PLOS ONE

RESEARCH ARTICLE

Computational exploration of *cis*-regulatory modules in rhythmic expression data using the "Exploration of Distinctive CREs and CRMs" (EDCC) and "CRM Network Generator" (CNG) programs

Pavlos Stephanos Bekiaris☯, Tobias Tekath☯, Dorothee Staiger, Selahattin Danisman*

RNA Biology and Molecular Physiology, Faculty of Biology, Bielefeld University, Bielefeld, Germany

☯ These authors contributed equally to this work.
* selahattin.danisman@uni-bielefeld.de

Citation: Bekiaris PS, Tekath T, Staiger D, Danisman S (2018) Computational exploration of *cis*-regulatory modules in rhythmic expression data using the "Exploration of Distinctive CREs and CRMs" (EDCC) and "CRM Network Generator" (CNG) programs. PLoS ONE 13(1): e0190421. https://doi.org/10.1371/journal.pone.0190421

Editor: Kentaro Yano, Meiji Daigaku - Ikuta Campus, JAPAN

Received: June 1, 2017

Accepted: December 14, 2017

Published: January 3, 2018

Data Availability Statement: All data used for this study have been published earlier and are referred to in the paper. All data produced in this study are given in the paper and supplemental files. The programs are deposited at https://sourceforge.net/projects/edcc/.

Funding: This work was supported by a core grant of Bielefeld University to D. Staiger. The funders had no role in study design, data collection and

Abstract

Understanding the effect of *cis*-regulatory elements (CRE) and clusters of CREs, which are called *cis*-regulatory modules (CRM), in eukaryotic gene expression is a challenge of computational biology. We developed two programs that allow simple, fast and reliable analysis of candidate CREs and CRMs that may affect specific gene expression and that determine positional features between individual CREs within a CRM. The first program, "Exploration of Distinctive CREs and CRMs" (EDCC), correlates candidate CREs and CRMs with specific gene expression patterns. For pairs of CREs, EDCC also determines positional preferences of the single CREs in relation to each other and to the transcriptional start site. The second program, "CRM Network Generator" (CNG), prioritizes these positional preferences using a neural network and thus allows unbiased rating of the positional preferences that were determined by EDCC. We tested these programs with data from a microarray study of circadian gene expression in *Arabidopsis thaliana*. Analyzing more than 1.5 million pairwise CRE combinations, we found 22 candidate combinations, of which several contained known clock promoter elements together with elements that had not been identified as relevant to circadian gene expression before. CNG analysis further identified positional preferences of these CRE pairs, hinting at positional information that may be relevant for circadian gene expression. Future wet lab experiments will have to determine which of these combinations confer daytime specific circadian gene expression.

Introduction

Temporal and spatial regulation of gene expression is a common process in eukaryotic organisms. Transcription factor-mediated control of gene expression has been studied for decades and involves complex interplays between DNA and proteins. Transcription factors bind to CREs, i.e. short sequences that are usually situated upstream of coding sequences, and affect the set-up of the transcriptional machinery. Today large numbers of CREs are known, e.g. in

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

humans [1,2], yeast [3], and plants [4,5]. However, CREs not only function as single elements, they also combine with other CREs. The sum of all CREs that convey specific gene expression is called *cis*-regulatory module (CRM) [6]. The gene expression patterns regulated by CRMs are highly dependent on the composition of these CRMs, i.e. the number of repeats of a specific CRE [7], the combination of CREs present [8], the spacing between CREs [9,10], and the CREs' positions within the CRM [9,10]. In plants, CRMs control the expression of genes that are involved in the cell cycle, photosynthesis, development of the male germline, stress response, and circadian gene expression [5,11–13].

Circadian gene expression denotes rhythmic expression of a gene that follows a rhythm of about 24 hours (from 'circa diem' = 'about a day'). The circadian clock, a biological timekeeper that consists of proteins controlling each other in regulatory feedback loops, maintains this rhythm even under free-running conditions, i.e. when there is no external rhythm, a *Zeitgeber*, indicating the begin of a day. In Arabidopsis, up to 90% of all genes display rhythmic behavior under at least one light/temperature regime [14]. Rhythmic expression under free-running conditions has been shown in up to 36% of all genes [15], covering a plethora of physiological processes including photosynthesis [16], starch metabolism [17], growth [18], flowering time determination [19,20] and regulation of the plant immune system [21,22].

Several CREs are known to confer circadian gene expression. The evening element (AAAATATCT) was identified based on its over-representation in circadianly regulated genes that exhibited maximum expression in the subjective evening [23,24]. The morning element (AAAAAATCT) was identified in a mutational analysis of the *PSEUDO-RESPONSE REGULATOR 5* promoter, a clock gene that is involved in repression of the core clock genes *CIRCADIAN CLOCK ASSOCIATED 1 (CCA1)* and *LATE ELONGATED HYPOCOTYL* during the day [25,26]. Michael and colleagues conducted bioinformatics analyses of microarray experiments in which Arabidopsis was subjected to 11 different rhythmic conditions (e.g. photocycles, thermocycles, short days, long days). Here they identified the protein box (ATGGGCC), the telobox (AAACCCTT) and the starch box (AAGCCC) elements as CREs that confer midnight-specific gene expression and that are conserved between Arabidopsis, rice and poplar [14]. The so-called Hormone-up-in-Dawn (HUD) element (CACATG) was found to be over-represented in genes that respond to brassinosteroid and auxin treatments and in genes that are expressed preferentially at dawn [27].

The identification of CREs that correlate with specific gene expression has long been established [28–31]. For example, Bussemaker and colleagues detected new regulatory motifs in the upstream regions of genes by correlating the presence of these motifs with genome-wide gene expression in *Saccharomyces cerevisiae* [28]. Another tool, called *'in silico* expression analysis', determines which genes contain a given CRE and compares the expression of these genes in microarray data [31]. With the help of this program, the authors were able to determine that a CGACTTTT sequence was involved in the response of Arabidopsis to infection with the fungus *Botrytis cinerea* [31]. In another approach, the MEME suite [30] was used to detect over-represented CREs in rhythmically expressed genes and further gene expression profiles were compared with a neural network approach [32]. The respective calculations were so computation intensive that a supercomputer was used for this study [32]. Most programs focus on the detection and analysis of single CREs, although it is long established that CREs affect gene expression in a combinatorial manner. Studies to identify and analyze CRMs are less straightforward. For this, Hidden Markov models have been successfully used in simulated and real data sets of fruitflies and humans [29]. Also, Hidden Markov Models have been used to identify CRMs by analyzing correlations between binding sites and multispecies comparisons in yeast and fruitfly experimental data [33]. CRMs were further detected using position weight matrices [34,35], Monte Carlo methods [36], phylogenetic approaches [37], and chromatin signatures and neural networks, respectively [38].

We propose a simpler approach to determine candidate CREs and CRMs that may confer specific gene expression. This approach reliably analyzes the potential of millions of CRMs in a relatively short time. It uses programs that run on a table-top computer and can be used by users with minimal bioinformatics knowledge. These two programs are called "Exploration of Distinctive CREs and CRMs" (EDCC) and "CRM Network Generator" (CNG). EDCC correlates the presence and positions of known CREs/CRMs with gene expression data, and CNG further assesses the importance of positional features within CRMs that were determined by EDCC. We tested the performance of these programs using data from a circadian microarray experiment of *Arabidopsis thaliana* seedlings [14]. EDCC identified both known and candidate CREs and CRMs in circadian gene expression control. CNG analysis shows that some of the identified CRE pairs occur at specific locations in the promoters of downstream genes, indicating functional CRMs in circadian gene expression.

Results

Principle of EDCC analysis

We designed two programs to analyze whether user-determined CREs and CRMs correlate with specific expression patterns and thus, whether they may be involved in regulation of the specific gene expression (Fig 1). The first program, EDCC, correlates candidate CREs and CRMs with gene expression patterns, and compares this with the expression pattern of all genes under different experimental conditions. For pairs of CREs, EDCC further determines whether they are positioned at a specific distance to each other, whether they are positioned in a specific order towards the transcriptional start site (TSS), and whether the two CREs are positioned at a specific distance to the TSS.

EDCC uses three initial data sets: gene expression data, promoter sequences of the respective genes, and a list of CREs and CRMs defined by the user. The gene expression data needs to be categorized over the different treatments that the user wants to analyze. Only genes that are differentially expressed between treatments will be included in the analysis. EDCC categorizes each gene according to its maximum gene expression, and each gene is categorized in only one condition. EDCC then plots the percentage of genes per category, which results in the background distribution (Fig 2A). Queried with a CRE/CRM, EDCC determines the promoters that contain the motifs and the expression category that the respective genes belong to. EDCC then plots the percentage of genes that contain the CRE/CRM per category, resulting in a distribution of expression maxima (DEM) which is specific for each given CRE/CRM (Fig 2B). This DEM is then compared to the background distribution. A CRE/CRM that has no effect on gene expression in the analyzed conditions should lead to a DEM that is similar to the background distribution (Fig 2B). Inversely, a CRE/CRM that affects genes towards expression under a specific treatment or condition should lead to a shift between the DEM and the background distribution (Fig 2B). EDCC determines a threshold at which a CRE/CRM is identified as a candidate by calculating the DEMs of a large number of random CREs and determining the standard deviation from the mean for each category. A CRE that correlates with a DEM that differs from the background by at least one standard deviation in one or more conditions is identified as a candidate CRE (Fig 2C). EDCC also allows for more conservative approaches by increasing the threshold to a multifold of the standard deviation.

Testing EDCC on circadian microarray data

We tested EDCC with data of a circadian microarray experiment, in which Arabidopsis seedlings were entrained for nine days in a 12 h dark/12 h light cycle and then transferred into continuous light [14]. Seedlings were harvested every four hours for 48 hours, beginning at

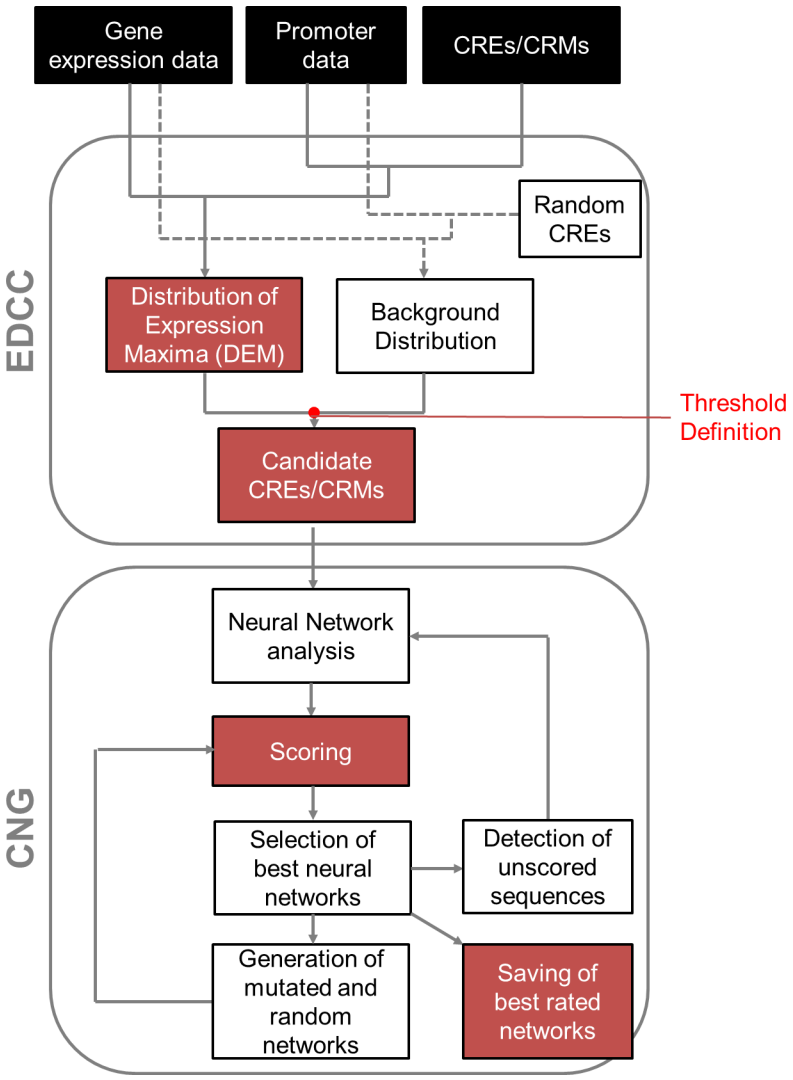

Fig 1. Flowchart of EDCC and CNG analyses. The flowchart shows which data input is needed for EDCC and CNG analyses, the principle behind their functions and the outputs of the two programs. Further detailed graphics explain the calculations of EDCC and CNG in Supporting figures S1 and S2 Figs, respectively.

https://doi.org/10.1371/journal.pone.0190421.g001

Zeitgeber Time 0 (ZT0), i.e. the hour at which the lights are switched on. Gene expression for each time point was determined using an Affymetrix *Arabidopsis* ATH1 gene chip (E-MEXP-1304) [14]. We identified circadianly expressed genes using ARSER [39] and categorized the genes into six categories according to the respective peak expression times. We found that 3561 genes (10% of the TAIR10 genome annotation) were expressed circadianly under these experimental conditions. A majority of these exhibited peak expression between ZT8 and ZT12 (26%), i.e. before the subjective dusk (Fig 3A). This was followed by the category ZT20-ZT0 (18%), i.e. just before dawn, with all other categories exhibiting lower percentages (Fig 3A). This background distribution was queried with random CREs of 5–8 bp lengths to determine the standard deviation and hence the threshold for further EDCC analyses. To test the optimum number of CREs for background models, we queried EDCC with 10, 50, 100, 500, and 1000 random CREs, respectively, and analyzed the difference between the background distribution and the DEM of the randomized CREs. This difference decreased with a higher

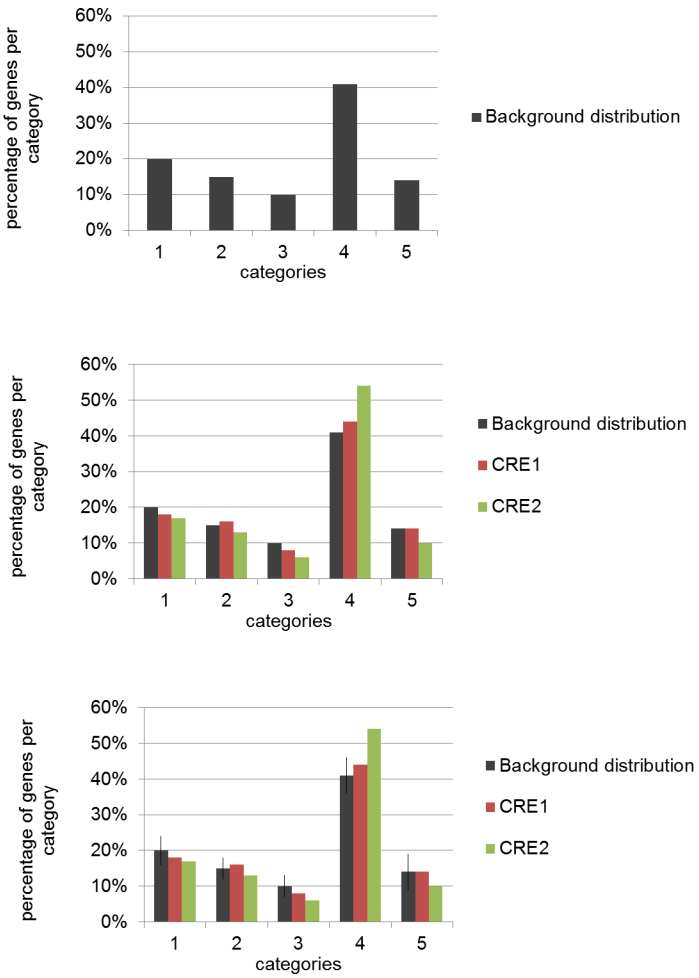

Fig 2. Principle of EDCC analysis. A) Presented is the background distribution of gene expression across five generic categories. B) DEM of two exemplary CREs compared to the background distribution. Genes containing CRE1 (red) do not correlate with a shift in the DEM, whereas genes containing CRE 2 (green) do. C) Addition of standard deviations after analysis of random CREs allows establishing thresholds for the determination of candidate CREs.

https://doi.org/10.1371/journal.pone.0190421.g002

number of queries (Fig 3B). As the difference between 100 and 1000 queries was negligible, we decided to further use 100 random CREs to determine EDCC thresholds (Fig 3C).

Testing EDCC with known circadian clock CREs

After having established a random background with thresholds for the circadian microarray experiment, we tested CREs that are known to confer circadian gene expression, i.e. the evening element, the morning element, the three midnight elements and the HUD-domain [14,23,25,27]. Genes containing the evening element and the telobox element (AAACCCTT) exhibited DEMs that differed from the background at ZT8-12 (evening) and ZT16-20 (midnight), respectively (Fig 4). As the evening element indeed confers evening specific gene expression [23], this indicates that EDCC is able to correctly identify CREs that may be involved in circadian gene expression and the time point that is affected by the CRE. The evening element is marked "candidate" in the EDCC analysis even when using a threshold of three standard deviations, correctly indicating the strength of the evening element as a CRE conferring evening specific circadian gene

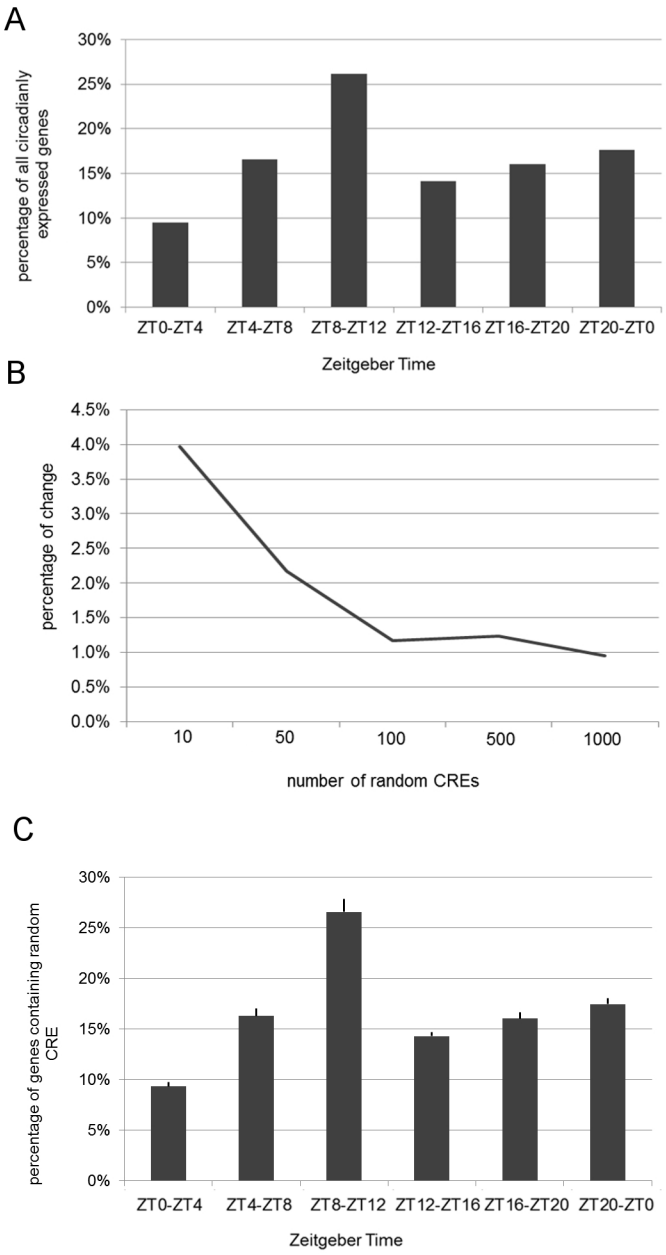

Fig 3. EDCC analysis of circadian microarray data. A) Distribution of maximum gene expression times of circadianly genes expressed in Arabidopsis seedlings [14], which was used as background distribution for the EDCC analysis. Distribution is shown as percentage of all circadianly expressed genes. Maxima are categorized in six categories, i.e. ZT0-ZT4 (morning), ZT4-ZT8 (midday), ZT8-ZT12 (evening), ZT12-ZT16 (early night), ZT16-ZT20 (midnight), ZT20-ZT0 (before dawn). B) Decrease of standard deviations of randomized CREs in percent plotted against the number of randomized CREs used (10, 50, 100, 500, and 1000 random CREs, respectively). C) Mean DEM of random CREs after 100 iterations, including standard deviations.

https://doi.org/10.1371/journal.pone.0190421.g003

expression. EDCC also correctly identifies the telobox element as a CRE that confers midnight specific gene expression between ZT16 and ZT20 [14]. All other tested CREs were not indicated as candidates by EDCC, which means that the EDCC analysis is more conservative than other types of analysis.

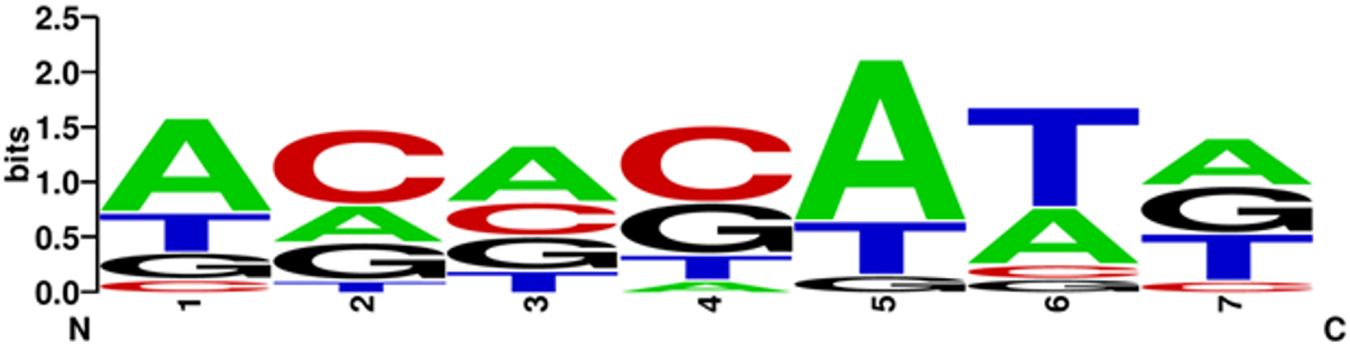

Fig 4. DEM for genes containing the evening and telobox elements compared to randomized background. Shaded areas indicate one to three standard deviations distance from the background.

https://doi.org/10.1371/journal.pone.0190421.g004

Testing EDCC performance with 1755 single CREs

We then tested EDCC performance using 1755 CREs that are known in plants [5]. We only counted CREs that were present in at least 10 promoters to prevent a false positive effect on the DEM. We ran the analysis five times and found 182.8 candidate CREs on average, i.e. 10.4% of all queries were identified as candidate CREs for at least one time point (S1 Table). Although EDCC creates a new random background in each run, 98% of the CREs that were found overlapped in all five iterations. We also calculated the quartile dispersion coefficient [40] and found a 0.27% variation between runs, indicating that the results generated by EDCC are extremely consistent.

We then ran the same test under more conservative conditions. In the first approach, we increased the number of promoters that a CRE must be present in to 15, 20, and 30, respectively, and ran each test five times. This led to smaller numbers of candidate CREs (Fig 5A; S2 Table). In each case, the overlap among the five iterations of the analysis was large, i.e. 98%, 100%, and 100%, respectively. In the second approach, we increased the threshold to two, three or four standard deviations, respectively. This dramatically reduced the number of candidate CREs (Fig 5B). Also here, we found a high overlap between the individual runs. At a distance of minimum three standard deviations, we found only one consistent candidate CRE: GACGTGTA, which has been described as an abscisic acid (ABRE) binding response element [41]. The list of CREs that were found to be candidates in all five analyses with a threshold of two standard deviations is given in Table 1. Non-surprisingly, the evening element was one of the candidates that were identified by the EDCC analysis. Further candidate elements that have been found are involved in light-controlled or circadian gene expression, e.g. MYB transcription factor binding sites, which are involved in the light responsiveness of enzymes of the flavonol biosynthetic pathway in Arabidopsis [42], and GATA and G box motifs, which belong to the earliest promoter elements found in light-regulated and circadian clock regulated genes [43,44]. Also a binding site for TCP transcription factors was found (Table 1). These transcription factors have recently been shown to bind to clock genes and affect their expression [45–47]. Abscisic acid (ABA) response elements, which are similar to the G box, have been found several times by the EDCC analysis (Table 1). ABA signaling has been found to be connected to the circadian clock in several studies [48–50]. In case of non-annotated CREs, we used agriGO to determine the enrichment of gene ontology (GO) terms [51] (Table 1).

Analysis of pairwise CRE combinations

We then analyzed the simplest type of CRMs: pairwise combinations of CREs. Here, we combined each of the 1755 CREs with each other, leading to 1,540,890 tested combinations,

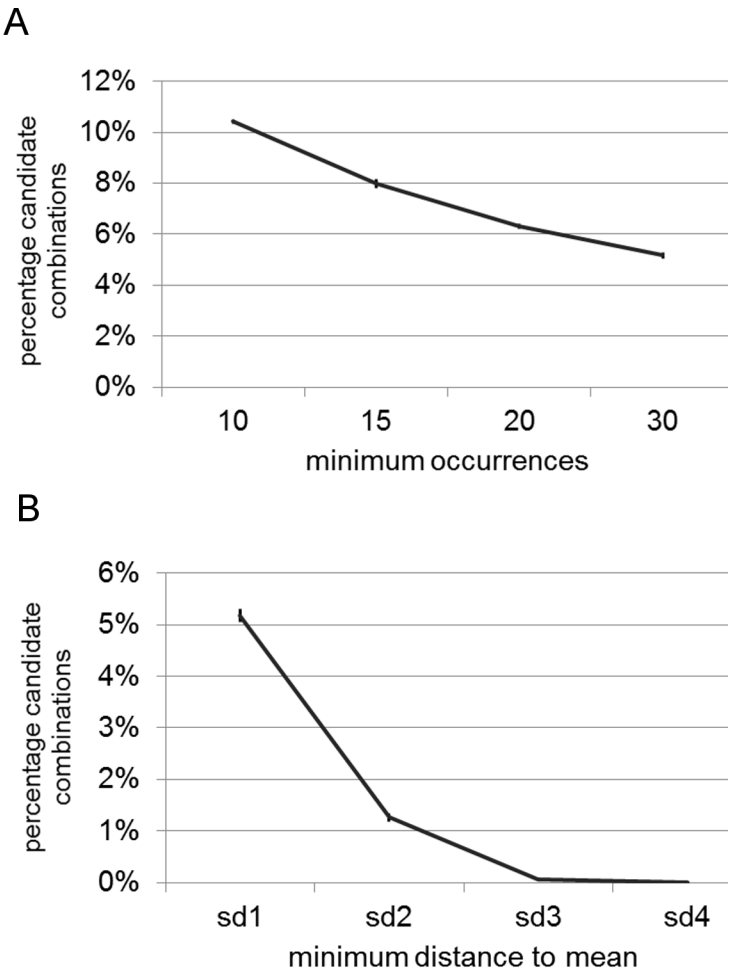

Fig 5. Number of candidate CREs under different EDCC settings. A) Graph depicting the decrease of candidate CREs when increasing the minimum number of promoters a CRE has to be present in. B) Graph depicting the decrease of candidate CREs when increasing the thresholds from one to four standard deviations (sd).

https://doi.org/10.1371/journal.pone.0190421.g005

including homotypic combinations. Analogous to the tests with single CREs, we first estimated the conditions under which the test needed to be conducted. Under the least conservative conditions (10 occurrences, one standard deviation threshold), we found on average 192,010.6 candidate combinations (12.46% of all combinations). Increasing the number of minimum occurrences to 15, 20 and 30 led to a decrease of candidate CRMs analogous to the case in single CREs (Fig 6A). A stepwise increase of the threshold distance from the background from one to six standard deviations led to a strong decrease in the number of candidate CRMs, respectively (Fig 6B). Under the most restrictive conditions—at minimum six standard deviations and minimum 30 hits in promoters—only one combination remained: the evening element together with a Dc3 Promoter-Binding Factor-1 and 2 (DPBF1&2) element, which first has been described as an ABA responsive element in the promoter of the carrot *Dc3* gene [67].

21 candidate CRMs were found with a threshold of five standard deviations and a minimum occurrence of 30 promoters in all five repetitions of the EDCC analysis (Table 2). The evening element was present in six candidate CRMs. The evening element was found in combinations with the LEAFY consensus site motif [68], the DPBF1&2 binding site motif

Table 1. Single CREs that were identified as candidates with a threshold of at least two standard deviations.

single sequence	interesting timepoints	sum of matches	Annotation or agriGO enrichment	Reference
AAAATATCT	ZT8-ZT12	267	evening element	[23]
AACCTACC	ZT20-ZT0	63	MYB binding site promoter	[52]
AATATTTTTATT	ZT4-ZT8	36	AT1BOX AT-1 box (AT-rich element)	[43,53,54]
AAWGTATCSA	ZT20-ZT24	32	Wound-responsive element	[55]
ATCCAACC	ZT4-ZT8	81	MYB1 binding site motif	[56]
ATCCTACC	ZT16-20, ZT20-ZT24	33	MYB1 binding site motif	[56]
CAATGATTG	ZT8-12, ZT16-ZT20	35	ATHB5 binding site motif	[57]
CACCTACC	ZT8-ZT12, ZT20-ZT0	42	MYB1 binding site motif	[56]
CACGCAAT	ZT8-ZT12	33	Sequence found in auxin responsive genes of Soybean	[58]
CAGAAGATA	ZT16-ZT20	44	GATA motif binding factor	[59]
CCAGGTGG	ZT16-ZT20	38	Class I TCP binding site in rice	[60]
GACGTGTA	ZT16-ZT20	48	ABRE-like binding site motif	[41]
GATGAYRTGG	ZT12-ZT16	39	opaque-2 binding site of maize b-32 type I ribosome-inactivating protein gene	[61]
GCGGCAAA	ZT16-ZT20	37	E2F binding site in tobacco Ribonucleotide reductase gene promoter	[62]
MAGGTAAGT	ZT8-ZT12	56	*cis*-element in exon-intron splice junctions of plant introns	[63]
MCACGTGGC	ZT4-ZT8	80	G box/Conserved sequence upstream of light-regulated genes	[64]
NCCCGCCA	ZT16-ZT20	68	enriched in GOs DNA replication, DNA-dependent DNA replication, DNA metabolism etc	
TAACTCGTT	ZT4-ZT8, ZT8-ZT12, ZT16-ZT20, ZT0-ZT4	32	MYB2 binding site motif	[65]
TAACTGGTT	ZT12-ZT16	55	MYB2 binding site motif	[65]
TACGTGGA	ZT4-ZT8	63	ABRE-like binding site motif	[41]
TACGTGTC	ZT16-ZT20	59	ABRE-like sequence found in rice	[66]

Annotations are from AtCOEcis [5], alternatively enriched GO terms according to agriGO [51] are given.

https://doi.org/10.1371/journal.pone.0190421.t001

described above, an undefined motif (AATNCCNC), elements that were found in genes that are involved in glucosyltransferase activity (ATGGCNNC), calmodulin regulated protein kinase activity and ATPase activity (GAANGAGA), and in auxin signaling (ACACATG), respectively (Table 2). Other candidate CRE combinations contained G boxes together with an element that is overrepresented in metal homeostasis genes, and the ABRE-like motif (GACGTGTA) together with an undefined motif (CNANAGAA). Also here, unannotated CREs were subjected to GO term analysis using agriGO [51].

Mutational analysis of a CRE pair: An example

Finding the evening element represented in six of the 21 CRE pairs led us to an interesting question: is it possible that EDCC identifies a CRE pair as a candidate only because one of the two CREs would be identified as a candidate in any case? This might lead to false positive CRE pairs. We tested whether the evening element/DPBF1&2 binding element (ACACATG) pair is specific by generating mutations within both CREs and subjecting these to EDCC analysis. We generated one million unique CRE pairs including 0 to 16 mutations from the original pair each. Of the one million pairs, only 13 pairs performed comparably to the original pair in the EDCC analysis. All other mutant combinations did not correlate with a shift in peak expression times. Of the 13 mutations, none included a mutated evening element, indicating that mutation of the evening element may have a stronger effect on evening-specific gene

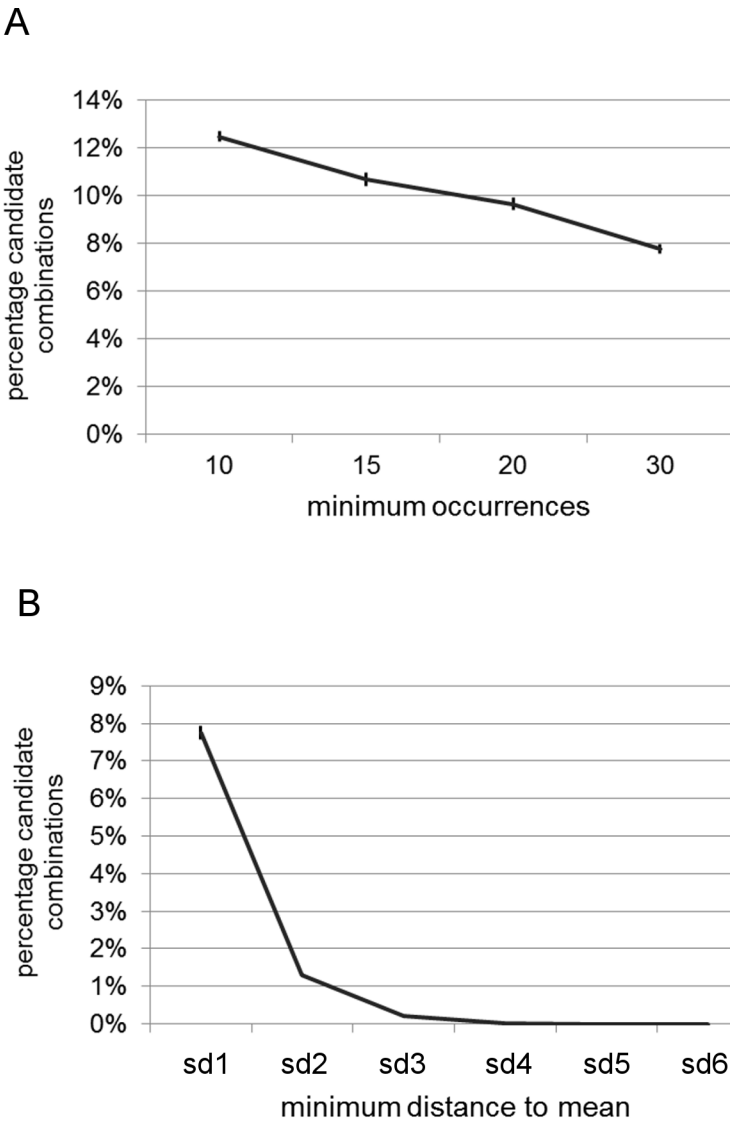

Fig 6. Analysis of pairwise CRE correlation with circadian gene expression. A) Graph depicting the decrease in candidate CRMs when increasing minimum number of promoters the CRM must be present in. B) Graph depicting the decrease of candidate CRMs with increasing thresholds (sd: multifold of standard deviation from background).

https://doi.org/10.1371/journal.pone.0190421.g006

expression than mutation of the DPBF1&2 element. This also indicated that indeed the evening element may be more important for the specific gene expression conferred by the CRE pair than the DPBF1&2 element. We were further able to determine which nucleotides of the DPBF1&2 element correlated with a better performance in the EDCC analysis, i.e. positions 1, 4, 5 and 6 of the ACACATG sequence (Fig 7). It is however not possible to finally decide whether one of the two elements is irrelevant for a possible function as a CRM without resorting to wetlab experiments, which were beyond the scope of this study.

Gene ontology analysis of pairwise CRE combinations

The EDCC output includes a list of Arabidopsis Genome Initiative (AGI) identifiers for all those genes that contain a CRE or CRM in their promoters. A GO analysis was conducted

Table 2. Candidate CRMs in circadianly expressed genes.

no.	sequence combination	interesting timepoints	sum of matches	annotation+citation element 1	annotation+citation element 2
1	AAAATATCT, ATGGCNNC	ZT8-ZT12	31	evening element [23]	enriched in GO glucosyltransferase activity
2	AAAATATCT, CCAGTG	ZT8-ZT12	38	evening element [23]	LFY consensus binding site motif [68]
3	AAAATATCT, GAANGAGA	ZT8-ZT12	56	evening element [23]	enriched in GO calmodulin regulated protein kinase activity and ATPase activity
4	AATNCCNC, AAAATATCT	ZT8-ZT12	52	undefined	evening element [23]
5	ACACATG, AAAATATCT	ZT8-ZT12	30	DPBF1&2 binding site motif [67]	evening element [23]
6	ACACCGG, AAGNGTNG	ZT12-ZT16	30	DPBF1&2 binding site motif [67]	enriched in GO calmodulin regulated protein kinase activity
7	ACANTACN, ATCCAACC	ZT4-ZT8	44	undefined	MYB1 binding site motif [52]
8	ACANTACN, MCACGTGGC	ZT4-ZT8	34	undefined	G box; Conserved sequence upstream of light-regulated genes [64]
9	AGNGATAN, MCACGTGGC	ZT4-ZT8	33	enriched in metal ion homeostasis	G box; Conserved sequence upstream of light-regulated genes [64]
10	ANCACATG, AAAATATCT	ZT8-ZT12	36	enriched in auxin stimulus	evening element [23]
11	ATACGTGT, TAACAAA	ZT0-ZT4	40	Z-DNA-forming sequence found in the Arabidopsis chlorophyll a/b binding protein gene (cab1) promoter; Involved in light-dependent developmental expression of the gene [69]	MYBGAHV Central element of gibberellin (GA) response complex (GARC) in high-pI alpha-amylase gene in barley (H.v.) [70]
12	ATGNTTCA, ACGTGGC	ZT16-ZT20	39	enriched in GO protein serine/threonine kinase activity	enriched in GO glucan biosynthesis, chloroplast part
13	CATGCATG, AGNAACAA	ZT4-ZT8	34	RY-repeat motif; Binding site of FUS3; TRAB1, bZIP transcription factor, interacts with VP1 and mediates ABA-induced transcription [71]	n/a
14	CATGCATG, NGCNTGAA	ZT4-ZT8	30	RY-repeat motif; Binding site of FUS3; TRAB1, bZIP transcription factor, interacts with VP1 and mediates ABA-induced transcription [71]	n/a
15	CCNNCACN, GTGATCAC	ZT0-ZT4	32	n/a	PIATGAPB found in the Arabidopsis thaliana GAPB gene promoter; Mutations resulted in reductions of light-activated gene transcription [72]
16	CNANAGAA, GACGTGTA	ZT16-ZT20	32	n/a	ABRE-like binding site motif [41]
17	CTCATTTN, AGATCCAA	ZT4-ZT8	30	n/a	AG-motif found in the NtMyb2 gene promoter; AGP1binding site [73]
18	GACGTGTA, CNNACANC	ZT16-ZT20	30	ABRE-like binding site motif [41]	n/a
19	TCNTNAGA, CAAAACGC	ZT16-ZT20	31	n/a	CDA1ATCAB2 CDA-1 binding site in DtRE (dark response element) f of chlorophyll a/b-binding protein2 gene in Arabidopsis [74]
20	TGTCACA, TGAGTCA	ZT4-ZT8	31	motif found cucumisin gene promoter in melon fruits [75]	Required for endosperm-specific expression [76]
21	TGTGNGNA, TAGTGGAT	ZT4-ZT8	32	enriched in GO external encapsulating structure organization and biogenesis, cell wall biogenesis	negative regulatory region in promoter region of Brassica napus (B.n.) extA extensin gene [77]

Candidate CRE pairs are present in at least 30 promoters and correlate with DEMs that deviate from the background by at least five standard deviations. Annotations are as given by AtCOEcis [5], alternatively, enriched GO terms according to agriGO [51] are given, n/a depicts CREs without annotation or enriched GO term.

https://doi.org/10.1371/journal.pone.0190421.t002

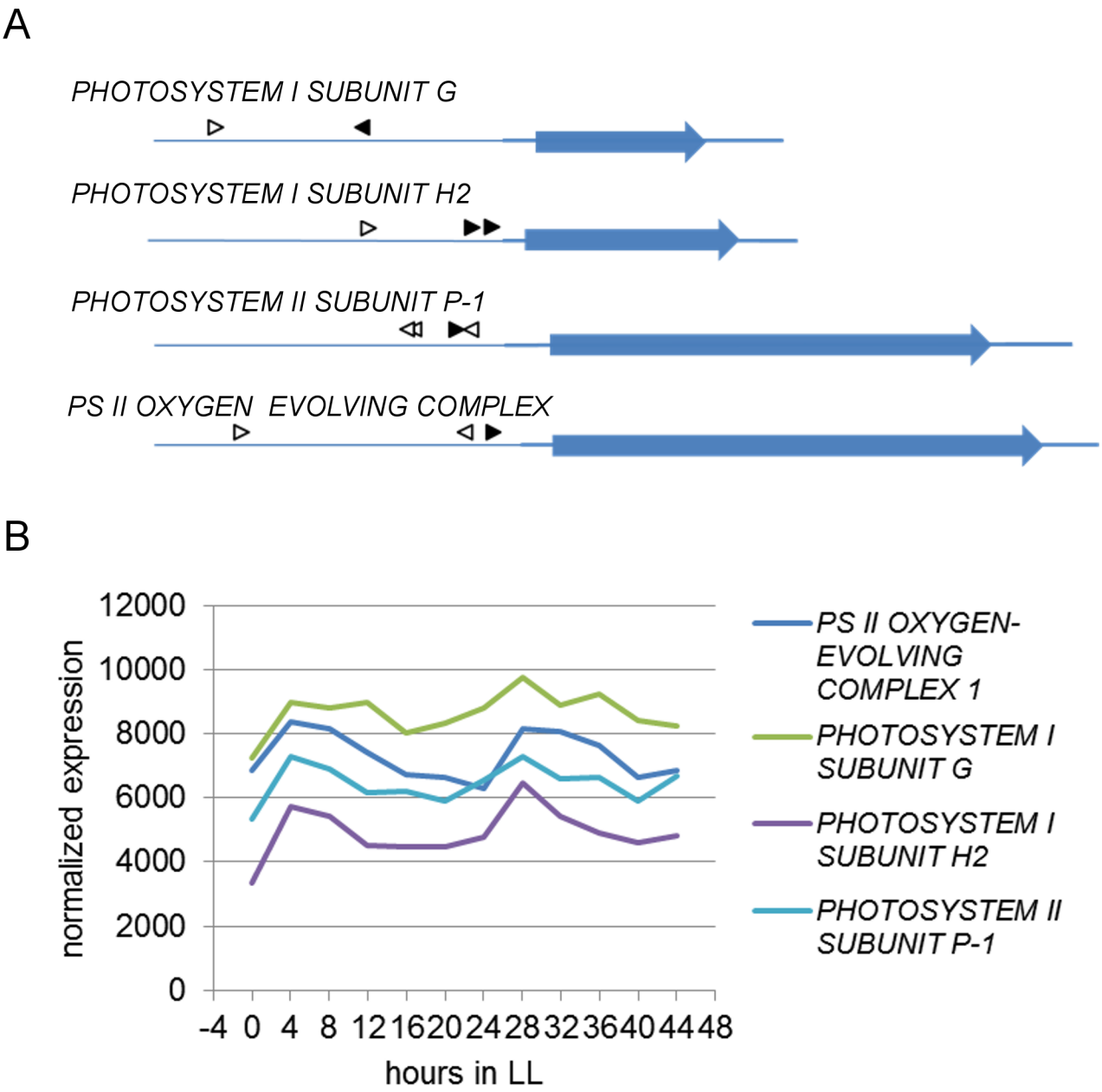

Fig 7. Position weight matrix of nucleotides in the DPBF1&2 element that correlate with a shift in the DEM when combined with the evening element. The size of the letters at each position indicate which bases lead to a decrease in the performance of the CRE pair when mutated prior to the EDCC analysis. That means that changing the adenine on position 5 to any other nucleotide led to a decreased correlation of the mutated CRE pair with time point specific gene expression in almost all cases.

https://doi.org/10.1371/journal.pone.0190421.g007

with the genes that contain the 21 candidate CRMs [14]. Amongst the biological processes, the sequence combinations no. 3, 8, 9, 11 and 16 were most interesting, as they included processes that are known to be under the control of the circadian clock, i.e. shoot morphogenesis, photosynthesis, the regulation of defense response and the response to light stimuli (Table 3). Interestingly, six of the 21 combinations were enriched in the GO term chloroplast, i.e. the gene products of genes containing these CRE are more often located in the chloroplast than expected.

Table 3. Gene ontology (GO) analysis of pairwise CRE combinations.

no.	sequence combination	GO Biological Process	GO Cellular Component
1	AAAATATCT,ATGGCNNC	N/A	plastid, chloroplast, intracellular part, plastid part
2	AAAATATCT,CCAGTG	metabolic process, response to cadmium ion, response to inorganic substance	N/A
3	AAAATATCT,GAANGAGA	shoot morphogenesis, regulation of cellular process	N/A
4	AATNCCNC,AAAATATCT	N/A	N/A
5	ACACATG,AAAATATCT	N/A	N/A
6	ACACCGG,AAGNGTNG	response to stress, glycoside metabolic process,	N/A
7	ACANTACN,ATCCAACC	N/A	N/A
8	ACANTACN,MCACGTGGC	photosynthesis, cystein metabolic process	chloroplast thylakoid membrane
9	AGNGATAN,MCACGTGGC	regulation of defence response	chloroplast part
10	ANCACATG,AAAATATCT	N/A	N/A
11	ATACGTGT,TAACAAA	response to light stimulus, organic acid biosynthetic process	N/A
12	ATGNTTCA,ACGTGGC	cellular carbohydrate metabolic process	N/A
13	CATGCATG,AGNAACAA	N/A	N/A
14	CATGCATG,NGCNTGAA	N/A	cell wall; external encapsulating structure
15	CCNNCACN,GTGATCAC	cellular protein catabolic process	chloroplast thylakoid membrane
16	CNANAGAA,GACGTGTA	response to external stimulus, photosynthesis light reaction, alcohol metabolic process	chloroplast stroma
17	CTCATTTN,AGATCCAA	catalytic activity	N/A
18	GACGTGTA,CNNACANC	cellular protein complex assembly, photosynthesis light reaction, response to external stimulus, cellular amino acid biosynthetic process	chloroplast part
19	TCNTNAGA,CAAAACGC	N/A	N/A
20	TGTCACA,TGAGTCA	ubiquitin-dependent protein catabolic process; ligase activity	N/A
21	TGTGNGNA,TAGTGGAT	N/A	plasma membrane

https://doi.org/10.1371/journal.pone.0190421.t003

Comparison with other approaches

There are few approaches that work similarly to EDCC and CNG. However, circadian gene expression has been subject of earlier studies on CREs and CRMs. In an earlier analysis, Ding and colleagues used a frequent mining pattern [78] based approach to identify sequence combinations that frequently co-occur in Arabidopsis and poplar promoters [79]. We compared the 21 combinations we found to correlate with a shift in the DEM of circadianly expressed genes and compared these with the combinations which were found by Ding and colleagues. Here, we found that 4 out of 21 CRMs are over-represented in Arabidopsis and poplar promoters (Table 4). Note that Ding and colleagues only used CREs from the PLACE database

Table 4. Overlap between EDCC analysis and combinations found in an earlier analysis.

Combination found in this study	Combination found by Ding et al.
MCACGTGGC/ACANTACN	MCACGTGGC/CcaTACatt
CATGCATG/AGNAACAA	CATGCATG/gctaAACAAt
CCNCACN/GTGATCAC	CCnnnnnnnnnnncCACg/GTGATCAC CAAACACC/GTGATCAC
TCNTNAGA/CAAAACGC	TCaTcttctt/CAAAACGC

Partially overlapping CREs contain large and small letters. The large letters indicate nucleotides that are identic between the CRE analysed by EDCC and the CRE analysed by Ding et al [79].

https://doi.org/10.1371/journal.pone.0190421.t004

[80], which is a subset of the AtCOEcis database that we used for this study [5]. Thus, it is likely that more combinations that we found in our analysis are over-represented in Arabidopsis promoters.

Another study found 10 CREs that correlated with diurnal and circadian gene expression in Arabidopsis [32]. For this they used MEME [30] but as the analysis with MEME is very computation intensive, the authors had to use a supercomputer [32]. We analyzed the 10 CREs they found using EDCC, and identified only CCACGTG as a candidate. EDCC determined that the motif deviates from the background at ZT0-ZT4 (at the start of the day), whereas the authors of the previous study only identified two sets of genes that contained this motif but displaying different expression patterns.

Both comparisons indicate that EDCC may be more conservative than other approaches to correlate gene expression with presence of CREs.

Analyzing positional attributes of candidate CRE combinations

EDCC determines three positional features between CREs: Over-representation of specific distances between two CREs, the distance of the closest of two CREs to the TSS, and a specific order of the two CREs in respect to the TSS. Depending on the number of identified 'candidate' CRE pairs, this leads to a large number of positional features that need to be evaluated by the user. To prevent user-bias, we introduced a neural network generator that categorizes the positional features and allows for unbiased scoring of the data: CNG.

CNG is able to classify a large amount of CRE pairs at once by using two-class neural networks. We used the 21 candidate CRE pairs that were identified in the previous EDCC analysis to perform the CNG analysis (S3 Table). CNG was run eight times resulting in 7.125 networks, respectively.

One exemplary CNG network includes eight CRE pairs, of which six showed significant overrepresentation of a specific order between the two CREs and the TSS (Fig 8A). None of the combinations showed a preference for a specific distance between the individual CREs (Fig 8B), and most combinations are positioned close to the TSS (Fig 8C). CNG summarizes the analysis of all three positional features in a scatterplot matrix, in which each point represents a specific CRE pair (Fig 9). One of the pairs that showed strong order preference and a tendency to be close to the TSS consists of a G box (MCACGTGGC) [64] and an undefined ACANTACN motif. Genes containing this CRE pair are enriched in the GO term photosynthesis. Four of the genes containing this combination belong to the photosystems I and II, respectively. These were the genes *PHOTOSYSTEM I SUBUNIT G*, *PHOTOSYSTEM I SUBUNIT H2*, *PHOTOSYSTEM II SUBUNIT P-1*, and *PS II OXYGEN-EVOLVIN G COMPLEX 1* (Fig 10A). They all exhibit their maximum expression between ZT4 and ZT8, i.e. in the middle of the subjective light phase (Fig 10B). In the promoters of these and 30 other genes, the G box motif is positioned closer to the TSS than the ACANTACN motif ($p = 3.86 \cdot 10^{-5}$).

Discussion

EDCC correctly identifies known circadian clock promoter elements

Although a plethora of programs exist that allow deciphering of the influence of *cis*-regulatory elements on gene expression, most programs are either complicated to handle or cannot be used for large data sets, especially if statistical calculations are included. For example, the analysis of more than 1.5 million pairwise CRE combinations would suffer from a large multiple comparison error, or require large computing power. Here, we introduce the EDCC and CNG programs, which allow simple and fast identification of a large number of CREs and CRMs which may influence gene expression.

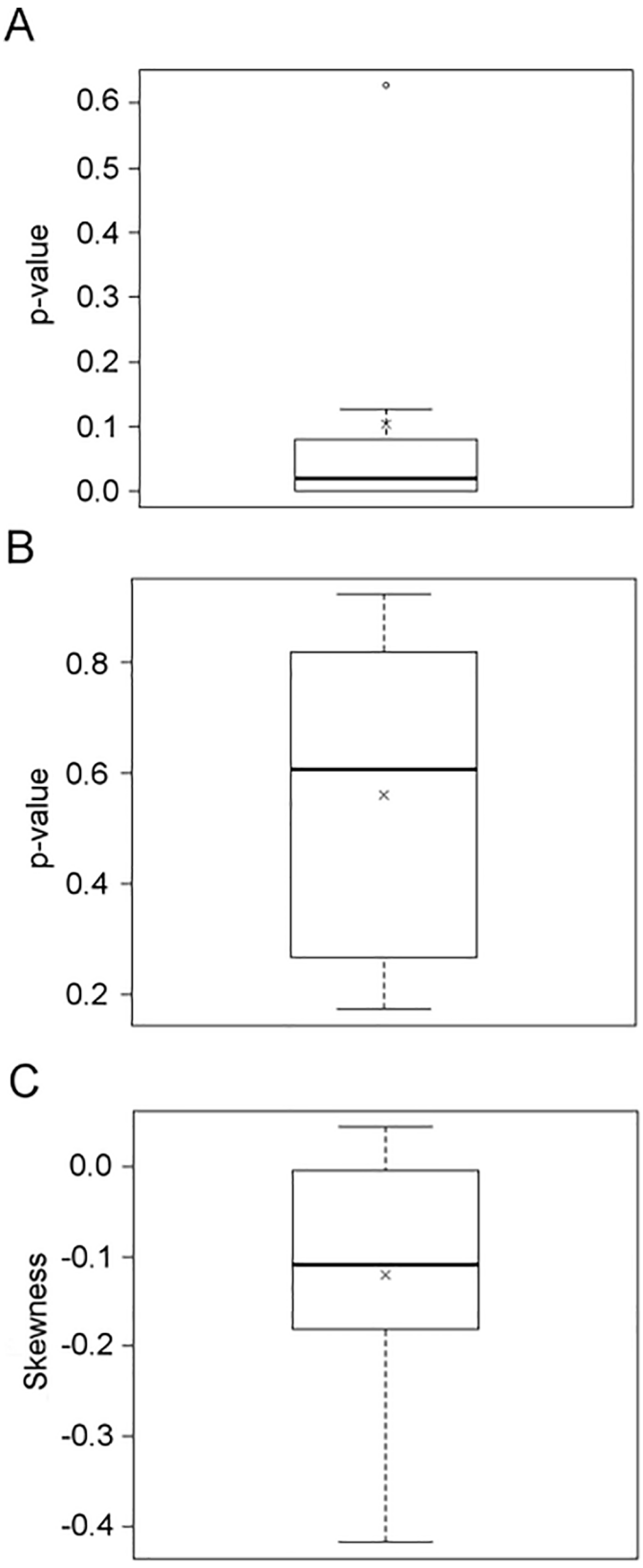

Fig 8. Representative output of the CNG analysis. A) Distribution of p-values for binomial order test. B) Distribution of p-values for distance G-test. C) Distribution of Bowley skewness analysis.

https://doi.org/10.1371/journal.pone.0190421.g008

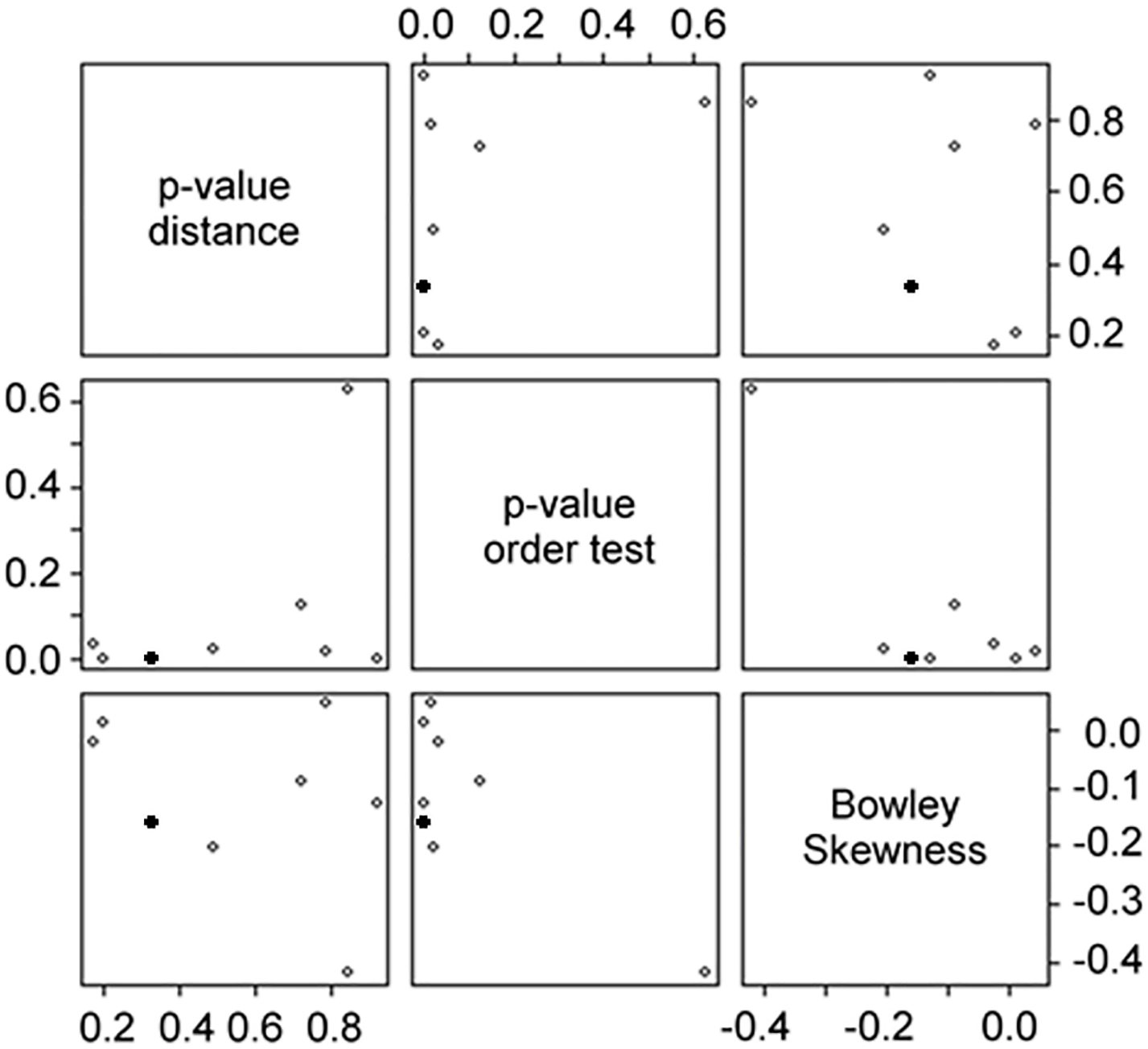

Fig 9. Scatterplot matrix summarizing the representative neural network analysis of three positional attributes. Each dot represents one CRE pair. Filled dots represent gene pairs that indicate the G box/ACANTACN pair, which is present in four photosystem genes and correlated with midday specific gene expression.

https://doi.org/10.1371/journal.pone.0190421.g009

EDCC determines whether the presence of a CRE or CRM in promoters correlates with a specific expression pattern. For this, the expression data needs to be categorized into different treatment conditions prior to the EDCC analysis. EDCC compares the DEM of genes containing queried CREs/CRMs with the background distribution. With each analysis, EDCC runs a large set of random CREs and determines their standard deviation from the background. This standard deviation serves as the threshold at which a queried CRE is marked as a candidate.

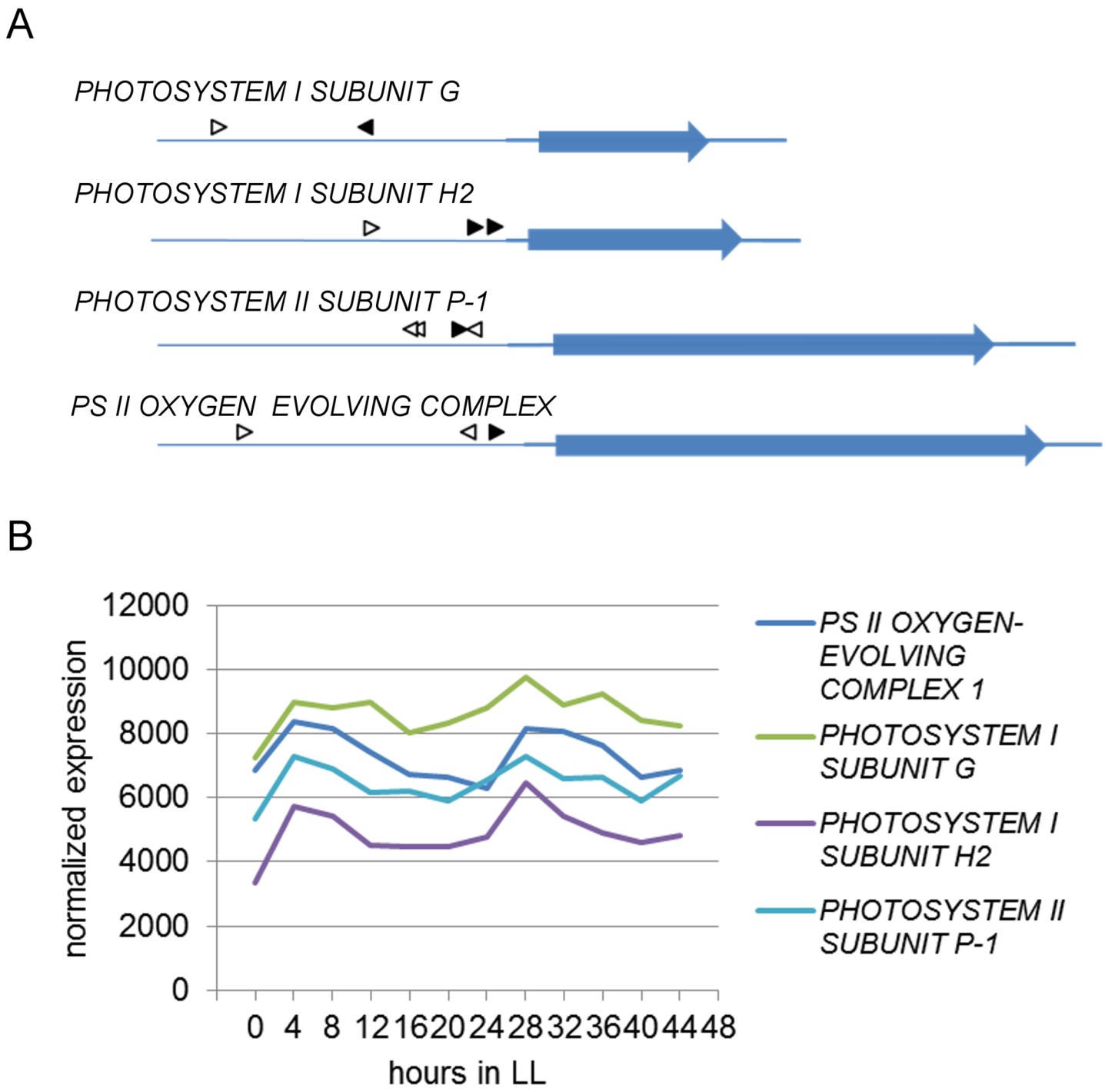

Fig 10. Positions of ACANTACN and G box motifs in photosystem subunit gene promoters and correlation with circadian gene expression. A) Positions of ACANTACN (white arrows) and the G box MCACGTGGC (black arrows) CREs in promoters of photosystem subunit genes. Blue arrows indicate CDS in 5'-3' direction (introns are ignored), thicker blue lines indicate 5' and 3' UTRs. Thin blue line represents 1000 bp upstream region of the TSS. B) Circadian gene expression of the same photosystem genes as given by DIURNAL [81].

https://doi.org/10.1371/journal.pone.0190421.g010

In our study of CREs and CRM in circadian gene expression [14] we were able to identify only two of the known CREs as candidates, i.e. the evening element [23] and one of the known three midnight elements [14]. The morning element [25], two of the midnight elements [14]

and the HUD-domain [27] were not found in the EDCC analysis. This means that i) EDCC is generally conservative and may generate false negatives, ii) the given positive controls have a small effect on circadian gene expression, and/or iii) the positive control CREs would have been discovered when using another circadian microarray experiment. As we mainly wanted to avoid discovery of false positive CREs, we were satisfied with the performance of EDCC on the positive control queries and continued to analyze all CREs given in the AtCOEcis database [5].

EDCC finds both known and unknown CREs and CRE combinations that correlate with circadian gene expression

We used EDCC to analyze 1755 CREs with circadian microarray data and to identify candidate elements that correlate with gene expression at a specific time of the day. In one of the most conservative approaches we found 21 candidate CREs, which included the evening element [23], MYB1 and MYB2 binding site motifs [52,56,65], a wound responsive element [55], a TCP binding site [60], a GATA motif [59], ABA response element binding sites [41,66], and a G box element [64]. Whereas the Myb-domain transcription factors CCA1 and LATE ELONGATED HYPOCOTYL are involved in the regulation of the core clock, the homologs MYB1 and MYB2 have not been shown to be involved in circadian gene expression yet. Furthermore, MYB1 and MYB2 were both found to influence ABA signaling and responses [82,83]. ABA is a phytohormone that is essential in plant developmental processes as well as plant stress responses. Genes that are expressed rhythmically during a day-night cycle are overrepresented among ABA responsive genes [84] and ABA response to drought is gated by the circadian clock core component TIMING OF CAB EXPRESSION 1 [48,50,85]. Conversely, ABA treatment lengthens the circadian expression period of circadian clock genes [86]. Thus, it is fitting that the EDCC analysis identifies ABA response elements as candidate CREs in the regulation of circadian gene expression. Class I TCP transcription factors have been identified to control circadian gene expression, especially via binding to the promoter of the core clock gene *CCA1* [45–47,87]. In sum, these findings point out that EDCC indeed is able to identify candidate CREs that may confer specific gene expression.

In a next step, we used EDCC to analyze over 1.5 million pairs of CREs that were created by pairing each of the 1755 CREs with each other. We found a plethora of potential CRE pairs that correlate with daytime specific gene expression. The strongest effect was seen in the co-occurrence of the evening element with the DPBF1&2 binding site motif (ACACATG). Although first defined in carrots [67], a similar site (ACACNNG) has been found in Arabidopsis, where the motif is bound by the bZIP class transcription factor ABA-INSENSITIVE 5 (ABI5) [88], again pointing out the close association of ABA signaling with the circadian clock. No indications exist as yet to what the function of this pairwise combination is, and it would be one of the first CRE pairs to study in wetlab experiments after the EDCC analysis. Some positions within a CRE are less important for its function than others, leading to annotated CREs containing ambiguity code. When mutating the evening element/ DPBF1&2 binding site motif pair, we found that all positions of the evening element were important for EDCC to define the pair as a candidate. For the DPBF1&2 binding site motif, we found several variations which allowed us to indicate specific positions that are important for its presumed function. It would be interesting to determine whether these positions are indeed important for the evening element/ DPBF1&2 binding site motif pair to confer daytime specific gene expression, however this was beyond the scope of this study. This example also highlights another potential of EDCC: the EDCC program is able to analyze CREs with ambiguity code. For this, EDCC first analyzes the component CREs (e.g. AAAGA and AAAAA when

calculating of AAARA) and then summarizes the results. EDCC would thus also be able to determine which of the component CREs correlates stronger with a specific expression pattern, allowing the identification of important positions. We have not tested this, but it would be an interesting future experiment.

When applying less restrictive conditions to the analysis of 1.5 million CRE pairs, EDCC identified more candidate CRE pairs. These often included at least one CRE that was previously found in circadian or light-responsive gene regulation, e.g. the evening element [23], a G box [89], a Z-DNA-forming sequence [90], or a dark responsive element [74]. In a previous study, Ding and colleagues used a frequent pattern mining approach to determine which CRE pairs are over-represented in Arabidopsis and poplar promoters [79]. When comparing our CRE pairs with those, we found that four CRE pairs were similar. Hence, these four combinations not only coincide often in plant promoters, they also correlate with specific peak circadian expression times of the respective genes. In summary, the EDCC program was able to not only detect CREs that are known to control circadian gene expression, further analysis also allowed to detect secondary CREs that are likely to influence circadian gene expression in combination with the previously known CREs. After validating these in wetlab approaches, it will be interesting to analyze, how they influence expression of target genes and what kinds of protein complexes bind to these.

CNG scoring of positional CRE/CRM offers an unbiased approach to analyzing large-scale EDCC outputs

EDCC not only determines interesting secondary CREs, it also calculates positional features, as CRE positions are an important feature of CRE-mediated gene control [9,10,91]. The positional features calculated are: the distance of two CREs to each other, the distance of a CRM to the TSS, and the orientation of two CREs regarding which one is closer to the TSS. To prevent user-bias, we created the CNG program, which scores these positional features using a neural network. We used the CNG program to analyze CRE pairs that were found by EDCC. In a representative network scored by CNG we found the combination of a G box element with a ACANTACN sequence. This combination was found in 34 gene promoters and correlates with gene expression in the middle of the subjective day. One of the reasons that this combination was included by CNG is that the ACANTACN element is mostly positioned 5' of the G box. We found this combination to be very prominent in the promoters of four photosystem subunit genes that are all expressed in the middle of the day. This indicates that this CRE combination indeed may affect day time specific gene expression. To our knowledge, this is the first description of this potential CRE pair and it would be interesting to validate these findings in wetlab experiments.

Possibilities of EDCC and CNG and comparison with other approaches

The EDCC and CNG analysis have certain limitations, which will be discussed here. First of all, EDCC is designed to work with gene expression data, in which each gene exhibits maximum gene expression in one expression category. Circadian data was an ideal test case, as circadianly expressed genes exhibit a defined peak in contrast to other treatments or conditions. We see possible applications of this program in deciphering regulation of organ growth processes. For example, the identity of Arabidopsis floral organs is controlled by the presence of different MADS box transcription factors, each controlling different sets of genes (for a review, see [92]). These may be identified using the EDCC and CNG programs. In principle, any expression data that follows an OR logic, is suitable to be analyzed with the programs presented here. Furthermore, we have limited the analysis of CRMs to pairs of CREs. EDCC is in

principle able to analyze combinations with more than two individual CREs, however the determination of positional features would not be possible yet. For example, the order of the two CREs in relation to the TSS is calculated using a binomial order test. A variation of EDCC with a multinomial test would be able to conduct the analysis. Also, the number of positional features that are calculated by EDCC can be increased. Such possible features are e.g. the number of repeats of a CRE, non-traditional positions like introns or downstream sequences, and the orientation of CREs, amongst others. EDCC is already able to include orientations of the CREs, but for this study we allowed CREs to appear in all possible orientations.

Whereas many programs were developed to identify CREs and CRMs in data sets, we designed a program that works with a user-identified list of CREs and CRMs. The simple approach of EDCC to correlate CREs/CRMs with gene expression data is reliable without being hindered by multiple comparison errors or by a lack in computing infrastructure. EDCC and CNG both run on PCs using free software (R and Phython), allowing non-experts fast identification of candidate CREs that may confer specific expression under different treatments and conditions.

Ultimately, the EDCC analysis provides a starting point for further in depth analysis of CRMs in gene expression. We showed that EDCC correctly identifies candidate CREs that are known for their effect on circadian gene expression. EDCC further identified candidate single CREs and CRE pairs that were not known to affect circadian gene expression. Some of the pairs are found in specific positions upstream of the respective genes. In the future, wetlab experiments need to show whether the presence and positions of these CREs are also functionally linked to circadian gene expression.

Material and methods

Exploration of Distinctive CREs and CRMs (EDCC)

EDCC compares the expression of genes containing a queried CRE with the background distribution of all genes that are affected by specific treatments or conditions. The CNG program scores the positional features that EDCC determines for candidate CRE pairs, avoiding user bias. Both EDCC and CNG are available for download under the link https://sourceforge.net/projects/edcc/. A manual is given in S1 File.

EDCC and CNG both provide graphical user interfaces (GUI). Additionally, EDCC provides an additional command line interface. The application is licensed under Apache License Version 2.0. EDCC is written in Python 3 and CGN in Python 3 and R, which makes them compatible with Microsoft Windows, macOS and Unix-like systems.

EDCC allows combining multiple CREs of interest in one query, by using the separator (,). Combinations of two CREs are further analyzed in respect to their positional attributes. The EDCC/CNG programs are able to include complementary and inversed sequences to the query CREs when specified by the user. All combined queries are split into single CREs before being validity checked, expanded and matched against the selected database (S3 Fig). Expansion means that query CREs that contain ambiguity code are broken down into their component CREs (e.g. AAAGCC and AAAACC in case of a AAARCC query). *K*-mer based indexing is used to maintain a high speed of the analysis. Peak expression times of promoters that match with the queries are extracted from an expression database (see below). If the initial query consisted of multiple CREs and was therefore split prior to the analysis, the results of all CRE are combined.

EDCC identifies whether a given query correlates with a DEM that differs from the background. The background contains all genes that are differentially expressed under the experimental conditions. The threshold is calculated using a user-determined number of random

CREs (by default 100). EDCC calculates a DEM for each random CRE and determines a standard deviation for each expression category based on these DEMs. One standard deviation is the minimum threshold that is recommended in the EDCC analysis. As the random background is calculated in each run of EDCC, each run may produce slightly different results. To eliminate randomly occurring extreme variations, a default total of 100 backgrounds are produced per run and a query is termed 'candidate' when it deviates from the majority of the runs, respectively.

CREs that only occur in few promoters may exhibit distribution biases. Hence, the number of minimum matches a query has to meet is user-determined, but we do recommend using CREs that occur in at least 10 to 30 promoters. The default setting is 20 promoters.

Analysis of positional features of CRE pairs

EDCC calculates three different positional features per candidate CRE pair:

Distance test. A two-sided Kolmogorov-Smirnov test is used to determine whether two CREs prefer a specific distance towards each other. The distribution of expected distances is generated using a stochastic approach: at first, the probability that a CRE occurs in a promoter is calculated. Then for each CRE as many random numbers are generated as expected to occur in 1000 bp, which represents the length of the analyzed promoter regions. The probabilities of the CREs are subtracted from each other and the smallest absolute difference between the probabilities is taken by EDCC to determine the distance of the CRM elements in a promoter. This procedure is performed 10,000 times to calculate the distribution of expected distances.

Order test. To determine whether CRE pairs occur predominantly in a given order in relation to the TSS, a binomial test is performed with the null hypothesis that each possible order of the two given CREs occurs with the same probability. The formula for this test is given below:

$$p(X) = \frac{n!}{(n-X)!X!} \cdot (p)^X \cdot (q)^{n-X})$$

Here, p and q are equal to 0.5, n is the total of pairwise occurrences and X is the number of occurrences of one possible order.

Bowley skewness of CRM positions. We defined the position of a CRM as the smallest distance of its constituent CREs to the TSS. As CRMs are predominantly positioned near the TSS, we expected that the distribution of the single positions of a CRM in the affected promoters is left-skewed. The skewness is calculated with Bowley's coefficient of skewness. The value range lies between -1 and 1. Positive values indicate a right-skewed distribution; negative values a left-skewed distribution. The skewness coefficient is calculated as follows:

$$S = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

where Q1 is the first, Q2 the second and Q3 the third quartile of the position's distribution.

CRM Network Generator (CNG)

CNG uses an artificial two-class neural network to categorize and weigh positional features that were determined by EDCC (S2 Fig), thus precluding bias when manually assessing the EDCC output [33]. Via the CNG GUI, the user can change most parameters of the neural network generation. All neural networks created by CNG are feedforward networks that take the numeric results from the three statistical tests of EDCC as input. The networks consist of a

neuron with a sigmoid activation function for the input and a Heaviside activation function for the output [93].

The CNG is trained with three types of neural network training data: the output of EDCC (i.e. the sequences of interest), sequences that exhibit p-values of 1 and Bowley skewness of -1, 0, and 1 (the negative control), and random sequences. The random sequences are used to ensure that the network categorization does not become too broad within the numeric data range of the positive sequences, as very broad categorizations could simply include all positive sequences without performing categorizations based on their properties.

The network training follows an evolutionary approach in order to get a more start values-independent categorization than with classical backpropagation [79,94]. In order to be more controllable in respect to the small number of inputs and outputs, the CNG training method only evolves the weights and biases of a neural network, but not its structure. Each evolutionary training process of networks is separated in "cycles", which are separated again in "rounds". The currently trained networks are scored each round (see below). Afterwards, the networks are sorted according to their score, and the best rated networks are selected for the next round. Mutated variants of the currently best rated networks and new randomly created networks are generated and scored together with the best networks of the previous round. The mutations can be either single incremental or disruptive changes of the biases or the weights, or crosses of two of the best rated networks. One CNG cycle ends when the score of the best network does not increase for a user-defined number of rounds. The best network of the last round of a cycle is saved internally and can be visualized later.

In the next cycle, the newly generated networks are forced to include positive sequences that have not been categorized before to ensure that the new networks are not identical to previous ones, and to increase the total number of categorized sequences. The CNG analysis ends when all positive sequences were categorized at least once in a generated network.

Scoring of neural networks

The score of a network depends on the quantity of positive, negative and random sequences that are included in the network. If a network includes one of the negative sequences in its categorization, it gets the lowest score. If this is not the case, the network's score is calculated by dividing the number of positive sequences with the number of random sequences. If two or more networks have the same rating, the networks including more sequences are rated higher to avoid too narrow categories. The user can change most of the training process settings via the CNG user interface. This includes the fixed number of neurons of the hidden layer as well as all other numeric parameters to set or change the bias and the weights of the hidden layer's and the output layer's neurons during the training process. Each ongoing or finished training process, as well as each generated neural network, can be saved in a binary format. The binary files of training processes can be reloaded by the CNG.

Visual output of the CNG

The CNG user interface shows the results of an ongoing or finished neural network training process. These are documented in HTML files which include textual information and plots. The subsequently generated index file is the starting point for the visualization. It shows all settings of the training process as well as an overview of all generated networks. Each generated network is also described and visualized in its own HTML file. The binary files of the training process and the single networks are automatically created with the HTML report. The index HTML file also shows the differences of the categorized sequences of the networks. This is done by generating a distance matrix of all generated networks. A value of 1 means that no

sequences can be found in both categories, whereas a value of 0 means that all sequences of the smaller category are included in the larger category. This distance matrix is visualized as a 2D plot using the "symmetric SMACOF" multidimensional scaling method [95]. Additionally, the index file also shows whether a correlation between each of the input data of the categorized sequences of each particular network was detected using Spearman's correlation coefficient. The HTML files describing the single networks show a scatterplot of the input data, boxplots of the single data, the categorized sequences as well as a table containing the all included CREs and the genes in which they occur. The gene identifiers are provided in separate text files. These gene identifier lists allow subsequent analyses, such as GO analyses.

Experimental data

The programs were tested using published data of a circadian microarray experiment (E-MEXP-1304) [14]. In this experiment Arabidopsis seedlings were grown for 9 days in a 12 hours light/12 hours dark regime and subsequently transferred to continuous light. Samples were taken every four hours for 48 hours after transfer to continuous light. We analyzed the continuous light experiment with the ARSER package and a significance cut-off of q = 0.05 [39]. Genes that exhibited circadian gene expression were categorized according to their peak expression time (ZT0-ZT4, ZT4-ZT8, ZT8-ZT12, ZT12-ZT16, ZT16-ZT20, ZT20-ZT0). Arabidopsis sequence data including 1000 bp upstream of the TSS for all coding and non-coding genes represented in TAIR10 was used to query for CREs, respectively [96]. We used 1755 CREs as given in the AtCOEcis database to test the programs [5]. These 1755 CREs include known motifs from PLACE [80] and AGRIS [4] and *de novo* motifs that were identified by homology between Arabidopsis and poplar [5]. Based on this collection we created a dataset in which each CRE was paired with a second CRE (disregarding the order), resulting in a query dataset of 1,540,890 CRE combinations.

Supporting information

S1 Fig. Schematic representation of EDCC analysis. Legend indicates input data, processes, and output of the EDCC analysis.
(TIF)

S2 Fig. Schematic representations of two-class neural networks generated by CNG. Neurons are shown as circles, numeric inputs as rectangles. All of these networks take the Bowley Skewness of a CRM's positions, the p value of the distance test of the CRM and the p value of the order test of the CRM as numeric input. The activation function of the n neurons in the sole hidden layer of these networks is the sigmoid function $(t) = \frac{1}{1+e^{-t}}$. For each of these neurons, the parameter for the activation function is the sum of the neuron's bias value with t. t is the sum of the weighted numeric inputs. Each hidden layer neuron has its own weight w for each numeric input. The output layer consists of one neuron. This neuron has the Heaviside function h as activation function. As parameter for h, the sum of the neuron's bias b and t is used. In this case, t is the sum of the weighted outputs of the hidden layer's neurons.
(TIF)

S3 Fig. Expansion of a CRE by EDCC. Handling of ambiguity code by EDCC. First, the ambiguity code is unscrambled into the component four bases. In the second step, complementary and inverse CREs are determined. Then, EDCC analysis is performed for each component CRE and the results united.
(TIF)

S1 Table. Candidate single CREs identified by EDCC. 1755 CREs [5] were analyzed for correlation with a shift in circadian peak expression time. The table depicts all CREs that were found as candidates in five runs and occurred at least 10 times in Arabidopsis promoters.
(PDF)

S2 Table. Candidate single CREs under conservative settings. EDCC analysis of 1755 CREs for correlation with a shift in circadian peak expression time in Arabidopsis. The number of minimum occurrences was increased to 15, 20, and 30, respectively. Given are all CREs that were found as candidates in five runs.
(PDF)

S3 Table. Candidate CRE pairs that were used for CNG analysis. Given are 21 CRE pairs that have been found to correlate with a shift in peak expression time of circadianly expressed genes in Arabidopsis. All listed pairs occurred in at least 30 promoters and deviated from the background by at least five standard deviations in all five EDCC runs.
(PDF)

S1 File. EDCC and CNG manual. The manual is also available as.html file under https://sourceforge.net/projects/edcc/files/edcc_cng.zip/download.
(DOCX)

Acknowledgments

We thank Martin Lewinski for useful discussions about the methods used in this paper.

Author Contributions

Conceptualization: Pavlos Stephanos Bekiaris, Tobias Tekath, Selahattin Danisman.

Data curation: Selahattin Danisman.

Formal analysis: Tobias Tekath.

Investigation: Pavlos Stephanos Bekiaris.

Methodology: Pavlos Stephanos Bekiaris, Tobias Tekath, Selahattin Danisman.

Project administration: Dorothee Staiger, Selahattin Danisman.

Software: Pavlos Stephanos Bekiaris, Tobias Tekath.

Supervision: Selahattin Danisman.

Writing – original draft: Selahattin Danisman.

Writing – review & editing: Pavlos Stephanos Bekiaris, Tobias Tekath, Dorothee Staiger, Selahattin Danisman.

References

1. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 2013; 23: 777–788. https://doi.org/10.1101/gr.152140.112 PMID: 23482648
2. Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res. 2011; 21: 775–789. https://doi.org/10.1101/gr.110254.110 PMID: 21372179
3. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, Santos D, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription

regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 2014; 42: D161–D166. https://doi.org/10.1093/nar/gkt1015 PMID: 24170807

4. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. Plant Physiol. 2006; 140: 818–829. https://doi.org/10.1104/pp.105.072280 PMID: 16524982
5. Vandepoele K, Quimbaya M, Casneuf T, Veylder LD, Peer YV de. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. Plant Physiol. 2009; 150: 535–546. https://doi.org/10.1104/pp.109.136028 PMID: 19357200
6. Howard ML, Davidson EH. cis-Regulatory control circuits in development. Dev Biol. 2004; 271: 109–118. https://doi.org/10.1016/j.ydbio.2004.03.031 PMID: 15196954
7. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 2010; 20: 565–577. https://doi.org/10.1101/gr.104471.109 PMID: 20363979
8. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001; 29: 153–159. https://doi.org/10.1038/ng724 PMID: 11547334
9. Vardhanabhuti S, Wang J, Hannenhalli S. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res. 2007; 35: 3203–3213. https://doi.org/10.1093/nar/gkm201 PMID: 17452354
10. Kulkarni MM, Arnosti DN. cis-Regulatory Logic of Short-Range Transcriptional Repression in Drosophila melanogaster. Mol Cell Biol. 2005; 25: 3411–3420. https://doi.org/10.1128/MCB.25.9.3411-3420.2005 PMID: 15831448
11. Vandepoele K, Casneuf T, Van de Peer Y. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. Genome Biol. 2006; 7: R103. https://doi.org/10.1186/gb-2006-7-11-r103 PMID: 17090307
12. Peters B, Casey J, Aidley J, Zohrab S, Borg M, Twell D, et al. A cis-regulatory module in the transcription factor DUO1 promoter. Plant Physiol. 2016; pp.01192.2016. https://doi.org/10.1104/pp.16.01192
13. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, et al. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci. 2011; 108: 14992–14997. https://doi.org/10.1073/pnas.1103202108 PMID: 21849619
14. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network Discovery Pipeline Elucidates Conserved Time-of-Day–Specific cis-Regulatory Modules. PLoS Genet. 2008; 4: e14. https://doi.org/10.1371/journal.pgen.0040014 PMID: 18248097
15. Michael TP, McClung CR. Enhancer Trapping Reveals Widespread Circadian Clock Transcriptional Control in Arabidopsis. Plant Physiol. 2003; 132: 629–639. https://doi.org/10.1104/pp.021006 PMID: 12805593
16. Cumming B, Wagner E. Rhythmic Processes in Plants. Annu Rev Plant Physiol. 1968; 19: 381–416. https://doi.org/10.1146/annurev.pp.19.060168.002121
17. Graf A, Schlereth A, Stitt M, Smith AM. Circadian control of carbohydrate availability for growth in Arabidopsis plants at night. Proc Natl Acad Sci. 2010; 107: 9458–9463. https://doi.org/10.1073/pnas.0914299107 PMID: 20439704
18. Sellaro R, Pacín M, Casal JJ. Diurnal Dependence of Growth Responses to Shade in Arabidopsis: Role of Hormone, Clock, and Light Signaling. Mol Plant. 2012; 5: 619–628. https://doi.org/10.1093/mp/ssr122 PMID: 22311777
19. Johansson M, Staiger D. SRR1 is essential to repress flowering in non-inductive conditions in Arabidopsis thaliana. J Exp Bot. 2014; 65: 5811–5822. https://doi.org/10.1093/jxb/eru317 PMID: 25129129
20. Streitner C, Danisman S, Wehrle F, Schöning JC, Alfano JR, Staiger D. The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana. Plant J. 2008; 56: 239–250. https://doi.org/10.1111/j.1365-313X.2008.03591.x PMID: 18573194
21. Zhang C, Xie Q, Anderson RG, Ng G, Seitz NC, Peterson T, et al. Crosstalk between the Circadian Clock and Innate Immunity in Arabidopsis. PLoS Pathog. 2013; 9: e10033770.
22. Korneli C, Danisman S, Staiger D. Differential Control Of Pre-Invasive And Post-Invasive Antibacterial Defense By The Arabidopsis Circadian Clock. Plant Cell Physiol. 2014; 55: 1613–1622. https://doi.org/10.1093/pcp/pcu092 PMID: 24974385
23. Harmer SL, Hogenesch JB, Straume M, Chang H-S, Han B, Zhu T, et al. Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. Science. 2000; 290: 2110–2113. https://doi.org/10.1126/science.290.5499.2110 PMID: 11118138
24. Staiger D, Apel K. Circadian clock-regulated expression of an RNA-binding protein in Arabidopsis: characterisation of a minimal promoter element. Mol Gen Genet. 1999; 261: 811–819. https://doi.org/10.1007/s004380050025 PMID: 10394919

25. Harmer SL, Kay SA. Positive and Negative Factors Confer Phase-Specific Circadian Regulation of Transcription in Arabidopsis. Plant Cell. 2005; 17: 1926–1940. https://doi.org/10.1105/tpc.105.033035 PMID: 15923346

26. Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua N-H, Sakakibara H. PSEUDO-RESPONSE REGULATORS 9, 7, and 5 Are Transcriptional Repressors in the Arabidopsis Circadian Clock. Plant Cell. 2010; 22: 594–605. https://doi.org/10.1105/tpc.109.072892 PMID: 20233950

27. Michael TP, Breton G, Hazen SP, Priest H, Mockler TC, Kay SA, et al. A Morning-Specific Phytohormone Gene Expression Program underlying Rhythmic Plant Growth. PLoS Biol. 2008; 6: e225. https://doi.org/10.1371/journal.pbio.0060225 PMID: 18798691

28. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet. 2001; 27: 167–174. https://doi.org/10.1038/84792 PMID: 11175784

29. Bailey TL, Noble WS. Searching for statistically significant regulatory modules. Bioinformatics. 2003; 19: ii16–ii25. https://doi.org/10.1093/bioinformatics/btg1054 PMID: 14534166

30. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37: W202–W208. https://doi.org/10.1093/nar/gkp335 PMID: 19458158

31. Bolívar JC, Machens F, Brill Y, Romanov A, Bülow L, Hehl R. "In silico expression analysis", a novel PathoPlant web tool to identify abiotic and biotic stress conditions associated with specific cis -regulatory sequences. Database. 2014;2014. https://doi.org/10.1093/database/bau030 PMID: 24727366

32. Janaki C, Joshi RR. Motif Detection in Arabidopsis: Correlation with Gene Expression Data. In Silico Biol. 2004; 4: 149–161. PMID: 15107020

33. Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory modules. Bioinformatics. 2003; 19: i292–i301. https://doi.org/10.1093/bioinformatics/btg1040 PMID: 12855472

34. Johansson Ö, Alkema W, Wasserman WW, Lagergren J. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics. 2003; 19: i169–i176. https://doi.org/10.1093/bioinformatics/btg1021 PMID: 12855453

35. Hu J, Hu H, Li X. MOPAT: a graph-based method to predict recurrent cis -regulatory modules from known motifs. Nucleic Acids Res. 2008; 36: 4488–4497. https://doi.org/10.1093/nar/gkn407 PMID: 18606616

36. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007; 8: R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271

37. Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefèbvre C, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 2006; 16: 656–668. https://doi.org/10.1101/gr.4866006 PMID: 16606704

38. Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics. 2010; 26: 1579–1586. https://doi.org/10.1093/bioinformatics/btq248 PMID: 20453004

39. Yang R, Su Z. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. Bioinformatics. 2010; 26: i168–i174. https://doi.org/10.1093/bioinformatics/btq189 PMID: 20529902

40. Bonett DG. Confidence interval for a coefficient of quartile variation. Comput Stat Data Anal. 2006; 50: 2953–2957. https://doi.org/10.1016/j.csda.2005.05.007

41. Shinozaki K, Yamaguchi-Shinozaki K. Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. Curr Opin Plant Biol. 2000; 3: 217–223. https://doi.org/10.1016/S1369-5266(00)80068-0 PMID: 10837265

42. Hartmann U, Sagasser M, Mehrtens F, Stracke R, Weisshaar B. 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 Mol Biol. 2005; 57: 155–171. https://doi.org/10.1007/s11103-004-6910-0 PMID: 15821875

43. Terzaghi W, Cashmore AR. Light-Regulated Transcription. Annu Rev Plant Physiol Plant Mol Biol. 1995; 46: 445–474. https://doi.org/10.1146/annurev.pp.46.060195.002305

44. Staiger D, Becker F, Schell J, Koncz C, Palme K. Purification of tobacco nuclear proteins binding to a CACGTG motif of the chalcone synthase promoter by DNA affinity chromatography. Eur J Biochem. 1991; 199: 519–527. https://doi.org/10.1111/j.1432-1033.1991.tb16150.x PMID: 1714388

45. Pruneda-Paz JL, Breton G, Para A, Kay SA. A Functional Genomics Approach Reveals CHE as a Component of the Arabidopsis Circadian Clock. Science. 2009; 323: 1481–1485. https://doi.org/10.1126/science.1167206 PMID: 19286557

46. Pruneda-Paz JL, Breton G, Nagel DH, Kang SE, Bonaldi K, Doherty CJ, et al. A Genome-Scale Resource for the Functional Characterization of Arabidopsis Transcription Factors. Cell Rep. 2014; 8: 622–632. https://doi.org/10.1016/j.celrep.2014.06.033 PMID: 25043187

47. Wu J-F, Tsai H-L, Joanito I, Wu Y-C, Chang C-W, Li Y-H, et al. LWD–TCP complex activates the morning gene CCA1 in Arabidopsis. Nat Commun. 2016; 7: 13181. https://doi.org/10.1038/ncomms13181 PMID: 27734958

48. Legnaioli T, Cuevas J, Mas P. TOC1 functions as a molecular switch connecting the circadian clock with plant responses to drought. EMBO J. 2009; 28: 3745–3757. https://doi.org/10.1038/emboj.2009.297 PMID: 19816401

49. Hermans C, Vuylsteke M, Coppens F, Craciun A, Inzé D, Verbruggen N. Early transcriptomic changes induced by magnesium deficiency in Arabidopsis thaliana reveal the alteration of circadian clock gene expression in roots and the triggering of abscisic acid-responsive genes. New Phytol. 2010; 187: 119–131. https://doi.org/10.1111/j.1469-8137.2010.03258.x PMID: 20406411

50. Seung D, Risopatron JPM, Jones BJ, Marc J. Circadian clock-dependent gating in ABA signalling networks. Protoplasma. 2012; 249: 445–457. https://doi.org/10.1007/s00709-011-0304-3 PMID: 21773710

51. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010; 38: W64–W70. https://doi.org/10.1093/nar/gkq310 PMID: 20435677

52. Sablowski RW, Moyano E, Culianez-Macia FA, Schuch W, Martin C, Bevan M. A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. EMBO J. 1994; 13: 128–137. PMID: 8306956

53. Castresana C, Garcia-Luque I, Alonso E, Malik VS, Cashmore AR. Both positive and negative regulatory elements mediate expression of a photoregulated CAB gene from Nicotiana plumbaginifolia. EMBO J. 1988; 7: 1929–1936. PMID: 2901343

54. Ueda T, Pichersky E, Malik VS, Cashmore AR. Level of expression of the tomato rbcS-3A gene is modulated by a far upstream promoter element in a developmentally regulated manner. Plant Cell. 1989; 1: 217–227. https://doi.org/10.1105/tpc.1.2.217 PMID: 2535544

55. Palm CJ, Costa MA, An G, Ryan CA. Wound-inducible nuclear protein binds DNA fragments that regulate a proteinase inhibitor II gene from potato. Proc Natl Acad Sci U S A. 1990; 87: 603–607. PMID: 2405385

56. Menkens AE, Cashmore AR. Isolation and characterization of a fourth Arabidopsis thaliana G-box-binding factor, which has similarities to Fos oncoprotein. Proc Natl Acad Sci U S A. 1994; 91: 2522–2526. PMID: 8146148

57. Johannesson H, Wang Y, Engström P. DNA-binding and dimerization preferences of Arabidopsis homeodomain-leucine zipper transcription factors in vitro. Plant Mol Biol. 2001; 45: 63–73. https://doi.org/10.1023/A:1006423324025 PMID: 11247607

58. Ulmasov T, Liu ZB, Hagen G, Guilfoyle TJ. Composite structure of auxin response elements. Plant Cell. 1995; 7: 1611–1623. https://doi.org/10.1105/tpc.7.10.1611 PMID: 7580254

59. Yin Y, Chen L, Beachy R. Promoter elements required for phloem-specific gene expression from the RTBV promoter in rice. Plant J. 1997; 12: 1179–1188. https://doi.org/10.1046/j.1365-313X.1997.12051179.x PMID: 9418055

60. Kosugi S, Ohashi Y. PCF1 and PCF2 Specifically Bind to cis Elements in the Rice Proliferating Cell Nuclear Antigen Gene. PLANT CELL ONLINE. 1997; 9: 1607–1619. https://doi.org/10.1105/tpc.9.9.1607 PMID: 9338963

61. Lohmer S, Maddaloni M, Motto M, Di Fonzo N, Hartings H, Salamini F, et al. The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. EMBO J. 1991; 10: 617–624. PMID: 2001677

62. Chabouté M-E, Clément B, Sekine M, Philipps G, Chaubet-Gigot N. Cell Cycle Regulation of the Tobacco Ribonucleotide Reductase Small Subunit Gene Is Mediated by E2F-like Elements. Plant Cell. 2000; 12: 1987–2000. PMID: 11041892

63. Brown JW. A catalogue of splice junction and putative branch point sequences from plant introns. Nucleic Acids Res. 1986; 14: 9549–9559. PMID: 3808952

64. Giuliano G, Pichersky E, Malik VS, Timko MP, Scolnik PA, Cashmore AR. An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. Proc Natl Acad Sci U S A. 1988; 85: 7089–7093. PMID: 2902624

65. Martin C, Paz-Ares J. MYB transcription factors in plants. Trends Genet. 1997; 13: 67–73. https://doi.org/10.1016/S0168-9525(96)10049-4 PMID: 9055608

66. Hattori T, Terada T, Hamasuna S. Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of cis-acting promoter elements required for regulation by abscisic acid and VP1. Plant J. 1995; 7: 913–925. https://doi.org/10.1046/j.1365-313X.1995.07060913.x PMID: 7599651

67. Kim SY, Chung H-J, Thomas TL. Isolation of a novel class of bZIP transcription factors that interact with ABA-responsive and embryo-specification elements in the Dc3 promoter using a modified yeast one-hybrid system. Plant J. 1997; 11: 1237–1251. https://doi.org/10.1046/j.1365-313X.1997.11061237.x PMID: 9225465

68. Wagner D, Sablowski RW, Meyerowitz EM. Transcriptional activation of APETALA1 by LEAFY. Science. 1999; 285: 582–584. PMID: 10417387

69. Yadav V, Kundu S, Chattopadhyay D, Negi P, Wei N, Deng X-W, et al. Light regulated modulation of Z-box containing promoters by photoreceptors and downstream regulatory components, COP1 and HY5, in Arabidopsis. Plant J. 2002; 31: 741–753. https://doi.org/10.1046/j.1365-313X.2002.01395.x PMID: 12220265

70. Gubler F, Kalla R, Roberts JK, Jacobsen JV. Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter. Plant Cell. 1995; 7: 1879–1891. PMID: 8535141

71. Nag R, Maity MK, DasGupta M. Dual DNA Binding Property of ABA insensitive 3 Like Factors Targeted to Promoters Responsive to ABA and Auxin. Plant Mol Biol. 2005; 59: 821–838. https://doi.org/10.1007/s11103-005-1387-z PMID: 16270233

72. Chan CS, Guo L, Shih MC. Promoter analysis of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase B subunit of Arabidopsis thaliana. Plant Mol Biol. 2001; 46: 131–141. PMID: 11442054

73. Sugimoto K, Takeda S, Hirochika H. Transcriptional activation mediated by binding of a plant GATA-type zinc finger protein AGP1 to the AG-motif (AGATCCAA) of the wound-inducible Myb gene NtMyb2. Plant J. 2003; 36: 550–564. https://doi.org/10.1046/j.1365-313X.2003.01899.x PMID: 14617085

74. Maxwell BB, Andersson CR, Poole DS, Kay SA, Chory J. HY5, Circadian Clock-Associated 1, and a cis-Element, DET1 Dark Response Element, Mediate DET1 Regulation of Chlorophyll a/b-Binding Protein 2 Expression. Plant Physiol. 2003; 133: 1565–1577. https://doi.org/10.1104/pp.103.025114 PMID: 14563928

75. Yamagata H, Yonesu K, Hirata A, Aizono Y. TGTCACA Motif Is a Novel cis-Regulatory Enhancer Element Involved in Fruit-specific Expression of thecucumisin Gene. J Biol Chem. 2002; 277: 11582–11590. https://doi.org/10.1074/jbc.M109946200 PMID: 11782472

76. Washida H, Wu C-Y, Suzuki A, Yamanouchi U, Akihama T, Harada K, et al. Identification of cis-regulatory elements required for endosperm expression of the rice storage protein glutelin gene GluB-1. Plant Mol Biol. 1999; 40: 1–12. https://doi.org/10.1023/A:1026459229671 PMID: 10394940

77. Trindade LM, Horvath BM, Bergervoet MJE, Visser RGF. Isolation of a Gene Encoding a Copper Chaperone for the Copper/Zinc Superoxide Dismutase and Characterization of Its Promoter in Potato. Plant Physiol. 2003; 133: 618–629. https://doi.org/10.1104/pp.103.025320 PMID: 12972661

78. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data Min Knowl Discov. 2007; 15: 55–86. https://doi.org/10.1007/s10618-006-0059-1

79. Ding J, Hu H, Li X. Thousands of Cis-Regulatory Sequence Combinations Are Shared by Arabidopsis and Poplar. Plant Physiol. 2012; 158: 145–155. https://doi.org/10.1104/pp.111.186080 PMID: 22058225

80. Higo K, Ugawa Y, Iwamoto M, Higo H. PLACE: A database of plant cis-acting regulatory DNA elements. Nucleic Acids Res. 1998; 26: 358–359. https://doi.org/10.1093/nar/26.1.358 PMID: 9399873

81. Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, et al. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. Cold Spring Harb Symp Quant Biol. 2007; 72: 353–363. https://doi.org/10.1101/sqb.2007.72.006 PMID: 18419293

82. Wang T, Tohge T, Ivakov A, Mueller-Roeber B, Fernie AR, Mutwil M, et al. Salt-Related MYB1 Coordinates Abscisic Acid Biosynthesis and Signaling during Salt Stress in Arabidopsis1. Plant Physiol. 2015; 169: 1027–1041. https://doi.org/10.1104/pp.15.00962 PMID: 26243618

83. Baek D, Chun HJ, Kang S, Shin G, Park SJ, Hong H, et al. A Role for Arabidopsis miR399f in Salt, Drought, and ABA Signaling. Moleucles Cells. 2015; 39: 111–118. https://doi.org/10.14348/molcells.2016.2188 PMID: 26674968

84. Mizuno T, Yamashino T. Comparative Transcriptome of Diurnally Oscillating Genes and Hormone-Responsive Genes in Arabidopsis thaliana: Insight into Circadian Clock-Controlled Daily Responses to Common Ambient Stresses in Plants. Plant Cell Physiol. 2008; 49: 481–487. https://doi.org/10.1093/pcp/pcn008 PMID: 18202002

85. Lee HG, Mas P, Seo PJ. MYB96 shapes the circadian gating of ABA signaling in Arabidopsis. Sci Rep. 2016; 6: 17754. https://doi.org/10.1038/srep17754 PMID: 26725725

86. Hanano S, Domagalska MA, Nagy F, Davis SJ. Multiple phytohormones influence distinct parameters of the plant circadian clock. Genes Cells. 2006; 11: 1381–1392. https://doi.org/10.1111/j.1365-2443.2006.01026.x PMID: 17121545

87. Giraud E, Ng S, Carrie C, Duncan O, Low J, Lee CP, et al. TCP Transcription Factors Link the Regulation of Genes Encoding Mitochondrial Proteins with the Circadian Clock in Arabidopsis thaliana. Plant Cell. 2010; 22: 3921–3934. https://doi.org/10.1105/tpc.110.074518 PMID: 21183706

88. Kim SY, Ma J, Perret P, Li Z, Thomas TL. Arabidopsis ABI5 Subfamily Members Have Distinct DNA-Binding and Transcriptional Activities. Plant Physiol. 2002; 130: 688–697. https://doi.org/10.1104/pp.003566 PMID: 12376636

89. Gendron JM, Pruneda-Paz JL, Doherty CJ, Gross AM, Kang SE, Kay SA. Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. Proc Natl Acad Sci U S A. 2012; 109: 3167–3172. https://doi.org/10.1073/pnas.1200355109 PMID: 22315425

90. Puente P, Wei N, Deng XW. Combinatorial interplay of promoter elements constitutes the minimal determinants for light and developmental control of gene expression in Arabidopsis. EMBO J. 1996; 15: 3732–3743. PMID: 8670877

91. Cai X, Hou L, Su N, Hu H, Deng M, Li X. Systematic identification of conserved motif modules in the human genome. BMC Genomics. 2010; 11: 567. https://doi.org/10.1186/1471-2164-11-567 PMID: 20946653

92. Immink RGH, Kaufmann K, Angenent GC. The "ABC" of MADS domain protein behaviour and interactions. Semin Cell Dev Biol. 2010; 21: 87–93. https://doi.org/10.1016/j.semcdb.2009.10.004 PMID: 19883778

93. Kilian J, Siegelmann HT. The Dynamic Universality of Sigmoidal Neural Networks. Inf Comput. 1996; 128: 48–56. https://doi.org/10.1006/inco.1996.0062

94. Ding S, Li H, Su C, Yu J, Jin F. Evolutionary artificial neural networks: a review. Artif Intell Rev. 2013; 39: 251–260. https://doi.org/10.1007/s10462-011-9270-6

95. de Leeuw J, Mair P. Multidimensional Scaling Using Majorization: SMACOF in R. J Stat Softw. 2009; 31. Available: https://ideas.repec.org/a/jss/jstsof/v031i03.html

96. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2011; 40: D1202–D1210. https://doi.org/10.1093/nar/gkr1090 PMID: 22140109