Rice (2008) 1:177–187
DOI 10.1007/s12284-008-9017-4

The Promoter Signatures in Rice LEA Genes Can Be Used to Build a Co-expressing LEA Gene Network

Stuart Meier · Chris Gehring · Cameron Ross MacPherson · Mandeep Kaur · Monique Maqungo · Sheela Reuben · Samson Muyanga · Ming-Der Shih · Fu-Jin Wei · Samart Wanchana · Ramil Mauleon · Aleksandar Radovanovic · Richard Bruskiewich · Tsuyoshi Tanaka · Bijayalaxmi Mohanty · Takeshi Itoh · Rod Wing · Takashi Gojobori · Takuji Sasaki · Sanjay Swarup · Yue-ie Hsing · Vladimir B. Bajic

Received: 3 July 2008 / Accepted: 31 October 2008 / Published online: 22 November 2008

Abstract Coordinated transcriptional modulation of large gene sets depends on the combinatorial use of *cis*-regulatory motifs in promoters. We postulate that promoter content similarities are diagnostic for co-expressing genes that function coherently during specific cellular responses. To find the co-expressing genes we propose an ab initio method that identifies motif families in promoters of target gene groups, map these families to the promoters of all genes in the genome, and determine the best matches of each of the target group gene promoters with all other promoters. When the method was tested in rice starting from a group of co-expressing Late Embryogenesis Abundant (LEA) genes, we obtained a promoter similarity-based network that contained candidate genes that could plausibly complement the function of LEA genes. Importantly, 73.36% of 244 genes predicted by our method were experimentally confirmed to co-express with the LEA genes in maturing rice embryos, making this methodology a promising tool for biological systems analyses.

Stuart Meier, Chris Gehring, Yue-ie Hsing, and Vladimir Bajic are the first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s12284-008-9017-4) contains supplementary material, which is available to authorized users.

S. Meier · C. R. MacPherson · M. Kaur · M. Maqungo · S. Muyanga · A. Radovanovic · B. Mohanty · V. B. Bajic (✉)
South African National Bioinformatics Institute (SANBI),
University of the Western Cape,
Cape Town, South Africa
e-mail: vlad@sanbi.ac.za
URL: www.sanbi.ac.za/people/faculty/professors/vlad-bajic/

C. Gehring
Department of Biotechnology,
University of the Western Cape,
Cape Town, South Africa

S. Reuben · S. Swarup
Department of Biological Sciences,
National University of Singapore,
Singapore, Singapore

M.-D. Shih · F.-J. Wei · Y.-i. Hsing
Institute of Plant and Microbial Biology, Academia Sinica,
Taipei, Taiwan

S. Wanchana · R. Mauleon · R. Bruskiewich
International Rice Research Institute (IRRI),
Metro Manila, Philippines

T. Tanaka · T. Itoh · T. Sasaki
National Institute of Agrobiological Sciences,
Tsukuba, Japan

R. Wing
Department of Plant Sciences, University of Arizona,
Tucson, USA

T. Gojobori
National Institute of Genetics,
Shizuoka, Japan

Keywords Transcription regulation · Co-expression · Co-regulation

Introduction

In the post-genomic sequencing era, computationally based tools are required to help decipher biologically meaningful information from the masses of sequence data generated. Computationally based homology comparisons are commonly used to infer gene functions based on similarities to previously functionally annotated genes. While extremely useful, homology comparisons are somewhat limited to the identification of 'more of the same' type of genes and fail to provide information regarding the temporal, spatial, and stimulus-specific context in which the gene is expressed and active. This problem is particularly apparent when considering within a genome large gene families which share high sequence similarity yet function within distinct cellular responses. Alternatively, although global gene expression studies, such as microarray, can reveal transcriptional responses of entire genomes in a single experiment, they only provide expression profiles at specific time points in response to a specific stimulus. Furthermore, the biological roles of many of the genes identified in large-scale expression studies are not well characterized and do not link genes to specific regulatory pathways since the expression profile can be a direct or indirect result of the treatments.

In eukaryotes, many cellular processes require the coherent participation of multiple gene products as evident by the co-expression of large sets of genes in response to specific stimuli [10, 27, 29]. Furthermore, a number of studies have shown that genes that have been confirmed to be co-expressed in response to a range of conditions have correlated functional relationships, including physical interactions between their proteins [1, 16, 17, 25, 34,]. Collectively, these studies indicate that cells possess a mechanism that coordinates the expression of genes that are involved in common functional responses.

According to the *cis*-regulatory logic [2, 6], the regulation of eukaryotic gene expression is critically dictated by the combinational presence (and effect) of regulatory motifs, or signatures, in their promoters which is necessitated by the specific binding requirements of transcription factors (TFs) [1, 2, 4, 6, 23]. Genomic sequences contain these regulatory motifs encoded mainly in promoter regions of individual genes.

We hypothesize that promoter content similarity can therefore be used to identify groups of co-expressed genes that function coherently during defined cellular processes, including changes in growth and development programs or environmental challenges. We report here a method we developed that based on promoter content similarity builds a putative transcriptional regulatory network of co-expressed genes. We use the term 'network' to define a group of genes that are linked by the fact they contain a common set of motifs in their promoters that we believe could be causative for their transcriptional regulation. The promoters of these genes would presumably bind common TFs thus connecting the genes into a putative transcriptional regulatory gene network.

We tested our method in rice using a group of late embryogenesis abundant (LEA) genes that were confirmed to be co-expressed in developing embryos. In plants, the LEA genes are believed to function in protecting cellular components during developmentally induced desiccation in embryos and during water deficit stress in vegetative tissue [9]. We, therefore, hypothesize that other genes that are determined to share the most similar promoter motif combinations with the LEA genes will function coherently with them in achieving a common cellular response, which will be manifested by their co-expression with the LEA genes. Experimental validation of our predictions shows that 73.36% of the 244 predicted genes co-express with the LEA genes. In addition, a literature analysis indicated that the function of many of the genes could plausibly complement the function of the LEA genes.

Results

Method outline

We have developed a method that builds a putative transcriptional network of co-expressed genes based on them sharing highly similar promoter contents. The network building relies on a reference target gene group (TGG) that is defined in terms of being co-expressed in response to a specific biological condition. A typical example could be a cluster of co-expressing genes identified in a microarray expression experiment. The promoters of these genes are then collectively assessed for the presence of specific signatures in the form of specific motif combinations that we believe could be causative for their transcriptional responses. The signatures identified in each of the individual promoters of the TGG are then mapped to other promoters in the genome. These signatures thus serve to identify other genes that share the most similar promoter content and thus have the greatest potential to be co-regulated with each gene of the TGG. The method generates a putative transcriptional network that contains groups of candidate genes that we predict will co-express and function coherently with the TGG in producing a common cellular response. This method extends the regulatory relationships of a TGG to other

candidate genes and thus links them to a well-defined biological response providing insights into the biological context in which the gene(s) functions.

The method described above briefly consists of the following steps (details of which, related to the implementation we made, are given in the "Methods" section):

1. Determine the target gene group based on their co-expression in a common systemic response.
2. Identify promoters for the TGG genes.
3. Identify enriched motif families in the promoters of the TGG.
4. Map identified motif families to all promoters of the genome. Overlapping of mapped motifs is allowed.
5. For each of the promoters of the TGG, search for other promoters in the genome that share the highest number of the mapped motifs with the individual TGG promoter. We hypothesize that genes associated with these identified promoters have a high probability to co-express with the genes in TGG under the same biological conditions.

Identification of TGG and construction of a putative co-expressing gene network

We tested our method on the recently sequenced rice genome and used 31 annotated LEA genes as the TGG. These LEA genes were all determined to be co-expressed in mature rice embryos as determined from massively parallel signature sequencing (MPSS) expression data (see Supplementary File 1). The Dragon Motif Builder (DMB) program was used to identify 30 enriched motif families in the promoters of the LEA genes (Table 1). For each of the motif families, the consensus motif was determined. The PATCH program of the Transfac database suite indicated that 21 of the 30 identified consensus motifs conform to known plant *cis*-elements and 19 of these contain sequences that correspond to binding sites for known plant TFs (Table 1) some of which have been shown to regulate the expression of LEA genes (Table 1 and Supplementary File 2).

The presence and abundance of the 30 motif families in the individual promoters of the TGG was used to build promoter signatures for each of the individual LEA genes. These signatures were then used to map to the most similar promoters in the genome and thus identify other genes that have the greatest potential to be co-regulated with the LEA genes. A summary of the average spatial distribution of each of the 30 identified motifs in the promoters of the predicted genes is depicted in Fig. 1. This analysis identified an additional 244 genes that shared the highest number of common motifs with each of the individual promoters of the LEA genes (see Supplementary File 3). A complete network diagram (see Supplementary File 4) was constructed to illustrate the edge relationship between the TGG and the predicted genes. Figure 2 illustrates such a relationship between a single LEA gene and its neighbors in the network.

A detailed literature search that was performed for 110 of the 244 identified genes indicated that the function of many of the genes, which possessed functional descriptions, could reasonably be linked to embryo development and water deficit stress responses (Supplementary File 5).

Experimental validation of predicted gene co-expression

Experimental validation of our method was obtained using semi-quantitative reverse transcriptase polymerase chain reaction (RT-PCR) and MPSS expression analysis to determine if the predicted genes are co-expressed with the LEA genes in maturing embryos. The results show that (based on RT-PCR and MPSS) 179 (73.36%) out of the 244 genes tested were co-expressed with the LEA genes (see Supplementary Files 6 and 7). A more detailed analysis revealed a strong positive correlation between the number of motifs shared between genes and the percentage that were co-expressed (Fig. 2). We found that 100% of the predicted genes that shared 27 or more motifs with the LEA genes were co-expressed with the LEA genes, compared to the 73.36% for the overall prediction success. This analysis thus provides compelling experimental support for our method since it illustrates an extremely high correlation coefficient between the number of shared motifs and co-expression (correlation coefficient=0.97) when we consider genes with 22 or more shared motifs.

Further, in order to test whether the proportion of our predicted genes found to be expressed in maturing embryos was significantly greater than that for the whole rice genome, we performed a global MPSS expression analysis to determine the percentage of all non-transposable element (TE) genes that are expressed in maturing rice embryos. According to TIGR v.5, there are 41,047 non-TE genes in the rice genome. Using MPSS analysis of matured rice embryos, we found that 27.99% (11,488) non-TE genes are expressed in maturing rice embryos with a TPM≥1, and 20.07% (8241) non-TE genes are expressed with a TPM≥4 (TPM stands for 'transcripts per million'). Consequently, the enrichment of the experimentally confirmed genes that co-express with LEA genes in our computationally predicted gene set, relative to those from the whole rice genome that express in maturing embryos, is characterized by the p values of 1.90e−044 (TPM≥1) and 1.37e−067 (TPM≥4). These p values represent the p values corrected for multiplicity testing (see details in "Methods"). Therefore, the successful

Table 1 Identified Consensus Promoter Motifs in Original LEA Genes and the Plant TFs That Were Predicted to Bind to Them in the PATCH Program

	Consensus motif pattern	Species/gene identifier	Position	Score	Predicted plant TF	Site binding sequence
1	GAGAAGAAG	AT$PHYA_01	2 (−)	100	CAMTA3	TCTTCT
2	GGCGCGYGG	AT$AVP1_01	2 (−)	91.7	(VOZ1&2)2, CAMTA1	ACGCGC
		RICE$ZB8_02	3 (+)(−)	91.7	CBT	CGCGCG
		AS$CBT_01	3 (+)	91.7	CBT	CGCGCG CACGCG
		MAIZE $ADH1P_01&03	5 (+)	90.0	No match	CGTGG
		DAUCE$DC3_04	3 (−)	91.7	DPBF-1, DPBF-2	CACGCG
3	CCGTCGWCC	AT$H4_05	1 (+)	100		CCGTCG
		AT$COR15A_01	3 (−)	90	ANT, CBF1, CBF2, DREB1A, ERFLP1, TSI1	CCGAC
		AT$RD29B_01	3 (−)	90	CBF1	CCGAC
		AT$COR78_01	3 (−)	90	ANT, CBF1, CBF2, DREB1A	CCGAC
		AT$COR15B_01	3 (−)	90	CBF1, CBF2, DREB1A	CCGAC
		RAPE$BN115_01	3 (−)	90	CBF17, CBF5	CCGAC
		AT$FL0521F13_01	3 (−)	91.7	DREB1A	GCCGAC
		BAR$HVA1_03	3 (−)	90	CBF1, CBF2	CCGAC
		GOSHI$LEAD113_01	3 (−)	90	DBP1	CCGAC
		AT$COR78_01	3 (−)	90	ANT, CBF1, CBF2, DREB1A	CCGAC
		AT$COR15A_03	3 (−)	91.7	DREB1A	GCCGAC
		AS$CEF1_02	3 (−)	90	CEF1	CCGAC
		GOSHI$LEAD113_01	3 (−)	90	DBP1	CCGAC
4	GCGGAGAAG	No match				
5	GCVGGGCAG	MAIZE$ADH11S_06	3 (−)	90	GCBP-1, Sp1	GCCCC
6	AACADCAAA	WHEAT$CATHB_08	1 (−), 2 (−)	90	GAMYB	TTGTT
7	AGCAGCAGC	No match				
8	MCCGACGGC	AT$COR15A_03&04	1 (+)	91.7	DREB1A	GCCCAG
		MAIZE$DHN1_01	1 (+)	91.7	DBF1, DBF2	ACCGAC
		AS$TINY2_01	1 (+)	91.7	TINY2	ACCGAC
		HELAN$HSP176_02	1 (−)	91.7	No match	GTCGGT
		AT$COR15A_01	2 (+)	100	ANT, CBF1, CBF2, DREB1A, ERFLP1, TSI1	CCGAC
		RAPE$BN115_02	2 (+)	100	CBF17, CBF5	CCGAC
		AS$DREBLP1_01	2 (+)	100	DREBLP1	CCGAC
		GOSHI$LEAD113_01	2 (+)	100	DBP1	CCGAC
		AT$H4_05	3 (−)	100	No match	CCGTCG
9	ACACATACG	No match				
10	TTCMTTTCA	DAUCE$EXT_02	1 (−)	92.86	No match	AAATGAA
		POT$KST1_01	1 (−)	90	DOF1	AAAAG
		BAR$CPI_01	3 (−)	90	PBF, SED	AAAGG
		AT$WUSCHEL_01	4 (−)	91.7	No match	TGAAAA
11	AWATTATAT	No match				
12	CGGCGSCGG	AT$HLS1_01	2 (−)	91.7	ATERF7, ERF-1,2,3,4,5, ERFLP1	GCCGCC
		TO$NP24PP_0	2 (−)	91.7	ERF-1,2,3,4	GCCGCC
		AS$GCCBOX_02	2 (−)	91.7	Pti4	GCCGCC
		AS$CEF1_01	2 (−)	91.7	CEF1	GCCGCC
		BAR$HVA1_04	3 (+)	90	CBF1	GCCGCC
		AT$H4_05	4 (−)	91.7	No match	CCGTCG
		AT$COR15A_0	5 (−)	90	ANT; CBF1,2; DREB1A, ERFLP1, TSI	CCGAC
		RAPE$BN115_01	5 (−)	90	CBF17, CBF5	CCGAC
13	CTTCTTCCT	No match				
14	AAAATAATA	SOYBN$VSPB_03	1 (−)			TATTTT

Table 1 (continued)

	Consensus motif pattern	Species/gene identifier	Position	Score	Predicted plant TF	Site binding sequence
15	AAATYGARA	AS$ARR10_17	2 (−)	90	ARR10	CGATT
16	AGAAGATCA	AT$PHYA_01	1 (−)	100	CAMTA3	TCTTCT
17	RCAGCAGCA	No match				
18	CGCGCGGCG	RICE$ZB8_02	1 (+)	100	CBT	CGCGCG
19	GTTAMATAT	AT$CAB2_03	1 (+)	90	GT-3a	GTTAC
		PV$PHS_03	2 (+)	90	No match	TTAAA
		RICE$ZB8_01	2 (−)	92.86	TBP2	TATTTAA
		MAIZE$PMS1_	3 (+)	92.86	No match	TAAATAT
20	TTGYTTAAT	WHEAT$CATHB	1 (+)	90	GAMYB	TTGTT
		AS$ARR10_18	2 (+)	90	ARR10	TGATT
		PEA$RS3A_03	3 (+)	91.67	GT-1, GT-1a, SBF-1	GGTTAA
		OAT$PHYA3_0	3 (+)	92.86	No match	GGTTAAT
		RICE$PHYA_0	3 (+)	92.86	GT-1, GT-2	GGTTAAT
		PV$PHS_03	4 (+)	91.67	No match	TTTAAT
		PV$PHS_03	4 (−)	90	No match	TTAAA
21	TGTACTCSC	TO$LAP171A_	3 (−)	100	JAMYC2	GAGTA
22	MSGATGRTG	BARL$CAB11_12	2 (−)	90	MCB1, MCB2	CATCC
23	AGCACACAT	No match				
24	CMAAAAGCT	AS$PF1_01&02	2 (−)	90	PF1	TTTTT
		POT$KST1_01	3 (+)	100	DOF1	AAAAG
25	CGGCTCGCC	No match				
26	GAATGGATG	WHEAT$CAB1_	4 (−)	100	MCB1, MCB2	ATCCA
		BARL$CAB11&12	5 (−)	100	MCB1, MCB2	CATCC
27	ATCAAGGAA	AT$ATBZIP60	2 (+)	100	No match	TCAAG
28	TGGCGCCGC	No match				
29	GCCGSGGCC	MAIZE$ADH1P&11	2 (−)	91.67	No match	CCCCGG
		MAIZE$ADH1P	3 (+)	90	No match	CGTGG
		AS$mEMBP_17	3 (−)	91.67	EmBP-1a	GCCACG
		MAIZE$ADH11	4 (−)	90	GCBP-1, Sp1	GCCCC
30	AATTTTRGT	AS$PF1_01	3 (+)	90	PF1	TTTTT

Species/gene identifier represents species and gene acronyms ($) and consecutive site number in which the identified motif is found in plant genes. *Position* indicates the position and strand within the consensus motif where the TF is predicted to bind; *score* is a measure of the match between the consensus sequence and the known binding site sequence with 100 being perfect

expression rate of 73.4% of our predicted genes is significantly higher (see *p* values) using both cut-off criteria and provides strong support for the method applied here.

Discussion

The promoter regions of eukaryotic genes contain important regulatory elements that are largely responsible for coordinating their transcriptional responses [2, 6]. We have developed a method that, based on promoter content similarity, constructed a putative network of genes that we predicted to be co-expressed with LEA genes in maturing rice embryos. Experimental verification of our predictions determined that 179 (73.36%) out of 244 of the predicted genes co-expressed with the LEA genes in maturing rice embryos. This value is significantly greater than the proportion of all rice genes that were determined, based on MPSS experimental data, to be expressed in maturing rice embryos, being 27.99% (TPM≥1) and 20.07% (TPM≥4). These findings are consistent with a number of other studies in plants that have used promoter motif analysis to link gene groups to defined biological processes [11, 31].

In plants, the LEA genes are believed to function in protecting cellular components during developmentally induced desiccation in embryos and during water deficit stress in vegetative tissue [9]. We identified a group of 31 LEA genes that were experimentally determined (MPSS) to be co-expressed in maturing rice embryos (Supplementary File 1), and using ab initio methodology, we identified 30

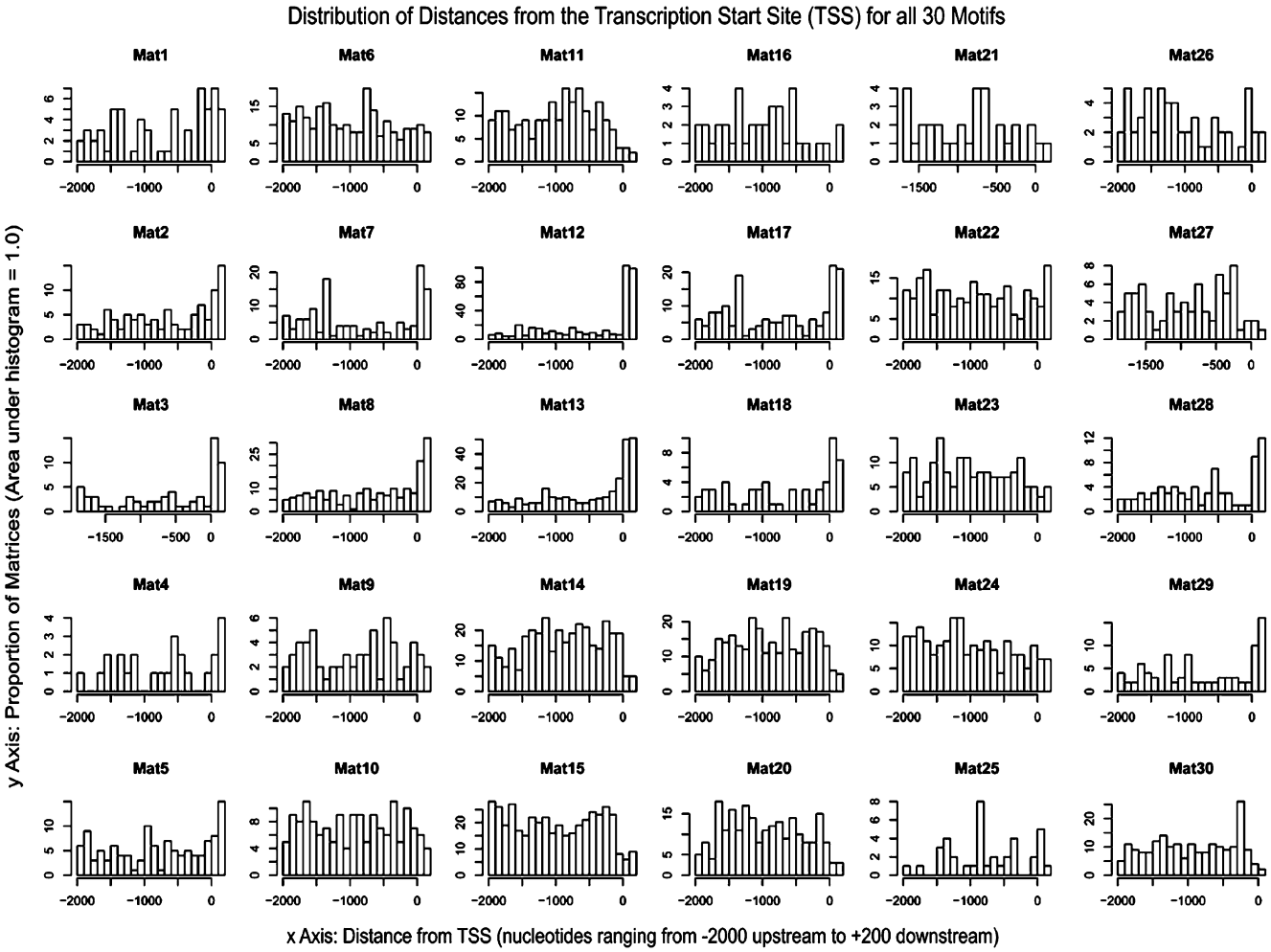

Fig. 1 The average spatial distribution of all 30 identified motifs relative to the TSS across promoters of all 244 predicted genes.

enriched motif families in the promoters of these genes (Table 1). These motifs we believe could be causative for co-expression of the LEA genes in maturing embryos.

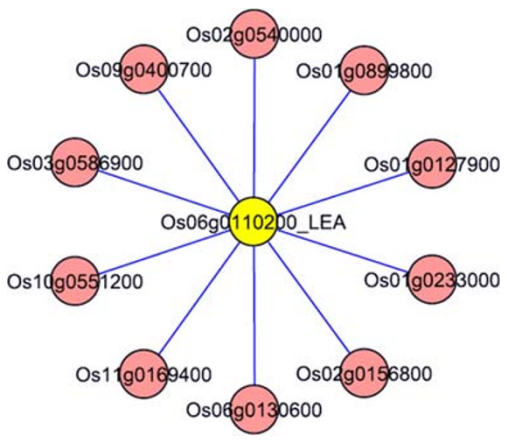

Fig. 2 Network diagram depicting the link/edge relationship between a single LEA gene (*yellow*) and its associated predicted genes (*purple*) that share the highest number of common promoter motifs.

An analysis of the consensus family motifs using the PATCH program in the Transfac database indicated that 19 of the 30 consensus motifs contain sequences that correspond to experimentally confirmed binding sites for specific plant TFs (Table 1). A number of these TFs correspond to those that are well-established regulators of transcriptional responses during water-deficit-related abiotic stresses and embryo development which are both well-established conditions that induce the expression of LEA genes in plants (see Supplementary File 2 for description of TFs) [24]. In brief, according to the PATCH program, the sequences of some of these motifs correspond to abscisic acid (ABA) response elements (ABRE, ABA being a key abiotic stress-activated plant hormone) and dehydration response elements (DRE) which are considered master switches in regulating drought-, cold-, and high salt-responsive gene expression in plants including that of LEA genes [8]. Additionally, a number of motifs that regulate endosperm-specific gene expression were also identified including DNA binding with one finger (DOF)

class prolamine-box binding factors (PBF [33]) and MYB class GAMYB TFs [7]. The identification of these motifs in the promoters of the LEA genes is consistent with their being representative of promoter elements that would regulate the transcription of LEA genes and other genes regulated during abiotic stresses and during embryo development.

The occurrences of each of these motifs in the promoter of each LEA gene were used to build a promoter signature for each individual LEA gene. The signature for each LEA gene was then used to identify other genes in the rice genome that contained the most similar signature (by way of the highest number of shared motifs) and thus, have the greatest potential to be co-regulated with the LEA genes. This analysis identified an additional 244 rice genes that were included in a putative co-expressing LEA gene network. There was an enrichment of some motifs in promoter regions ranging from 0 to 200 nucleotides downstream of the transcription start site (TSS; Fig. 1). This observation is consistent with a study in *Arabidopsis* which documented that promoters have a compact nature [31].

A detailed literature search that was performed for 110 of the 244 identified genes indicated that the function of many of the genes, which possessed functional descriptions, could reasonably be linked to LEA gene functions during embryo development and water deficit stress responses (Supplementary File 5).

The putative LEA co-expressing gene network included a number of genes encoding lipid transfer/seed storage proteins, lipolytic enzymes, and amino acid transporters which may be involved in the building/mobilization of storage reserves during seed embryonic development. Further, numerous A1 peptidases were also present which have been shown to be expressed in developing seed pods and be involved in the proteolytic processing and maturation of seed storage proteins in numerous plant species including rice [12] and additionally have proteolytic roles during water deficit stress [5].

The list also included genes involved in abiotic stress signaling including ABA-inducible kinases and some well-characterized components of the phosphatidylinositol second messenger signaling pathway [26], cellular protection and detoxification, photosynthesis, ion transport, and cell cycle regulators. It is also worth noting that 44 hypothetical proteins with unknown functions were identified and confirmed to be co-expressed with the LEA genes thus linking them to a specific biological response. These genes are thus interesting candidates for future studies investigating systemic late embryogenesis and/or drought-response-related genes that can be targeted for biotechnological interventions.

As previously stated, experimental validation of our putative LEA gene network determined that 179 of the 244 genes (73.36%) co-expressed with the LEA genes in maturing rice embryos. The high success rate of our predictions is put into perspective when considering other studies that have attempted to identify groups of co-responsive genes based on the presence of specific *cis*-elements in their promoters. Attempts to identify ABA-responsive genes in plants have reported success rates of 67.5% in *Arabidopsis* [35] and 49% in rice [22], with the latter being considered particularly high by the authors. It is also noteworthy that the success rate reported for *Arabidopsis* was based on their top 40 predicted genes and not all predicted genes as in our study. Further, both these studies were dependent on knowledge of well-defined experimentally determined *cis*-elements for their analysis.

The high success rate of our study also compares quite favorably with similar studies performed in non-plant organisms. In *Drosophila*, the identification of Dorsal responsive genes based on the presence of known *cis*-elements yielded a 34% success rate [18], while in the nematode, *Caenorhabditis elegans*, the use of defined *cis*-elements that are characteristic to target gene promoters reported an overall success rate of 72% for 57 arbitrarily selected predictions of interneuron AIY-expressed genes [32]. That analysis, however, required the use of defined AIY motifs and phylogenetic footprinting over genomic sequence data from two nematodes to identify candidate genes. In comparison, our method predicted 244 genes using genomic sequence data from a single organism. Contrary to Wenick and Hobert [32], all of our predictions were experimentally tested, with no selection bias, and 73.36% of genes were confirmed to co-express with the LEA genes. As noted above and depicted in Fig. 3, the

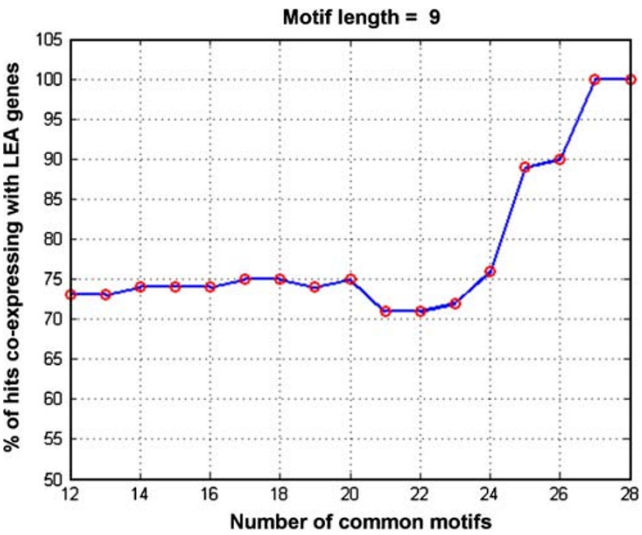

Fig. 3 Correlation between the percentage of predicted genes confirmed to co-express in the maturing embryo and the number of motifs they share with the original LEA genes.

success rate of our method increases up to 100% if we select the top ranked genes, i.e., those that share 27 or more common motifs. The strong positive correlation between motif number and co-expression provides compelling evidence that supports the biological relevance of the identified motifs. Furthermore, the positive expression of 73.36% of our predicted genes significantly exceeded that from all rice genes determined to express in maturing rice embryos by MPSS, being 20.1% (when using the TPM≥4 cut-off used for LEA and predicted genes). Even when applying a less stringent positive expression criterion (TPM≥1), only 28.0% of rice genes expressed positive, providing compelling support for our method as also demonstrated with the previously determined *p* values for the enrichment of experimentally confirmed expression in our predicted gene set.

We were intrigued to determine if the sole presence of our motifs that correspond with ABRE and DRE in the promoters of the predicted 244 genes could alone account for the found expression in mature embryos. Our results show that out of 179 genes co-expressing with LEA genes, 72.07% (129/179) contain motifs related to DRE or ABRE or both. At the same time, for the predicted genes that were not co-expressed with the LEA genes (65), we observe that 69.23% (45/65) contain motifs that correspond to DRE or ABRE or both (see Supplementary File 9). We therefore conclude that the presence of DRE or ABRE motifs, or both, in promoters is not itself sufficient to account for the accurate prediction of co-expression. Thus, other motifs that we identified appear to be required and may act synergistically with these to secure specific gene co-expression.

The higher overall success rate of our study may reflect the limitations in other methods that are dependent on using known and exclusive types of *cis*-elements in predicting co-responsive genes. Since not all *cis*-elements are known and transcription regulation most likely results from the presence and a combinational use of multiple transcription factor binding sites [21], computational identification of such sites can provide a rapid and cost-effective method for identifying groups of co-expressing genes with high success. Additionally, the use of computationally derived motifs allows a global spectrum of application of the method since it can be applied to any biological process occurring within a eukaryotic organism without relying on or being restricted to well-studied processes in well-studied organisms that have experimentally confirmed *cis*-elements available.

This computationally based prediction technique is particularly useful and applicable to newly sequenced eukaryotic genomes from species for which there is little global expression data available. The technique can be used to build putative transcriptional networks of genes based on promoter motif content similarity, which we predict to function coherently in response to defined biological conditions. Thus, both genes with and without known assigned functions can be linked to specific biological processes based on their promoter similarities and their predicted co-expression (under specific conditions) with genes of well-defined functions. Although this study was performed in rice, we believe it can be applied to a wide range of eukaryotes, including other plant species, animals, humans, and fungi since gene transcription is critically regulated by the combinational presence and use of specific *cis*-regulatory sequences in the promoter regions of genes in eukaryotic organisms in general [30].

In comparison to other approaches, our method (a) predicts co-expressing genes by selecting the best matches for each promoter of the TGG relying on the specific promoter motif combinations, (b) does not require previously defined models of transcription factor binding sites or knowledge of specific transcription factors that control TGG, (c) uses sequence data of only one genome, and (d) is applicable to any genome.

Conclusions

In summary, we demonstrate that similarities in promoter composition, interpreted in terms of the pool and number of shared motifs, can be used to identify putative transcriptional networks of genes that co-express with rice LEA genes. A literature analysis indicates that many of these genes could plausibly function coherently with the LEA genes during developmentally induced desiccation in the embryo. This type of analysis can greatly contribute towards understanding the function of newly annotated genes since it can be used to functionally associate them with genes that have well-defined functions in specific biological processes. Further, it provides valuable information regarding the transcriptional regulation of functionally related gene networks which could greatly facilitate in biotechnological manipulations to improve cellular responses to specific biological conditions.

Methods

Target gene group and promoters

The first step of the analysis is the selection of a target group genes. In our case we identified 31 rice LEA genes through MPSS analysis [19] that were determined to be expressed in mature rice embryos (see Supplementary File 1). MPSS provides a comprehensive assessment of gene expression by generating short sequence tags, each 20 bp long, produced from a defined position for each transcript.

Promoter sequences for genes covering the region [−2,000, +200] relative to the transcription start site were obtained from the International Rice Genome Sequencing Project [14].

Motif identification

To identify motifs enriched in the promoter regions, we used the Dragon Motif Builder system (http://apps.sanbi.ac.za/MotifBuilder/index.php) [13]. In total, we identified 30 enriched motif families with motifs of nine nucleotides in length. We used the following parameters: method = EM2, threshold=0.875, and the random DNA background with equal proportion of the four nucleotides. The details about the algorithm of DMB and a guide for interpretation of its results can be found on the system's website. The spatial distribution of all 30 motifs in the promoter regions of the predicted genes was determined (Fig. 1).

Determining promoter with similar content

We have used position weight matrix of each of the 30 motif families identified with DMB, and with the same threshold used for motif identification, we predicted motifs on the promoters of all rice genes. Then, for each of the promoters of genes from the TGG, we searched all other promoters that shared with it the highest number of common annotated promoter motifs. We have limited the number of predicted promoters/genes to the top three promoters that shared the highest number of common promoter motifs with the TGG. If it was not possible to limit the number of candidate promoters to three, we extended the set of associated promoters to include all those promoters that had the highest number of promoter elements. These associations were then used to generate a TGG-like transcriptional regulatory network (see Supplementary File 4).

Experimental confirmation of co-expression of predicted genes with the TGG

The rice cultivar Tainung 67 (*Oryza sativa* L. ssp. *japonica*) was grown in the paddy field at the Academia Sinica campus. Embryos were harvested and dissected from the seeds at 15–20 or 25 days after pollination (DAP) and designated as milky stage embryos (ME) and yellow stage embryos (YE), respectively. The total RNA was extracted using Trizol (Invitrogen). First-strand cDNA was synthesized using standard protocols and SuperScript III reverse transcriptase (Invitrogen). The PCR reaction was performed using the primers sets listed in Supplementary File 8. The amplification was performed using 30 or 35 cycles consisting of 15 s at 94°C, 30 s at 60°C, and 60 s at 72°C, following an initial denaturation cycle of 2 min at 94°C. The final extension step was performed at 72°C for 3 min. The PCR products were separated by electrophoresis and stained with ethidium bromide. The sample collection, RT-PCR, and gel analysis were all performed in duplicate.

The RT-PCR gel image (Supplementary File 7) intensities were graded using standard techniques. A value of 0 was assigned to genes when no PCR products were detected. Samples that gave positive products were assigned a value of 1 (weakest) to 5 (strongest). The values presented in Supplementary File 6 are the average values determined by three independent assessments.

Massively parallel signature sequencing data

RNA samples were extracted from ME and YE. The RNA samples were sent to Illumina Company for custom service of MPSS analysis [19]. The total tag number received from Illumina was 3,520,358. The raw number was normalized to a metric of TPM. Positive expression of LEA genes based on MPSS data was limited to a minimal signal of at least 4 TPM according to Brandenberger et al. [3].

The percentage of all genomic non-transposable element genes that are expressed in maturing rice embryos was determined using global MPSS expression analysis. This analysis was performed using a cut-off of at least 4 TPM (as used for positive selection of LEA genes and predicted genes) and also with the less stringent cut-off of at least 1 TPM.

Statistical test for enrichment

We calculate the p values for the enrichment of the experimentally confirmed genes that co-express with LEA genes in our computationally predicted gene set, relative to the whole rice genome. We used Fisher's exact right-side test based on hypergeometric distribution and corrected for multiplicity testing by the Bonferroni method. The parameters used are as follows:

Genes predicted to co-express with the LEA genes: n=244
Genes with experimentally confirmed expression (out of 244): k=179
Total number of genes in the rice genome: N=41,047
Total number of rice genes expressing in embryo (TPM≥1): K=11,488
Bonferroni correction factor=41,047
p value=4.62e−049, corrected for multiplicity testing p value=1.90e−044
Genes predicted to co-express with the LEA genes: n=244
Genes with experimentally confirmed expression (out of 244): k=179

Total number of genes in the rice genome: N=41,047
Total number of rice genes expressing in embryo (TPM≥4): K=8,241
Bonferroni correction factor=41,047
p value=3.33e−072, corrected for multiplicity testing p value=1.37e−067

Online database

We have created an online Dragon Database for Exploration of late embryogenesis abundant genes in rice (http://apps.sanbi.ac.za/dlea) to allow access to our results and data. Using Rice Annotation Project (RAP, eg: Os01g0159600) or TIGR (LOC_Os01g06630) identifiers, one can access the promoter details for individual genes. This provides information on the number of occurrences and spatial location of all motifs present in individual gene promoters. Further, the database also contains information generated with the DMB algorithm that illustrates the spatial distribution of the best motifs from each of the motif families in the promoters of the TGG (LEA genes). The site also contains links to the RAP database (http://rapdb.dna.affrc.go.jp/) that provides additional information on gene annotations [15, 20, 28].

Acknowledgments SM received postdoctoral fellowship from NBN; CG received support from NRF; CRM received support from SSABMI program; MK received postdoctoral fellowship from the Claude Leon Foundation; MM received support from NBN and NRF FA2006040900002; YIH received support from NSC; VBB received partial support from the DST/NRF Research Chair grant, NBN, and NRF grants FA2007051400013, ICD2006071000003, and FA2006040900002.

References

1. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 2004;5:18.
2. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell 2004;117:185–98.
3. Brandenberger R, Khrebtukova I, Thies RS, Miura T, Jingli C, Puri R, Vasicek T, Lebkowski J, Rao M. MPSS profiling of human embryonic stem cells. BMC Dev Biol 2004;4:10.
4. Brazma A, Jonassen I, Vilo J, Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale. Genome Res 1998;8:1202–15.
5. Cruz de Carvalho MH, d'Arcy-Lameta A, Roy-Macauley H, Gareil M, El Maarouf H, Pham-Thi AT, Zuily-Fodil Y. Aspartic protease in leaves of common bean (*Phaseolus vulgaris* L.) and cowpea (*Vigna unguiculata* L. Walp): enzymatic activity, gene expression and relation to drought susceptibility. FEBS Lett 2001;492:242–6.
6. Davidson EH, McClay DR, Hood L. Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci U S A 2003;100:1475–80.
7. Diaz I, Vicente-Carbajosa J, Abraham Z, Martinez M, Isabel-La Moneda I, Carbonero P. The GAMYB protein from barley interacts with the DOF transcription factor BPBF and activates endosperm-specific genes during seed development. Plant J 2002;29:453–64.
8. Dubouzet JG, Sakuma Y, Ito Y, Kasuga M, Dubouzet EG, Miura S, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. OsDREB genes in rice, *Oryza sativa* L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression. Plant J 2003;33:751–63.
9. Dure LI, Crouch M, Harada J, Ho T-HD, Mundy J, Quatrano R, Thomas T, Sung ZR. Common amino acid sequence domains among the LEA proteins of higher plants. Plant Mol Biol 1989;12:475–86.
10. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
11. Geisler M, Kleczkowski LA, Karpinski S. A universal algorithm for genome-wide in silico identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in *Arabidopsis*. Plant J 2006;45:384–98.
12. Hiraiwa N, Kondo M, Nishimura M, Hara-Nishimura I. An aspartic endopeptidase is involved in the breakdown of propeptides of storage proteins in protein-storage vacuoles of plants. Eur J Biochem 1997;246:133–41.
13. Huang E, Yang L, Chowdhary R, Kassim A, Bajic V. An algorithm for ab initio DNA motif detection. In: Bajic VB, Tan TW, editors. Information processing and living systems. Singapore: World Scientific; 2005. p. 611–4.
14. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature 2005;436:793–800.
15. Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R, Bureau T, Burr F, Costa DO, Fuks G, Habara T, Haberer G, Han B, Harada E, Hiraki AT, Hirochika H, Hoen D, Hokari H, Hosokawa S, Hsing YI, Ikawa H, Ikeo K, Imanishi T, Ito Y, Jaiswal P, Kanno M, Kawahara Y, Kawamura T, Kawashima H, Khurana JP, Kikuchi S, Komatsu S, Koyanagi KO, Kubooka H, Lieberherr D, Lin YC, Lonsdale D, Matsumoto T, Matsuya A, McCombie WR, Messing J, Miyao A, Mulder N, Nagamura Y, Nam J, Namiki N, Numa H, Nurimoto S, O'Donovan C, Ohyanagi H, Okido T, OOta S, Osato N, Palmer LE, Quetier F, Raghuvanshi S, Saichi N, Sakai H, Sakai Y, Sakata K, Sakurai T, Sato F, Sato Y, Schoof H, Seki M, Shibata M, Shimizu Y, Shinozaki K, Shinso Y, Singh NK, Smith-White B, Takeda J, Tanino M, Tatusova T, Thongjuea S, Todokoro F, Tsugane M, Tyagi AK, Vanavichit A, Wang A, Wing RA, Yamaguchi K, Yamamoto M, Yamamoto N, Yu Y, Zhang H, Zhao Q, Higo K, Burr B, Gojobori T, Sasaki T. Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. Genome Res 2007;17:175–83.
16. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein–protein interactions. Genome Res 2002;12:37–46.
17. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res 2004;14:1085–94.
18. Markstein M, Markstein P, Markstein V, Levine MS. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc Natl Acad Sci U S A 2002;99:763–8.

19. Nobuta K, Venu RC, Lu C, Belo A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang GL, Meyers BC. An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol 2007;25:473–7.
20. Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, Sasaki T. The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Res 2006;34:D741–744.
21. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001;29:153–9.
22. Ross C, Shen QJ. Computational prediction and experimental verification of HVA1-like abscisic acid responsive promoters in rice (*Oryza sativa*). Plant Mol Biol 2006;62:233–46.
23. Schulze A, Downward J. Navigating gene expression using microarrays—a technology review. Nat Cell Biol 2001;3:E190–195.
24. 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;3:217–23.
25. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.
26. Takahashi S, Katagiri T, Hirayama T, Yamaguchi-Shinozaki K, Shinozaki K. Hyperosmotic stress induces a rapid and transient increase in inositol 1,4,5-trisphosphate independent of abscisic acid in *Arabidopsis* cell culture. Plant Cell Physiol 2001;42:214–22.
27. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 1999;96:2907–12.
28. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei FJ, Hsing YI, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Haberer G, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Sakai Y, Nobushima S, Sakata K, Barrero RA, Sato Y, Souvorov A, Smith-White B, Tatusova T, An S, An G, OOta S, Fuks G, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Chaparro C, Piegu B, Panaud O, Echeverria M. The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res 2008;36:D1028–1033.
29. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999;22:281–5.
30. Tuch BB, Li H, Johnson AD. Evolution of eukaryotic transcription circuits. Science 2008;319:1797–9.
31. Vandepoele K, Casneuf T, Van de PY. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. Genome Biol 2006;7:R103.
32. Wenick AS, Hobert O. Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C. elegans*. Dev Cell 2004;6:757–70.
33. Wu C, Washida H, Onodera Y, Harada K, Takaiwa F. Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. Plant J 2000;23:415–21.
34. Yan X, Mehan MR, Huang Y, Waterman MS, Yu PS, Zhou XJ. A graph-based approach to systematically reconstruct human transcriptional regulatory modules. Bioinformatics 2007;23:i577–86.
35. Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS. Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*. Bioinformatics 2005;21:3074–81.