gene (Cuming et al., 2007). As discussed by Xu and Timco (2004), both the G-box and TA-rich region are required for Methyl Jasmonate (MeJA)-responsiveness in *Nicotiana tobacco*. The ERE element is the binding site of a family of Ethylene Response-Element-Binding-Proteins (EREBPs) and is present in promoters of various ethylene inducible genes and mediates ethylene response(Benavente and Alonso, 2006; Rawat et al., 2005).

In agreement with the identified CREs in this study, Boter et al, (2004) demonstrated that in dicotyledonous plants, *AtMYC2* possesses an important function in regulating the expression of different JA-dependent genes. *AtMYC2*is also reported as a transcription factor that functions in JA-ethylene defense responses; however, there is no evidence about the role(s) of *AtMYB2* in JA response system. Interestingly, JA and ethylene (ET) hormones can either cooperate or act as antagonists in regulation of different stress responses(Benavente and Alonso, 2006; Clarke et al., 2009). Response to ozone stress is an example of antagonistic interaction between JA and ET, where JA protects tissues from stress while ET enhances ozone-induced cell death (Tamaoki et al., 2003).

There are several signaling pathways involved in the plant response to heat stress. In addition, phytohormones, such as ABA, JA, ET and salicylic acid (SA), have been also implicated to play role(s) in heat stress signaling in different plants (Kotak et al., 2007). Clarke and colleagues (2009) demonstrated that JA acts in concert with SA to confer basal thermo tolerance in Arabidopsis. Moreover, there are some evidence that ethylene signaling pathways are involved in thermo-tolerance(Kotak et al., 2007). From the results of our expression and promoter analyses and previous knowledge about plant heat stress responses, it can be deduced that *AtMYC2* and

AtMYB2 genes are a part of the antagonistic response to heat, as we delineated a negative correlation between *AtMYC2* and *AtMYB2* expression under heat stress. One can suppose that in the early stages of heat stress and following ethylene accumulation, the *AtMYB2*responses to the hormone through the ERE elements. Increase in expression of *AtMYB2* can elevate the *AtMYC2* expression level via binding of *AtMYB2* to MBSs within *AtMYC2*promoter in the later stress stages. Furthermore, accumulation of JA in response to heat stress results in further increase in *AtMYC2* expression through G-box and TA-rich region within the promoter. Finally, antagonistic interaction between JA and ET and subsequent reduction in ET level alter *AtMYB2* expression trend.

5. Conclusion

Integrating comparative statistical based analysis of CREs with expression data can open a new vista in genome wide functional analysis. The probability distributions are very important to select data analysis method as the type of statistical analysis relies on the distribution of data set. By assigning the relevant distribution of CREs occurrences in promoter sequences as Poisson distribution, we established a powerful statistical approach for comparative promoter analysis in this study. It seems that Audic and Claverie test is appropriate to apply for this issue, since it is based on Poisson distribution.

There was a meaningful relationship between the results of in silico promoter analysis and the expression data in both of the case studies. Interactions of regulatory proteins with variable numbers of specific CRE on different promoters may constitute an important part of gene regulation mechanisms. The results of this study provide the required information for further experimental research such as genetic manipulation of given promoters and dissection of signaling pathways in an eukaryotic organism leading to improve our knowledge about the molecular mechanisms involved in responses to internal and external stimuli.

Acknowledgments

We would like to thank the staff of Shiraz University and The University of Adelaide for their valuable helps.

References

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., Yamaguchi-Shinozaki, K., 2003. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. The Plant Cell Online 15, 63-78.

Agarwal, M., Hao, Y., Kapoor, A., Dong, C.-H., Fujii, H., Zheng, X., Zhu, J.-K., 2006. A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance. Journal of Biological Chemistry 281, 37636-37645.

Alisoltani, A., Fallahi, H., Ebrahimi, M., Ebrahimi, M., Ebrahimie, E., 2014. Prediction of Potential Cancer-Risk Regions Based on Transcriptome Data: Towards a Comprehensive View. PLoS ONE 9, e96320, doi:10.1371/journal.pone.0096320.

Audic, S., Claverie, J.-M., 1997. The significance of digital gene expression profiles. Genome research 7, 986-995.

Babgohari, M. Z., Ebrahimie, E., Niazi, A., 2014. In silico analysis of high affinity potassium transporter (HKT) isoforms in different plants. Aquatic Biosystems 10, 9.

Bakhtiarizadeh, M., Moradi-Shahrbabak, M., Ebrahimie, E., 2014. Transcriptional regulatory network analysis of the over-expressed genes in adipose tissue. Genes & Genomics 36, 105-117, doi:10.1007/s13258-013-0145-x.

Bakhtiarizadeh, M. R., Moradi-Shahrbabak, M., Ebrahimie, E., 2013. Underlying functional genomics of fat deposition in adipose tissue. Gene 521, 122-128, doi:http://dx.doi.org/10.1016/j.gene.2013.03.045.

Benavente, L. M., Alonso, J. M., 2006. Molecular mechanisms of ethylene signaling in Arabidopsis. Molecular Biosystems 2, 165-173.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289-300.

Bernal, M., Casero, D., Singh, V., Wilson, G. T., Grande, A., Yang, H., Dodani, S. C., Pellegrini, M., Huijser, P., Connolly, E. L., 2012. Transcriptome sequencing identifies SPL7-regulated copper acquisition genes FRO4/FRO5 and the copper dependence of iron homeostasis in Arabidopsis. The Plant Cell Online 24, 738-761.

Bohm, G., Zech, G., 2010. Introduction to statistics and data analysis for physicists. DESY.

Borello, U., Ceccarelli, E., Giuliano, G., 1993. Constitutive, light-responsive and circadian clock-responsive factors compete for the different I box elements in plant light-regulated promoters. The Plant Journal 4, 611-619.

Brohée, S., Janky, R. s., Abdel-Sater, F., Vanderstocken, G., André, B., van Helden, J., 2011. Unraveling networks of co-regulated genes on the sole basis of genome sequences. Nucleic acids research 39, 6340-6358.

Bussemaker, H. J., Li, H., Siggia, E. D., 2001. Regulatory element detection using correlation with expression. Nature genetics 27, 167-174.

Clarke, S. M., Cristescu, S. M., Miersch, O., Harren, F. J., Wasternack, C., Mur, L. A., 2009. Jasmonates act with salicylic acid to confer basal thermotolerance in Arabidopsis thaliana. New Phytologist 182, 175-187.

Cohen, C. D., Klingenhoff, A., Boucherot, A., Nitsche, A., Henger, A., Brunner, B., Schmid, H., Merkle, M., Saleem, M. A., Koller, K.-P., 2006. Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. Proceedings of the National Academy of Sciences 103, 5682-5687.

Conceição, N., Cox, C., Simões, B., Viegas, M., Cancela, M., 2010. Comparative promoter analysis and its application to the identification of candidate regulatory factors of cartilage-expressed genes. Journal of Applied Ichthyology 26, 245-250.

Cuming, A. C., Cho, S. H., Kamisugi, Y., Graham, H., Quatrano, R. S., 2007. Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, Physcomitrella patens. New Phytologist 176, 275-287.

Deihimi, T., Niazi, A., Ebrahimie, E., 2013. Identification and expression analysis of TLPs as candidate genes promoting the responses to both biotic and abiotic stresses in wheat. Plant OMICS: Journal of Plant Molecular Biology & Omics 6.

Deihimi, T., Niazi, A., Ebrahimi, M., Kajbaf, K., Fanaee, S., Bakhtiarizadeh, M. R., Ebrahimie, E., 2012. Finding the undiscovered roles of genes: an approach using mutual ranking of coexpressed genes and promoter architecture-case study: dual roles of thaumatin like proteins in biotic and abiotic stresses. SpringerPlus 1, 1-10.

Dieterich, C., Grossmann, S., Tanzer, A., Röpcke, S., Arndt, P. F., Stadler, P. F., Vingron, M., 2005. Comparative promoter region analysis powered by CORG. BMC genomics 6, 24.

Ebrahimie, M., Esmaeili, F., Cheraghi, S., Houshmand, F., Shabani, L., Ebrahimie, E., 2014. Efficient and Simple Production of Insulin-Producing Cells from Embryonal Carcinoma Stem Cells Using Mouse Neonate Pancreas Extract, As a Natural Inducer. PLoS ONE 9, e90885, doi:10.1371/journal.pone.0090885.

Elemento, O., Slonim, N., Tavazoie, S., 2007. A universal framework for regulatory element discovery across all genomes and data types. Molecular cell 28, 337-350.

Engstrom, E. M., 2012. HAM proteins promote organ indeterminacy But how? Plant signaling & behavior 7, 227-234.

Engstrom, E. M., Andersen, C. M., Gumulak-Smith, J., Hu, J., Orlova, E., Sozzani, R., Bowman, J. L., 2011. Arabidopsis homologs of the petunia hairy meristem gene are required for maintenance of shoot and root indeterminacy. Plant physiology 155, 735-750.

Fan, M., Bai, M.-Y., Kim, J.-G., Wang, T., Oh, E., Chen, L., Park, C. H., Son, S.-H., Kim, S.-K., Mudgett, M. B., 2014. The bHLH Transcription Factor HBI1 Mediates the Trade-Off between Growth and Pathogen-Associated Molecular Pattern–Triggered Immunity in Arabidopsis. The Plant Cell Online 26, 828-841.

Foat, B. C., Houshmandi, S. S., Olivas, W. M., Bussemaker, H. J., 2005. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proceedings of the National Academy of Sciences of the United States of America 102, 17675-17680.

Gao, Z., Zhao, R., Ruan, J., 2013. A genome-wide cis-regulatory element discovery method based on promoter sequences and gene co-expression networks. BMC genomics 14, S4.

Goecks, J., Nekrutenko, A., Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11, R86.

Gómez-Porras, J. L., Riaño-Pachón, D. M., Dreyer, I., Mayer, J. E., Mueller-Roeber, B., 2007. Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. BMC genomics 8, 260.

Grinstead, C. C. M., Snell, J. L., 1997. Introduction to probability. American Mathematical Soc.

Hartmann, U., Sagasser, M., Mehrtens, F., Stracke, R., Weisshaar, B., 2005. Differential combinatorial interactions of cis-acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes. Plant molecular biology 57, 155-171.

Hosseinpour, B., Bakhtiarizadeh, M. R., Khosravi, P., Ebrahimie, E., 2013. Predicting distinct organization of transcription factor binding sites on the promoter regions: a new genome-based approach to expand human embryonic stem cell regulatory network. Gene 531, 212-219.

Hosseinpour, B., HajiHoseini, V., Kashfi, R., Ebrahimie, E., Hemmatzadeh, F., 2012. Protein Interaction Network of <italic>Arabidopsis thaliana</italic> Female Gametophyte Development Identifies Novel Proteins and Relations. PLoS ONE 7, e49931, doi:10.1371/journal.pone.0049931.

Jensen, M. K., Lindemose, S., Masi, F. d., Reimer, J. J., Nielsen, M., Perera, V., Workman, C. T., Turck, F., Grant, M. R., Mundy, J., 2013. ATAF1 transcription factor directly regulates abscisic acid biosynthetic gene< i> NCED3</i> in< i> Arabidopsis thaliana</i>. FEBS open bio 3, 321-327.

Jiao, Y., Lau, O. S., Deng, X. W., 2007. Light-regulated transcriptional networks in higher plants. Nature Reviews Genetics 8, 217-230.

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

Kazan, K., Manners, J. M., 2013. MYC2: the master in action. Molecular plant 6, 686-703.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. S., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241-254.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., Harter, K., 2007. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 50, 347-363.

Kotak, S., Larkindale, J., Lee, U., von Koskull-Döring, P., Vierling, E., Scharf, K.-D., 2007. Complexity of the heat stress response in plants. Current opinion in plant biology 10, 310-316.

L'Ecuyer, P., 2012. Random Number Generation. In: Gentle, J. E., et al., Eds.), Handbook of Computational Statistics. Springer Berlin Heidelberg, pp. 35-71.

Ladunga, I., 2010. Computational Biology of Transcription Factor Binding. Springer.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., Rombauts, S., 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Research 30, 325-327.

Li, G., Siddiqui, H., Teng, Y., Lin, R., Wan, X.-y., Li, J., Lau, O.-S., Ouyang, X., Dai, M., Wan, J., 2011. Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis. Nature cell biology 13, 616-622.

López-Juez, E., 2007. Plastid biogenesis, between light and shadows. Journal of experimental botany 58, 11-26.

Mahdi, L. K., Ebrahimie, E., Adelson, D. L., Paton, J. C., Ogunniyi, A. D., 2013. A transcription factor contributes to pathogenesis and virulence in Streptococcus pneumoniae. PloS one 8, e70862.

Mahdi, L. K., Deihimi, T., Zamansani, F., Fruzangohar, M., Adelson, D. L., Paton, J. C., Ogunniyi, A. D., Ebrahimie, E., 2014. A functional genomics catalogue of activated transcription factors during pathogenesis of pneumococcal disease. BMC genomics 15, 769.

Maruyama, K., Todaka, D., Mizoi, J., Yoshida, T., Kidokoro, S., Matsukura, S., Takasaki, H., Sakurai, T., Yamamoto, Y. Y., Yoshiwara, K., 2012. Identification of cis-acting promoter elements in cold-and dehydration-induced transcriptional pathways in Arabidopsis, rice, and soybean. DNA research 19, 37-49.

Mehrotra, R., Gupta, G., Sethi, R., Bhalothia, P., Kumar, N., Mehrotra, S., 2011. Designer promoter: an artwork of cis engineering. Plant molecular biology 75, 527-536.

Mikkelsen, M. D., Thomashow, M. F., 2009. A role for circadian evening elements in cold-regulated gene expression in Arabidopsis. The Plant Journal 60, 328-339.

Moghadam, A., Ebrahimie, E., Taghavi, S., Niazi, A., Babgohari, M., Deihimi, T., Djavaheri, M., Ramezani, A., 2013. How the Nucleus and Mitochondria Communicate in Energy Production During Stress: Nuclear MtATP6, an Early-Stress Responsive Gene, Regulates the Mitochondrial F1F0-ATP Synthase Complex. Molecular Biotechnology 54, 756-769, doi:10.1007/s12033-012-9624-6.

Mongkolsiriwatana, C., Pongtongkam, P., Peyachoknagul, S., 2009. In silico promoter analysis of photoperiod-responsive genes identified by DNA microarray in rice (Oryza sativa L.). J Nat Sci 43, 164-177.

Nakamichi, N., Kiba, T., Kamioka, M., Suzuki, T., Yamashino, T., Higashiyama, T., Sakakibara, H., Mizuno, T., 2012. Transcriptional repressor PRR5 directly regulates clock-output pathways. Proceedings of the National Academy of Sciences 109, 17123-17128.

Nejad, E. S., Askari, H., Hamzelou, S., Gholami, M., 2013. Regulation of core cell cycle genes by cis-regulatory elements in Arabidopsis thaliana. Plant Knowl J 2, 69-75.

Ó'Maoiléidigh, D. S., Wuest, S. E., Rae, L., Raganelli, A., Ryan, P. T., Kwaśniewska, K., Das, P., Lohan, A. J., Loftus, B., Graciet, E., 2013. Control of reproductive floral organ identity specification in Arabidopsis by the C function regulator AGAMOUS. The Plant Cell Online 25, 2482-2503.

Oh, E., Zhu, J.-Y., Bai, M.-Y., Arenhart, R. A., Sun, Y., Wang, Z.-Y., 2014. Cell elongation is regulated through a central circuit of interacting transcription factors in the Arabidopsis hypocotyl. eLife, e03031-e03031.

Ouyang, X., Li, J., Li, G., Li, B., Chen, B., Shen, H., Huang, X., Mo, X., Wan, X., Lin, R., 2011. Genome-wide binding site analysis of FAR-RED ELONGATED HYPOCOTYL3 reveals its novel function in Arabidopsis development. The Plant Cell Online 23, 2514-2535.

Park, P. J., 2009. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10, 669-680.

Pennacchio, L. A., Rubin, E. M., 2001. Genomic strategies to identify mammalian regulatory sequences. Nature Reviews Genetics 2, 100-109.

Pilpel, Y., Sudarsanam, P., Church, G. M., 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. Nature genetics 29, 153-159.

Qiu, P., 2003. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochemical and Biophysical Research Communications 309, 495-501.

Quinn, G. G. P., Keough, M. J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.

Ramezani, A., Niazi, A., Abolimoghadam, A., Zamani Babgohari, M., Deihimi, T., Ebrahimi, M., Akhtardanesh, H., Ebrahimie, E., 2013. Quantitative Expression Analysis of TaSOS1 and TaSOS4 Genes in Cultivated and Wild Wheat Plants Under Salt Stress. Molecular Biotechnology 53, 189-197, doi:10.1007/s12033-012-9513-z.

Rawat, R., Xu, Z.-F., Yao, K.-M., Chye, M.-L., 2005. Identification of cis-elements for ethylene and circadian regulation of the Solanum melongena gene encoding cysteine proteinase. Plant molecular biology 57, 629-643.

Romualdi, C., Bortoluzzi, S., Danieli, G., 2001. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Human Molecular Genetics 10, 2133-2141.

Romualdi, C., Bortoluzzi, S., d'Alessi, F., Danieli, G. A., 2003. IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. Physiological genomics 12, 159-162.

Rumsey, D., 2011. Statistics for dummies. Wiley. com.

Rushton, P. J., Reinstädler, A., Lipka, V., Lippok, B., Somssich, I. E., 2002. Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen-and wound-induced signaling. The Plant Cell Online 14, 749-762.

Saculinggan, M., Balase, E. A., 2013. Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers. Journal of Physics: Conference Series, Vol. 435. IOP Publishing, pp. 012041.

Sato, N., Ohshima, K., Watanabe, A., Ohta, N., Nishiyama, Y., Joyard, J., Douce, R., 1998. Molecular characterization of the PEND protein, a novel bZIP protein present in the envelope membrane that is the site of nucleoid replication in developing plastids. The Plant Cell Online 10, 859-872.

Schiessl, K., Muiño, J. M., Sablowski, R., 2014. Arabidopsis JAGGED links floral organ patterning to tissue growth by repressing Kip-related cell cycle inhibitors. Proceedings of the National Academy of Sciences 111, 2830-2835.

Schulze, S., Schäfer, B. N., Parizotto, E. A., Voinnet, O., Theres, K., 2010. LOST MERISTEMS genes regulate cell differentiation of central zone descendants in Arabidopsis shoot meristems. The Plant Journal 64, 668-678.

Segal, E., Widom, J., 2009. From DNA sequence to transcriptional behaviour: a quantitative approach. Nature Reviews Genetics 10, 443-456.

Shamloo-Dashtpagerdi, R., Razi, H., Lindlöf, A., Niazi, A., Dadkhodaie, A., Ebrahimie, E., 2013. Comparative analysis of expressed sequence tags (ESTs) from Triticum monococcum shoot apical meristem at vegetative and reproductive stages. Genes & Genomics, 1-11.

Sinha, S., Tompa, M., 2003. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic acids research 31, 3586-3588.

Storey, J. D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100, 9440-9445.

Stuurman, J., Jäggi, F., Kuhlemeier, C., 2002. Shoot meristem maintenance is controlled by a GRAS-gene mediated signal from differentiating cells. Genes & development 16, 2213-2218.

Tamaoki, M., Matsuyama, T., Kanna, M., Nakajima, N., Kubo, A., Aono, M., Saji, H., 2003. Differential ozone sensitivity among Arabidopsis accessions and its relevance to ethylene synthesis. Planta 216, 552-560.

Tao, Z., Shen, L., Liu, C., Liu, L., Yan, Y., Yu, H., 2012. Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis. The Plant Journal 70, 549-561.

Terasawa, K., Sato, N., 2005. Occurrence and characterization of PEND proteins in angiosperms. Journal of plant research 118, 111-119.

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., van Helden, J., 2012. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. nature protocols 7, 1551-1568.

Usadel, B., Nagel, A., Steinhauser, D., Gibon, Y., Bläsing, O. E., Redestig, H., Sreenivasulu, N., Krall, L., Hannah, M. A., Poree, F., 2006. PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC bioinformatics 7, 535.

Vedel, V., Scotti, I., 2011. Promoting the promoter. Plant Science 180, 182-189.

Werner, T., 2001. Target gene identification from expression array data by promoter analysis. Biomolecular engineering 17, 87-94.

Wittkopp, P. J., Kalay, G., 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13, 59-69.

Xiang, T., Benkrid, K., 2009. Mersenne Twister Random Number Generation on FPGA, CPU and GPU. Adaptive Hardware and Systems, 2009. AHS 2009. NASA/ESA Conference on, pp. 460-464.

Xu, Y., Johnson, C. H., 2001. A clock-and light-regulated gene that links the circadian oscillator to LHCB gene expression. The Plant Cell Online 13, 1411-1426.

Yamaguchi-Shinozaki, K., Shinozaki, K., 2005. Organization of< i> cis</i>-acting regulatory elements in osmotic-and cold-stress-responsive promoters. Trends in plant science 10, 88-94.

Zamani Babgohari, M., Niazi, A., Moghadam, A., Deihimi, T., Ebrahimie, E., 2013. Genome-wide analysis of key salinity-tolerance transporter (HKT1;5) in wheat and wild wheat relatives (A and D genomes). In Vitro Cellular & Developmental Biology - Plant 49, 97-106, doi:10.1007/s11627-012-9478-4.

Zhang, Y., Mayba, O., Pfeiffer, A., Shi, H., Tepperman, J. M., Speed, T. P., Quail, P. H., 2013. A quartet of PIF bHLH factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in Arabidopsis. PLoS genetics 9, e1003244.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.

Zheng, J., Wu, J., Sun, Z., 2003. An approach to identify over-represented cis-elements in related sequences. Nucleic acids research 31, 1995-2005.

Zhiponova, M. K., Morohashi, K., Vanhoutte, I., Machemer-Noonan, K., Revalska, M., Van Montagu, M., Grotewold, E., Russinova, E., 2014. Helix–loop–helix/basic helix–loop–helix transcription factor network represses cell elongation in Arabidopsis through an apparent incoherent feed-forward loop. Proceedings of the National Academy of Sciences 111, 2824-2829.

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

Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., Shiu, S.-H., 2011. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 108, 14992-14997.

Tables

Table 1. Brief information of the ChIP-seq experiments used in this study. (E=Experiment)

	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**
EBI accession	E-GEOD-39215	E-GEOD-43637	E-GEOD-45213	E-GEOD-45938	E-GEOD-49282	E-GEOD-51120	E-GEOD-51770	E-GEOD-53099	E-GEOD-36361
Chiped protein	*PIF3*	*ATAF1*	*SPL7*	*AGAMOUS*	*PRR7*	*IBH1*	*ARF1*	*HBI1*	*PRR5*
Reference	(Zhang et al., 2013)	(Jensen et al., 2013)	(Bernal et al., 2012)	(Ó'Maoiléidigh et al., 2013)	-	(Zhiponova et al., 2014)	(Oh et al., 2014)	(Fan et al., 2014)	(Nakamichi et al., 2012)

Table 2. Results for K-S goodness of fit test applied to the functional categories of all and sampled promoters. K-S statistics and the rank of distributions are represented. The table shows the results for the four best fitted distributions.

	D.Uniform		Geometric		Logarithmic		Poisson	
	K-S Statistics	Rank	K-S Statistics	Rank	K-S Statistics	Rank	K-S Statistics	Rank
All Arabidopsis promoters	0.368	3	0.356	2	0.319	1	0.765	4
Sampled promoters	0.371	3	0.298	2	0.194	1	0.691	4

Table 3. Summary of analysis of the ChIP-seq experiments used in this study (E=Experiment)

	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**
Number of peaks	3897	10605	2507	1699	3167	1828	2094	6122	6122
Number of enriched CREs	15	15	9	12	8	15	12	13	15
Range of CREs occurrence	1-27	1-42	1-54	1-33	1-12	1-10	1-17	1-12	1-12

Table 4. Ranking of the fitted distributions for both the datasets obtained from the sample of Arabidopsis promoters (P) and the ChIP-seq experiments (C).

	Statistical Distributions											
	Poisson		**D.Uniform**		**Geometric**		**Logarithmic**		**Neg.Binomial**		**Binomial**	
	P	**C**	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**
Rank 1	765	104	174	1	55	9	0	0	0	0	6	0
Rank 2	119	2	780	11	68	76	15	6	9	7	9	11
Rank 3	37	5	44	44	651	29	54	36	202	1	12	0
Rank 4	72	1	2	21	226	0	661	65	39	2	0	7
Rank 5	7	2	0	15	0	0	270	7	477	0	0	29
No fit.	0	0	0	22	0	0	0	0	273	104	973	70
Total	1000	114	1000	114	1000	114	1000	114	1000	114	1000	114

Table 5. Results of the Kruskal–Wallis test on the ranks of different distributions fitted to the number of occurrences of the CREs.

		Promoter samples			**ChIP-seq experiments**	
Distribution	**N**	**Expected under H_A**	**Mean score**	**N**	**Expected under H_A**	**Mean score**
Poisson	1000	2377500	936.6	114	27873	79.5
D.Uniform	1000	2377500	1374.5	92	22494	317.8
Geometric	1000	2377500	2548.5	114	27773	191.5
Logarithmic	1000	2377500	3653.3	114	27773	350
Neg.Binomial	727	1728442.5	3773.3	10	2445	226.6
Binomial	27	64192.5	1722.7	44	10785	386.8
DF		***5***			***5***	
Chi square		***3414.4571***			***320.7017***	
p-value		***0.0001***			***0.0001***	

Table 6. Results of the Mann-Whitney U test on the pairwise comparison of Poisson distribution with the remaining distributions including D. Uniform, Geometric, Negative Binomial and Binomial. P=the sample of 1000 Arabidopsis promoters, C=ChIP-seq experiment

		D.Uniform		**Geometric**		**Logarithmic**		**Neg.Binomial**		**Binomial**	
		P	**C**	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**
Poisson	U score	2.487E5	7.5E4	1.191E5	6.2E4	1.191E5	5.4E4	37265.5	5.2E4	24621	5.01E4
	P-value	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Table 7. The list of the significant CREs between the promoters of *AtHAM1* and *AtHAM2* identified by the AC pairwise test (q-value ≤ 0.01) performed in two states of with (q-values were shown in parentheses) and without TATA box motifs. ns=non-significant

Cis element	The occurrence of the CRE in the *AtHAM1* promoter	The occurrence of the CRE in the *AtHAM2* promoter	AC q-value	Chi-square q-value	Fishers exact q-value	Function
AAGAA-motif	8	1	0.001 (0.001)	0.001 (0.003)	0.001 (0.003)	PEND binding site
5UTRPy-rich stretch	8	2	0.002 (0.003)	0.004 (0.004)	0.003 (0.004)	conferring high transcription levels
CAAT-box	23	32	0.003 (0.005)	0.006 (0.035)	0.006 (0.035)	common cis-acting element
Skn-1_motif	0	4	0.004 (0.004)	0.003 (0.003)	0.004 (0.005)	required for endosperm expression
BoxI	3	1	0.005 (0.006)	0.017^{ns} (0.01)	0.007 (0.01)	involved in light responsiveness
GT1-motif	2	0	0.007 (0.008)	0.007 (0.007)	0.008 (0.007)	involved in light responsiveness
HSE	2	0	0.008 (0.009)	0.01 (0.008)	0.008 (0.008)	involved in heat stress responsiveness
TC-rich repeat	1	3	0.009	0.018^{ns}	0.018^{ns}	involved in defense and stress responsiveness
AT-rich repeat	0	2	0.01	0.01	0.011^{ns}	element for maximal elicitor-mediated activation
(TATA-box)	36	70	(0.002)	(0.001)	(0.001)	core promoter element
(G-Box)	4	3	(0.01)	(0.041^{ns})	(0.023^{ns})	involved in light responsiveness

Table 8. The list of the significant CREs between the promoters of *AtMYC2* and *AtMYB2* identified by the AC pairwise test (q-value ≤ 0.01) performed in two states of with (q-values were shown in parentheses) and without TATA box motifs. ns= non-significant

Cis element	The occurrence of the CRE in the *AtMYC2* promoter	The occurrence of the CRE in the *AtMYB2* promoter	AC q-value	Chi-square q-value	Fishers exact q-value	Function
ABRE	4	1	0.009 (0.009)	0.012^{ns} (0.012)	0.011^{ns} (0.011)	involved in the abscisic acid responsiveness
Box I	0	3	0.006 (0.006)	0.006 (0.006)	0.003 (0.003)	involved in light responsiveness
CAAT-box	30	33	0.002 (0.004)	0.008 (0.028^{ns})	0.007 (0.018^{ns})	conferring high transcription levels
ERE	0	3	0.007 (0.007)	0.006 (0.006)	0.003 (0.003)	ethylene-responsive element
G-Box	11	0	0.001 (0.001)	0.001 (0.035^{ns})	0.001 (0.001)	involved in light responsiveness
MBS	3	0	0.008 (0.008)	0.005 (0.005)	0.009 (0.009)	MYB binding site involved in drought-inducibility
TA-rich region	4	0	0.003 (0.003)	0.003 (0.003)	0.005 (0.005)	enhancer
TATA-box	84	90	0.002 (0.002)	0.009 (0.009)	0.006 (0.006)	common cis-acting element
Unnamed_4	8	3	0.005 (0.005)	0.008 (0.008)	0.008 (0.008)	-

Figures

Figure 1. The workflow diagram of the proposed method for identification of statistically differential CREs between two promoters. A: Determining the statistical distribution of number of CREs occurrences. B: Employing the pairwise test and verifying the biological significance of the identified CREs.

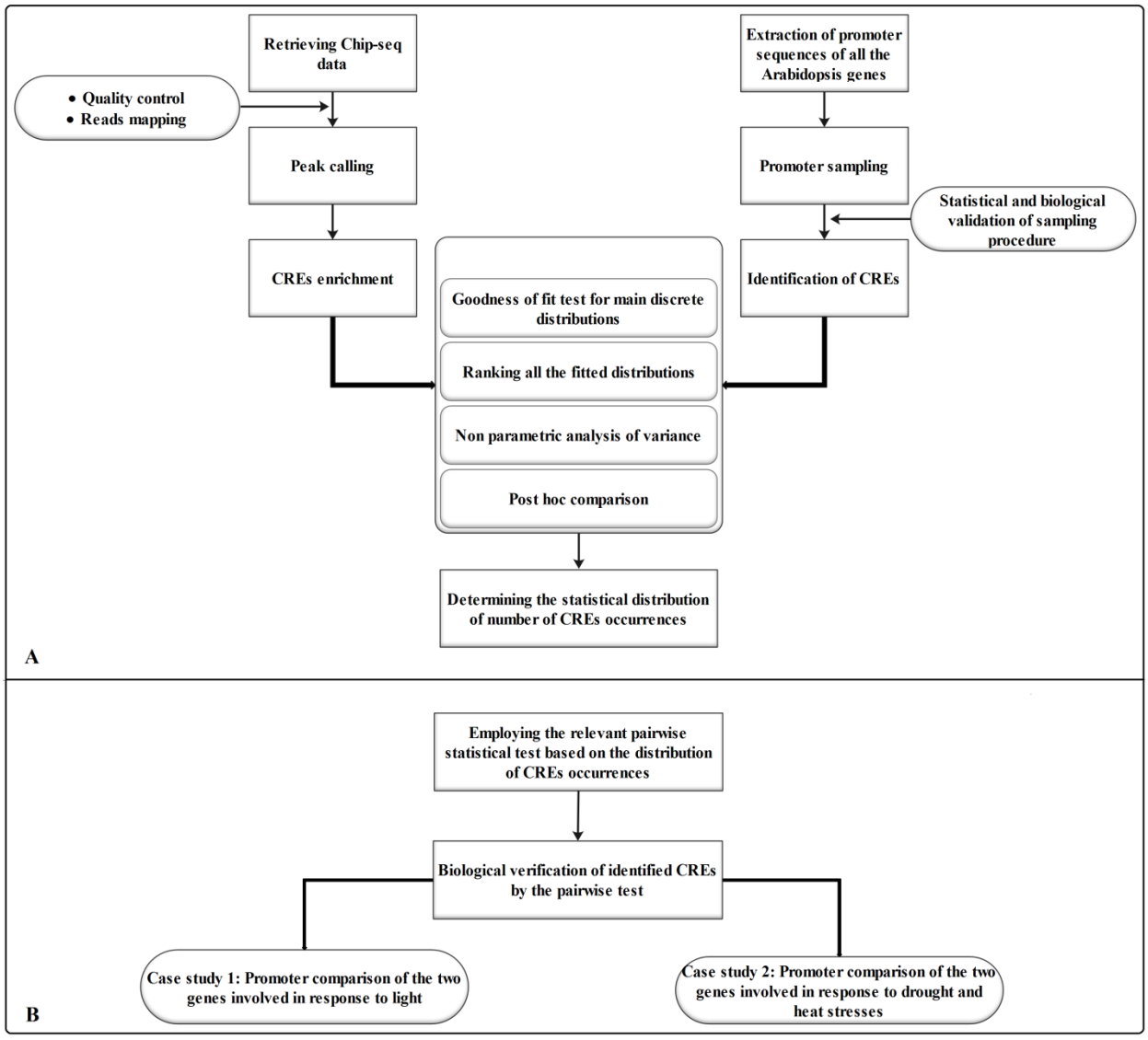


Figure 2. Cumulative distribution function (CDF) of the total and sampled promoter sequences. The red line shows the D. Uniform distribution of all promoters, and the black line shows the CDF line of the sampled promoters.

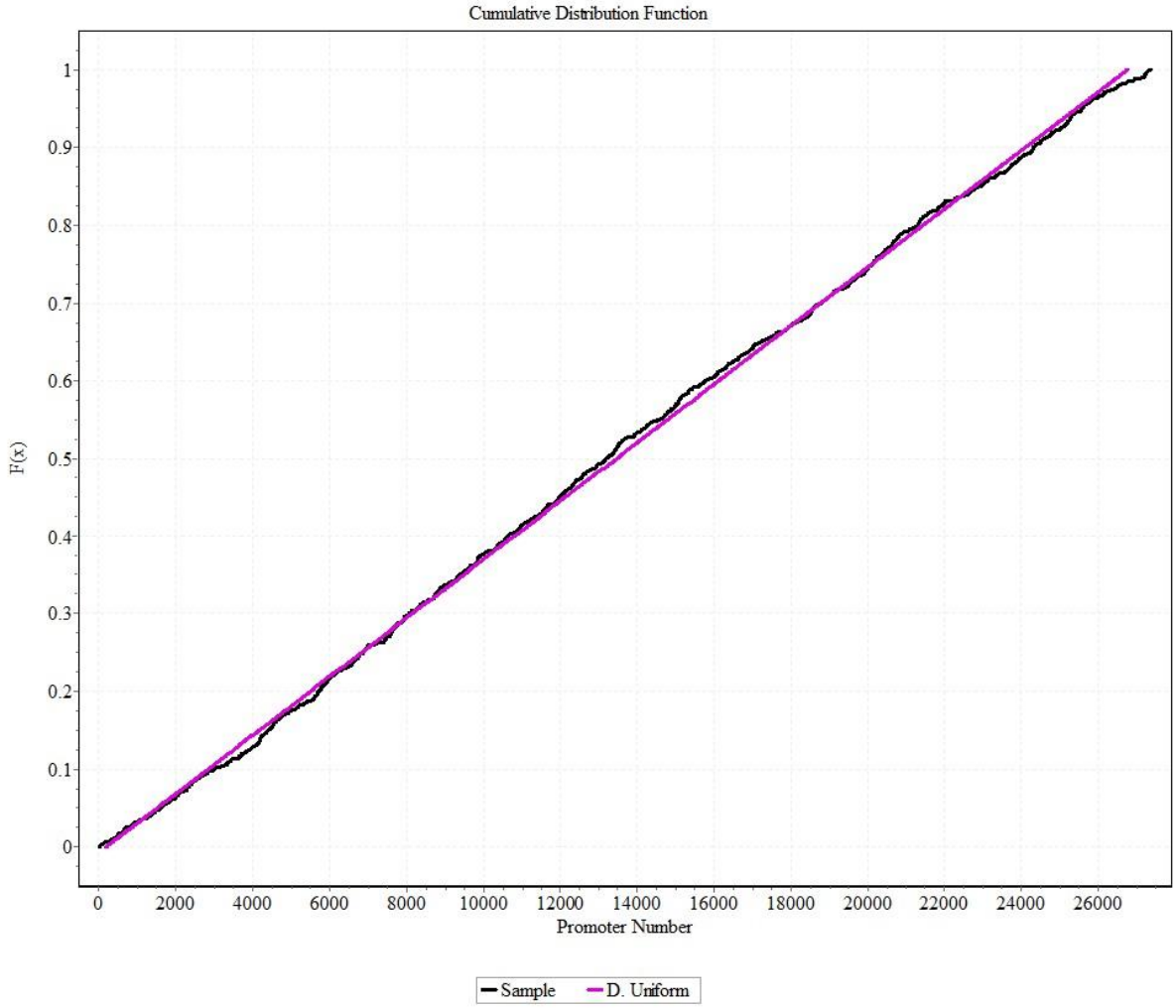


Figure 3. Comparative chart of functional categories of all and sampled Arabidopsis promoter IDs.

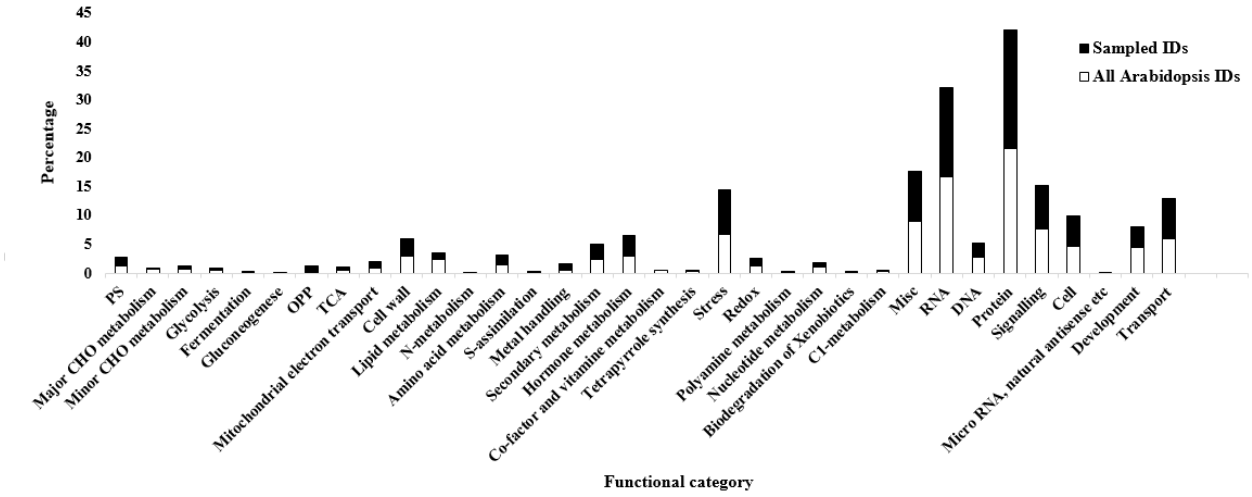
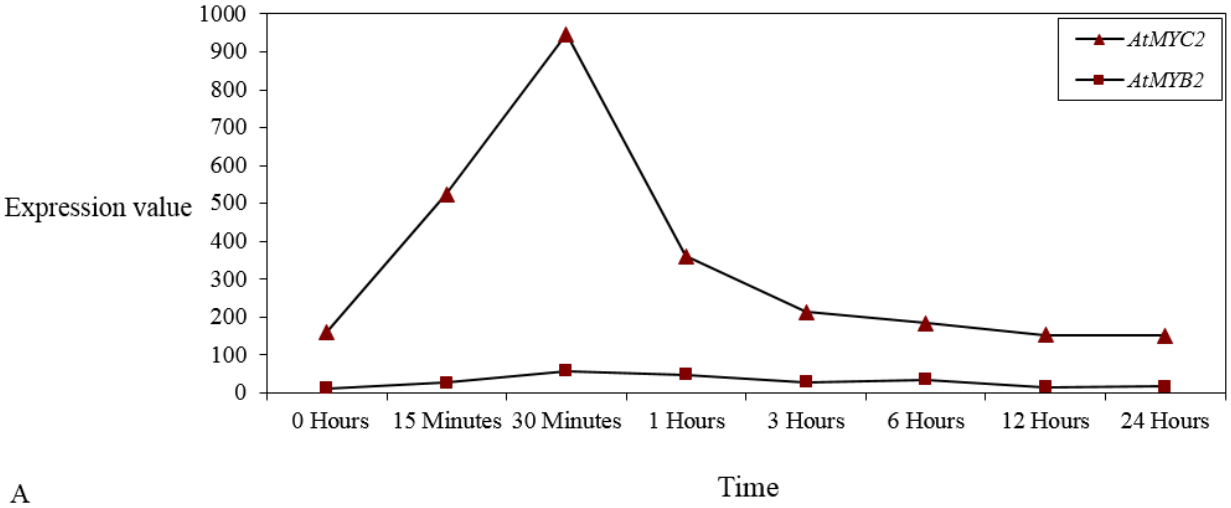
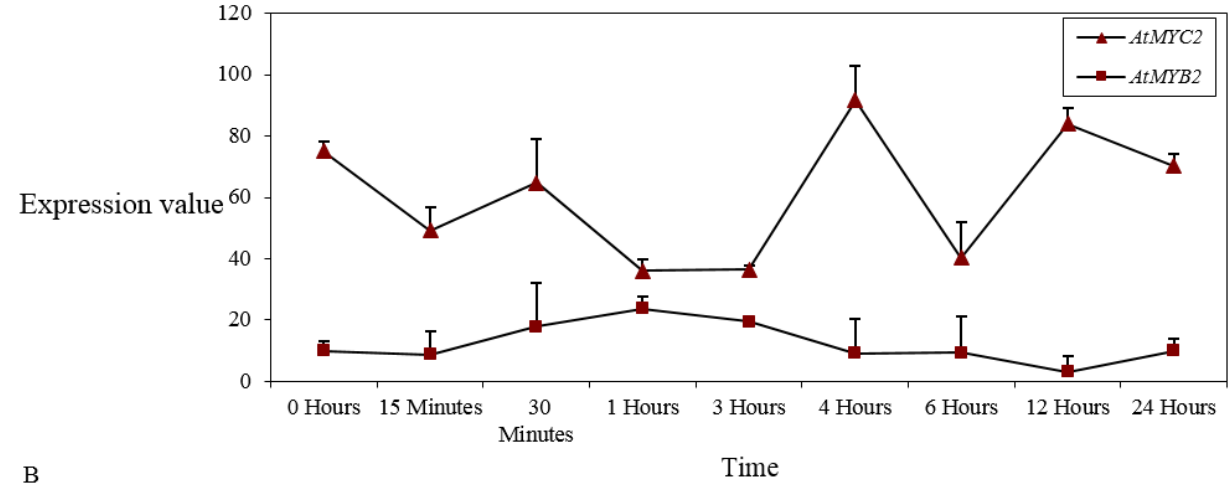


Figure 4. Expression profiles of *AtMYC2* and *AtMYB2*genes in drought (A) and heat (B) stress, respectively.



A



B