Plant Biotechnology Journal

Plant Biotechnology Journal (2014) **12**, pp. 1015–1026 doi: 10.1111/pbi.12206

Computational discovery of soybean promoter *cis*-regulatory elements for the construction of soybean cyst nematode-inducible synthetic promoters

Wusheng Liu[1], Mitra Mazarei[1], Yanhui Peng[1], Michael H. Fethe[1], Mary R. Rudis[1], Jingyu Lin[1], Reginald J. Millwood[1], Prakash R. Arelli[2] and Charles Neal Stewart, Jr.[1,*]

[1]*Department of Plant Sciences, The University of Tennessee, Knoxville, TN, USA*
[2]*Crop Genetics Research Unit, USDA-ARS-MRS, Jackson, TN, USA*

Received 10 February 2014;
revised 14 April 2014;
accepted 23 April 2014.
**Correspondence* (Tel +1 865 974 6487;
fax +1 865 946 1989;
email nealstewart@utk.edu)

Keywords: *de novo* motif discovery, plant synthetic biology, *pporRFP*, transgenic soybean hairy system, soybean cyst nematode, soybean.

Summary

Computational methods offer great hope but limited accuracy in the prediction of functional *cis*-regulatory elements; improvements are needed to enable synthetic promoter design. We applied an ensemble strategy for *de novo* soybean cyst nematode (SCN)-inducible motif discovery among promoters of 18 co-expressed soybean genes that were selected from six reported microarray studies involving a compatible soybean–SCN interaction. A total of 116 overlapping motif regions (OMRs) were discovered bioinformatically that were identified by at least four out of seven bioinformatic tools. Using synthetic promoters, the inducibility of each OMR or motif itself was evaluated by co-localization of gain of function of an orange fluorescent protein reporter and the presence of SCN in transgenic soybean hairy roots. Among 16 OMRs detected from two experimentally confirmed SCN-inducible promoters, 11 OMRs (i.e. 68.75%) were experimentally confirmed to be SCN-inducible, leading to the discovery of 23 core motifs of 5- to 7-bp length, of which 14 are novel in plants. We found that a combination of the three best tools (i.e. SCOPE, W-AlignACE and Weeder) could detect all 23 core motifs. Thus, this strategy is a high-throughput approach for *de novo* motif discovery in soybean and offers great potential for novel motif discovery and synthetic promoter engineering for any plant and trait in crop biotechnology.

Introduction

Microarray and next-generation sequencing technologies are the most powerful methods for high-throughput gene expression profiling. It is assumed that genes with similar mRNA expression profiles are likely to be co-regulated via the same signal transduction pathways (Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). Strongly co-expressed genes are more likely to have their promoters bound by common transcription factors (TFs; Yu *et al.*, 2003). To understand the mechanisms that regulate the correlated expression of genes, it is important and challenging to identify the transcription factor binding sites (TFBSs, also known as *cis*-regulatory elements or motifs) within the promoters of the co-regulated genes of interest. TFBSs are the functional motifs that determine temporal and spatial expression patterns. Database-assisted promoter analysis can be conducted for the identification of known motifs by submitting promoter sequences to the three main databases [i.e. PlantCARE (Lescot *et al.*, 2002), PLACE (Higo *et al.*, 1999) and TRANSFAC (Matys *et al.*, 2003)].

However, hunting for novel *cis*-regulatory elements in terms of developing computational tools for the prediction of such unknown motifs still seems much like searching for a needle in a haystack (D'haeseleer, 2006). Over the past 10 years, numerous computational tools have become available for *de novo* motif discovery, where nothing is assumed aside from the transcription factors or their preferred binding sites (Tompa *et al.*, 2005). These *de novo* motif discovery tools identify short DNA sequence motifs that are statistically overrepresented in the provided promoter regions of the co-regulated genes. The motif discovery algorithms of these tools can be divided into three distinct classes such as enumeration, deterministic optimization and probabilistic optimization (D'haeseleer, 2006). An enumerative algorithm such as used in MDScan (Liu *et al.*, 2002), Weeder (Pavesi *et al.*, 2004) or YMF (Sinha and Tompa, 2003) is a word-counting method that exhaustively searches for all possible motifs by counting the number of occurrences of all *n*-mers or consensus sequences in the target sequences. A deterministic optimization algorithm uses expectation maximization (EM) to simultaneously optimize a position weight matrix (PWM) description of each motif and the binding probabilities for its associated sites. The most popular EM algorithms are MEME (Bailey *et al.*, 2006), Improbizer (Ao *et al.*, 2004) and BioProspector (Liu *et al.*, 2001). A probabilistic optimization algorithm uses Gibbs sampling for a stochastic implementation of EM. Examples are MotifSampler (Thijs *et al.*, 2002), W-AlignACE (Chen *et al.*, 2008), ANN-Spec (Workman and Stormo, 2000), Consensus (Hertz and Stormo, 1999) and Oligo/dyad analysis (van Helden *et al.*, 1998, 2000).

A pivotal study comparing five different *de novo* motif discovery tools demonstrated that the accuracy of each tool is about 15%–25%, even though these tools are capable of predicting at least one motif correctly more than 90% of the time using large *Escherichia coli* datasets (Hu *et al.*, 2005). A more large-scale comparison between 13 different *de novo*

© 2014 Society for Experimental Biology, Association of Applied Biologists and John Wiley & Sons Ltd

motif discovery tools revealed that each method typically covers only a small subset of the motifs (thus relatively little overlap between tools) using 52 benchmarked datasets from fly, human, mouse and yeast (Li and Tompa, 2006; Tompa *et al.*, 2005). These discoveries led to the development of ensemble methods comprised of multiple motif finders, which provide improvements in accuracy, such as SCOPE (Carlson *et al.*, 2007; Chakravarty *et al.*, 2007), MProfiler (Altarawy *et al.*, 2009), MTAP (Quest *et al.*, 2008), MotifVoter (Wijaya *et al.*, 2008) and SAMF (Yanover *et al.*, 2009). However, the *de novo* motif discovery research in plants falls far behind from the above-mentioned progresses in other kingdoms. A recent good example is that Koschmann *et al.* (2012) used the binding site estimation suite of tools (BEST) consisting of five bioinformatics tools (i.e. MEME, BioProspector, BioOptimizer, AlignACE and Consensus) for the discovery of novel elicitor-responsive *cis*-regulatory elements in *Arabidopsis*, followed by functional analysis of the detected consensus sequences using synthetic promoters.

In the present study, we conducted a comprehensive bioinformatic analysis for *de novo* soybean cyst nematode (SCN, *Heterodera glycines* Ichinohe)-inducible motif discovery in the soybean [*Glycine max* (L.) Merr.] genome during a compatible interaction with SCN. SCN is an obligate, sedentary endoparasite of roots and the most damaging pathogen of soybean worldwide (Davis *et al.*, 2004; Endo, 1991; Gheysen and Mitchum, 2009; Kim *et al.*, 1987). During infection, the second-stage juveniles (J2) enter host roots and migrate intracellularly within the cortical tissue to the vascular cylinder. J2 then initiate the formation of specialized feeding structures that are called syncytia. Syncytia development in incompatible roots collapses 3–4 days postinfection (dpi), while it continues in compatible roots leading to feeding site formation from which the nematodes feed during their life cycles (Davis *et al.*, 2004; Endo, 1991; Gheysen and Mitchum, 2009; Kim *et al.*, 1987). Syncytium formation and maintenance are mediated via interactions between nematode secretions and changes in host gene expression (Davis *et al.*, 2004; Gheysen and Mitchum, 2009). Recently, several Affymetrix Soybean GeneChip microarray analyses studied soybean gene expression at different time points during the susceptible soybean–SCN interaction (Ithal *et al.*, 2007a,b; Klink *et al.*, 2007a,b; Mazarei *et al.*, 2011; Puthoff *et al.*, 2007). Here, we selected potentially co-regulated genes from the above-mentioned microarray datasets based on the expression profiles, followed by a comprehensive bioinformatic analysis for *de novo* motif discovery. Next, extensive functional analysis of the detected overlapping motif regions (OMRs, including the surrounding nucleotides) and the motifs alone was conducted in transgenic soybean hairy roots. We demonstrate that our ensemble strategy was a high-throughput approach for *de novo* SCN-inducible motif discovery in the soybean genome.

Results

De novo SCN-inducible motif discovery using an assemble strategy

Using the Affymetrix Soybean GeneChip assay, our previous study detected 675 genes whose expression was significantly induced in the soybean genome during a susceptible soybean–SCN interaction (Mazarei *et al.*, 2011). We compared these induced genes with those published from other microarray datasets studying the susceptible soybean–SCN interaction (Ithal *et al.*, 2007a,b; Klink *et al.*, 2007a,b; Puthoff *et al.*, 2007) and found 49 common genes from both of our transcriptome datasets (Mazarei *et al.*, 2011) and in at least one of the other five datasets (Ithal *et al.*, 2007a,b; Klink *et al.*, 2007a,b; Puthoff *et al.*, 2007). We assumed that genes with similar transcript profiles are likely to be co-regulated via the same signal transduction pathways (Schulze and Downward, 2001). Thus, 18 out of the 49 candidate genes were selected for *de novo* SCN-inducible motif discovery (Table 1). These candidate genes are mainly defence/stress-related, cell wall-related, and transcription factors and protein kinases (Table 1). Our assumption was that there were at least two out of the 18 promoters bound by the same TF(s); this was the underlying principle for *de novo* motif discovery analysis. Next, we chose two widely used and user-friendly bioinformatic tools from each class of the three *de novo* motif discovery algorithms, that is, MDscan and Weeder as representatives of enumerative algorithms; MEME and BioProspector as representatives of deterministic optimization; and MotifSampler and W-AlignACE as representatives of probabilistic optimization (Table 2). In addition, an ensemble method, SCOPE, was also used in this study (Table 2).

Because the binding sites for interacting transcription factors often co-localize to the same motif regions (i.e. modules; Zhou and Wong, 2004), we used an ensemble strategy to look for the OMRs that were detected by at least four out of the seven bioinformatic tools when running each bioinformatic tool independently and collecting the top ten motifs (Figure 1a). As a result, a total of 116 OMRs were discovered among the 18 promoter regions (Table S1). We found that most promoters contained 4–8 detectable OMRs, while the promoter of Glyma20g19200.1 harboured only one detectable OMR, and the promoter of Glyma04g40130.1 contained 14 (Table S1). In addition, the bioinformatic tools from different algorithms exhibited different capabilities to predict the 116 OMRs (Figure 2). The two representatives of enumerative algorithms—MDscan and Weeder—predicted 97.3% and 82.9% of the overall OMRs, respectively. However, only one computational tool from either deterministic optimization (MEME, 72.1%) or probabilistic optimization (W-AlignACE, 97.3%) exhibited a similar ability in predicting the overall OMRs as the two enumerative tools did. The other two tools, BioProspector (deterministic optimization) and MotifSampler (probabilistic optimization), only detected 23.4% and 45.9% of the overall OMRs, respectively. SCOPE, which uses three different algorithms, predicted 84.7% of the overall OMRs.

Time-course analysis of inducibility of seven out of 18 promoter regions by SCN infection

Before we tested the inducibility of the detected OMRs using SCN treatment, we selected seven out of the 18 promoter regions for time-course analysis of inducibility by SCN treatment using a transgenic soybean hairy root system generated via *Agrobacterium rhizogenes*-mediated genetic transformation as previously reported (Kereszt *et al.*, 2007; Lin *et al.*, 2013; Tables S2 and S3). These seven promoter regions were selected for promoter functional analysis because they contained at least seven OMRs, which, in total, accounted for almost half of the overall 116 OMRs detected among all of the 18 promoter regions (Table S1). The promoter region of Glyma04g40130.1 contained 14 OMRs but failed to be amplified by PCR for cloning; thus, it was not included in this study. The seven 1-kb-long promoter regions

Table 1 The 18 candidate genes selected for *de novo* motif discovery

Probe ID[†]	Gene ID[‡]	Fold change[†]	Gene ontology molecular function[†]	Induced in ref.[§]	
				3 dpi	8 dpi
Gma.11298.3.S1_a_at	**Glyma20g19200.1***	2.5471	Pectate lyase	(1)	(2)
GmaAffx.26533.1.A1_s_at	Glyma06g08860.1	3.3147	Bark storage protein	(3)	–
Gma.11336.2.S1_at	Glyma08g05820.1*	2.06177	*Arabidopsis* thaumatin-like protein 1	(4)	–
GmaAffx.12832.1.S1_at	Glyma18g18920.1	2.16193	Xyloglucan/xyloglucosyl transferase 33	(4)	–
GmaAffx.13717.1.S1_at	Glyma04g36520.1*	2.59805	Pectate lyase precursor	(4)	(2)
GmaAffx.50446.1.S1_at	Glyma09g41440.1	4.87596	Cationic peroxidase 1 precursor	(4)	–
Gma.529.1.S1_x_at	Glyma13g23090.1*	2.02918	Purple acid phosphatase-like protein	(5)	(2)
GmaAffx.5537.1.S1_at	Glyma01g04380.1	2.4529	Matrix metalloproteinase	(3)	–
Gma.5627.1.S1_at	Glyma20g27480.1	2.15976	Putative receptor-like protein kinase	(5)	(2)
GmaAffx.65048.1.S1_s_at	Glyma13g30950.1	2.23326	Protein kinase	(4)	(2)
GmaAffx.66986.1.S1_at	Glyma15g06140.1	2.19805	Arabinogalactan protein	(4)	–
GmaAffx.671.1.S1_at	**Glyma13g41160.1***	2.07901	Alpha-expansin 1	(4)	(2)
Gma.8546.1.S1_at	Glyma12g33530.1*	2.23542	Fasciclin-like arabino-galactan protein 4	(4)	–
GmaAffx.89045.1.A1_s_at	Glyma10g39760.1*	2.20908	Xyloglucan endotrans-glycosylase	(4)	–
GmaAffx.92003.1.S1_x_at	Glyma20g38570.1	2.31169	Chalcone-flavonone isomerase 1B-1	(5)	(5)
GmaAffx.93343.1.S1_s_at	Glyma09g41050.1	2.30124	WRKY70	(5)	(5)
Gma.987.1.S1_at	Glyma18g49570.1	2.88416	Expansin	(4)	–
Gma.9913.2.S1_a_at	Glyma04g40130.1	2.28741	WRKY70	(5)	(5)

[†]The data were selected from our microarray datasets (Mazarei *et al.*, 2011).

[‡]The gene IDs were acquired from the Affymetrix and Phytozome websites. The seven genes selected for promoter inducibility by SCN infection are indicated by asterisk marks (*). The two genes selected for inducibility analysis of the detected OMRs by SCN infection are in bold.

[§]References: (1) Klink *et al.* (2007a); (2) Puthoff *et al.* (2007); (3) Ithal *et al.* (2007a); (4) Ithal *et al.* (2007b); (5) Klink *et al.* (2007b).

Table 2 The seven bioinformatic tools used for *de novo* motif discovery

Tools	Algorithm	Motif model	Match model	Objective function	Parameters used besides defaults
MDscan	Greedy (enumerative)	String	PWM	MAP (hmm)	(i) Motif width 6, 8, 10, 12 bp; (ii) Used input sequences as background
Weeder	Enumerative	String	Mismatch	Pattern specificity	(i) Motif width 'large' (6, 8, 10, 12 bp); (ii) Motif appears in some genes; (iii) Used soybean intergenic sequences as background
MEME	Deterministic	Matrix	PWM	*P*-value	0 or 1 or any motifs per gene
BioProspector	Deterministic optimization	Matrix dyad	PWM	Motif overrepresentation (*z*-score)	Used input sequences as background
MotifSampler	Probabilistic (Gibbs)	Matrix	PWM	Log likelihood score	(i) Ran 100 times for the same parameters; (ii) Used soybean intergenic sequences as background
W-AlignACE	Probabilistic (Gibbs)	Matrix	PWM	Motif overrepresentation (MAP score)	None
SCOPE	BEAM PRISM SPACER	Matrix	PWM	Sig score	Used soybean intergenic sequences as background

PWM, position weight matrix.

were PCR-amplified and used to drive the expression of an orange fluorescent protein (OFP)—*pporRFP* from the hard coral *Porites porites* (Alieva *et al.*, 2008; Mann *et al.*, 2012)—in a binary vector pZP222, which contained a GFP screening marker, *35S::mGFP5-ER,* for transgenic hairy root selection (Figure 1b). The inducibility of each promoter region was evaluated by co-localization of gain-of-function orange fluorescence foci and the presence of SCN in transgenic hairy roots. Gain of function of the *pporRFP* reporter gene was determined by the presence of strong fluorescent spots when compared with its neighbouring regions. The negative control vector, pZP *-4635S::pporRFP-35S::GFP,* only showed basal expression before SCN infection and was not inducible by the nematode treatment (Figure S1). Three (i.e. Glyma08g05820.1, Glyma12g33530.1 and Glyma10g39760.1) out of the seven promoters showed strong expression in transgenic hairy roots before SCN infection

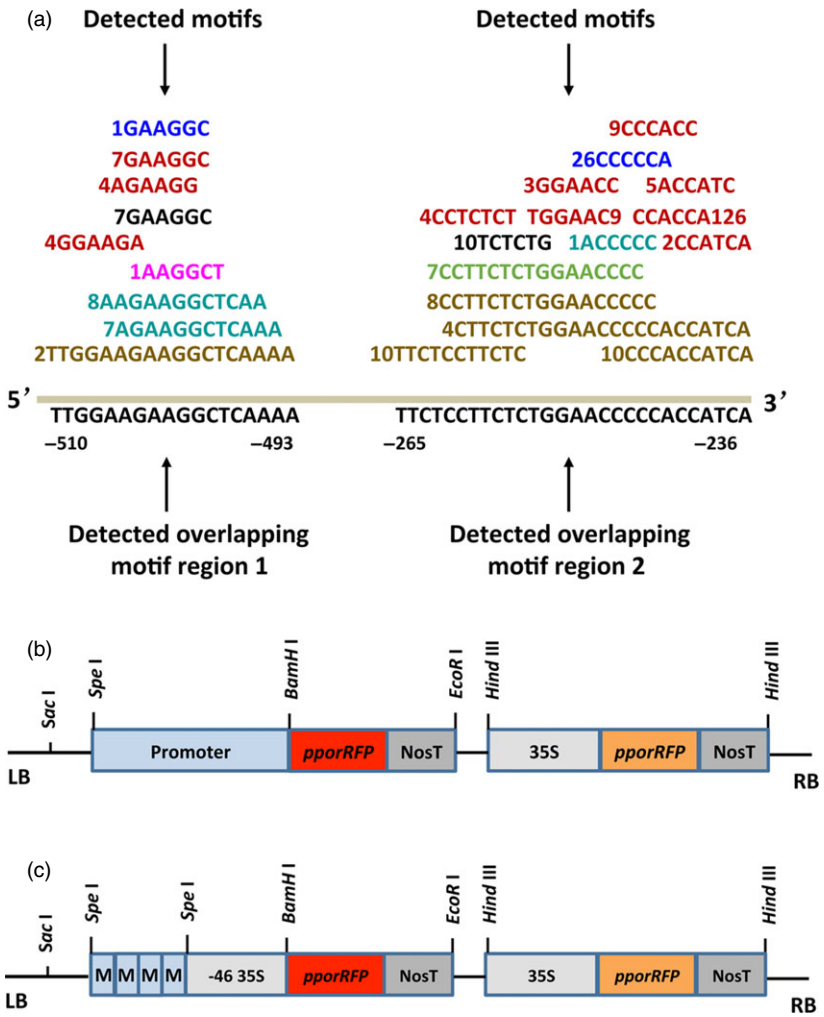

Figure 1 Illustration of the detected overlapping motif regions (OMRs) and the scheme of synthetic promoters. (a) Two OMRs were detected by at least four out of seven bioinformatic tools within the 1-kb-long promoter region (illustrated in a solid line with 5′ and 3′ ends indicated), and were then mapped into the promoter region with their positions being indicated by the numerical numbers underneath the promoter trunk. The detected motifs are shown in different colours above the two detected OMRs, with each colour representing the output from each tool. The numerical numbers before each detected motif indicate the ranking of that motif from the respective tool with one being the top motif. (b, c) Scheme of plasmids containing the *pporRFP* reporter gene driven by each of the seven 1-kb-long promoters (b) or 4× OMRs/core motifs (i.e. M) (c) in the pZP222 backbone harbouring *mGFP5-ER* driven by the CaMV 35S promoter.

(i.e. at time point 0; Table S3). In addition, promoters of Glyma08g05820.1, Glyma04g36520.1, Glyma13g23090.1 and Glyma10g39760.1 exhibited inducibility during the time course of SCN infection for about 4 weeks (Table S3). However, the promoters of Glyma20g19200.1 and Glyma13g41160.1 showed weak OFP expression before SCN treatment but were induced by the nematode infection at only 3 days postinfection (dpi; Table S3). The detected OMRs within these two promoters were chosen for further time-course analysis of inducibility by SCN infection.

Time-course analysis of inducibility of the detected 16 OMRs in two selected promoters by SCN infection as well as by wounding

A total of 16 OMRs were detected bioinformatically within these two promoters (i.e. Glyma20g19200.1 and Glyma13g41160.1; Table S1) and were chosen for the time-course analysis of their inducibility by SCN infection and for down-selection of the core motifs. The seven bioinformatic tools exhibited different abilities in predicting the 16 OMRs among these two promoters (Figure 2; Table S4). Most of the seven tools predicted 62.5%–87.5% of the 16 OMRs, with the exception being BioProspector and Motif-Sampler, which only predicted 31.3% and 37.5% of the 16 regions, respectively (Figure 2).

To conduct the functional analysis of the 16 OMRs in transgenic soybean hairy roots, each OMR was used to design its respective synthetic promoter by fusing tetramers (i.e. four head-to-tail copies) of each individual OMR into upstream of a minimal CaMV 35S promoter driving a *pporRFP* reporter in the pZP222 binary vector, which contains *35S::mGFP5-ER*. Eleven out of the 16 OMRs (i.e. 68.75%) were confirmed to be SCN-inducible in the early stage of SCN infection (i.e. at 3 dpi; Table 3). Among these, seven out of the 11 motifs (i.e. M2, M3, M4, M7, M10, M13 and M14) also exhibited a relatively weak inducibility at time points 10 and 17 dpi (Table 3). Considering the infection process in which J2 SCN penetrates roots, punctures

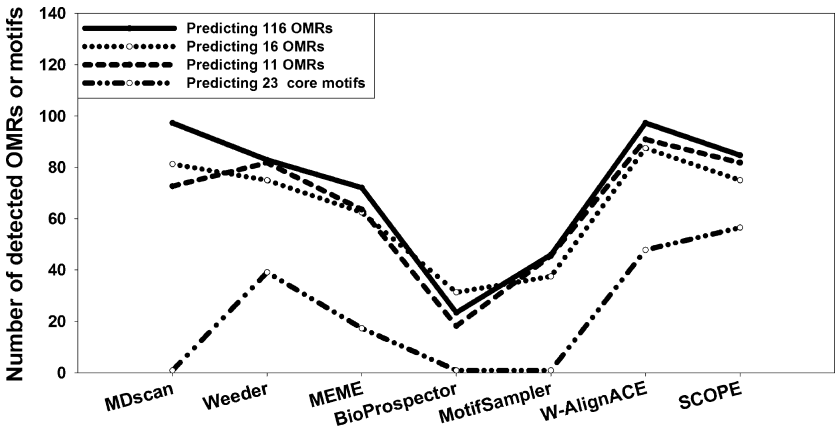

Figure 2 Capability of the seven bioinformatic tools to detect the overlapping motif regions (OMRs) and core motifs. *De novo* motif discovery was conducted using seven different bioinformatics tools among the 1-kb-long promoter regions of 18 co-regulated genes that were significantly induced in the early stages of soybean cyst nematode (SCN) infection during susceptible soybean–SCN interactions. Each bioinformatic tool was used to search for both strands with default parameters unless otherwise specified in Table 2. Either input sequences or soybean intergenic sequences were used as background sequences. The top 10 ranked motifs from the output of each bioinformatic tool were chosen for this study. An ensemble strategy was applied to look for the OMRs that were detected by at least four out of the seven bioinformatic tools within the same promoter regions when running each bioinformatic tool independently and collecting the top ten motifs (Figure 1a). The 116 OMRs were detected among the promoter regions of 18 co-regulated candidate genes. The 16 OMRs, 11 inducible OMRs and 23 core motifs were detected and down-selected on the two chosen promoter regions Glyma20g19200.1 and Glyma13g41160.

cells and migrates to initiate a feeding site, we tested whether wounding could affect promoter activity of the 11 inducible OMRs before down-selection of the core motifs. Our results revealed that wounding could not increase *OFP* expression when driven by each of the 11 inducible OMRs (Table 3). Therefore, the inducibility of the 11 OMRs was initiated from the infection alone rather than wounding.

Down-selection of the 11 inducible OMRs in two promoters for the characterization of core motifs

These 11 SCN-inducible OMRs were down-selected experimentally to further examine narrower OMRs (Table S5; Figure S2), which led to the discovery of a total of 23 core motifs of 5- to 7-bp length (Table 4; Figure 3). The 23 core motifs exhibited weak or no basal expression before SCN treatment (i.e. at time 0). Most of these core motifs showed strong inducibility in the early stage of SCN infection, that is, at time points 0.5–3 dpi. However, a few core motifs, such as M1.1.2.2, M2.3.1, M15.3.2 and M16.2.3, exhibited strong inducibility at time points 10–17 dpi. We also found that wounding alone could not increase *OFP* expression when driven by each core motif (Table 4).

Moreover, the DNA sequences of these 23 core motifs were used as query sequences to cross-check the PlantCARE (Lescot *et al.*, 2002) and PLACE (Higo *et al.*, 1999) databases that contain known *cis*-motifs in plants (Table 5). We found only 1 core motif (i.e. M2.3.1), containing a TATA box-like motif, in the PlantCARE database. While cross-checking the PLACE database, we found that 14 out of the 23 core motifs did not contain any known motifs, indicating that these 14 core motifs are novel in plants. The other core motifs contained known motifs in the PLACE database, which are mainly involved in the regulation of gene expression or defence (Table 5).

In addition, the seven bioinformatic tools exhibited differences in their ability to predict the 23 core motifs within the two promoters (Figure 2; Table S6). MDscan, BioProspector and MotifSampler showed the lowest prediction capability by detecting only 0.9% of the 23 core motifs. The best tool was SCOPE, which predicted 56.5% of the 23 core motifs, followed by W-AlignACE, Weeder and MEME, which detected 47.8%, 39.1% and 17.3% of the core motifs, respectively. Interestingly, we observed that a combination of the best three computational tools, that is, Weeder, W-AlignACE and SCOPE, detected all the 23 core motifs (Table S6).

Discussion

We attempted to combine bioinformatics with functional analysis using synthetic promoters for *de novo* SCN-inducible motif discovery in the soybean genome during a compatible interaction between soybean and SCN. We used a combination of seven different computational tools from different algorithms for *de novo* motif discovery due to the fact that almost all bioinformatic tools have limited ability in their sensitivity and precision at predicting the true motifs (Tompa *et al.*, 2005). These seven computational tools have been widely used and are very user-friendly. Because the underlying principles of regulatory mechanisms in plants are highly poorly understood (Tompa *et al.*, 2005) and *cis*-regulatory elements often occur in close proximity to each other and form functional modules, we applied an ensemble strategy to look for overlapping motif regions that were detected by at least four out of the seven bioinformatic tools. We experimentally confirmed that 11 out of the 16 OMRs within two selected promoters were SCN-inducible (Table 3). Thus, our ensemble strategy for *de novo* motif discovery significantly increased the prediction efficiency from 15% to 20% in Tompa *et al.* (2005) to 68.75% in this study. The down-selection of these 11 SCN-inducible OMRs led to the discovery of 23 novel core motifs, which were experimentally confirmed to be SCN-inducible (Table 4, Figure 3). According to the known PlantCARE and PLACE entries, 14 out of the 23 core motifs appear to be novel (Table 5). The other nine core motifs had never been shown to be SCN-inducible in the two plant motif databases, even though

Table 3 Time-course analysis of inducibility of the 16 detected OMRs within the two selected promoters of Glyma20g19200.1 and Glyma13g41160.1 by SCN infection

			Inducibility by SCN (dpi)					
OMRs	Nucleotide sequence (5′ to 3′)	Length (bp)	0	3	10	17	24	Wounding
1	TAAAATAAAGTTCTTTAATTTGTTTATTATTTAATTATTT	40	–	+++	–	–	–	–
2	TATATAATTAAGTGTTA	17	–	+++	++	–	–	–
3	AGGGATCGA	9	–	+++	++	–	–	–
4	AAAGCGAGGAAAAAAAAGTAAAAAGTAAAAAAGAGAAAGTGG	42	–	+++	++	+	–	–
5	GTGCGAGAGAGAGGGTAAAAAAGTGA	26	–	–	–	–	–	–
6	GAAGCAGAGAGGGTGGTTAATT	22	–	–	–	–	–	–
7	TTAGTTAACGGCGTGAGAGGGACGGCG	27	–	+++	++	+	–	–
8	GGCGTGAAAGGGTTGGGCGTTTGTGGGTTTAGTGTCTATAAAAAACCGTTA	51	–	–	–	–	–	–
10	CACTGAGAGCGAGGGG	16	–	+++	++	+	–	–
11	ACTACATTTATTTAAGTTCCAATTAAAGCTTCGTAATTACTACA	44	–	–	–	–	–	–
12	GCCTCGGGTTGGTTTCTTTC	20	–	–	–	–	–	–
13	AAAAAATAAACATTATTTGGTTTTA	25	–	+++	+++	++	+	–
14	TAAATTTTTTTTAACTTTAAAGTGCTATTAAACAAAAACAT	41	–	+++	++	++	–	–
15	GCAAAGTGAATTATTTTTTTTAAGAAAAAA	30	–	+++	–	–	–	–
16	TCGAGCAAATTAAAATTGAAAACAAA	26	–	+++	–	–	–	–
17	AAGAAAAAAAGAGTGA	16	–	+++	–	–	–	–

OMRs, overlapping motif regions; +++, strong induction (see the after staining image of M1.1.1 in Figure 3); ++, median induction (see the after staining image of M1.1.2.2 in Figure 3); +, weak induction (see the after staining image of M2.2.2 in Figure 3); –, no induction (see the after staining image of the negative control promoter -4635S in Figure 3); SCN, soybean cyst nematode.

Table 4 Time-course analysis of inducibility of the 23 core motifs within the two selected promoters by SCN infection and wounding

			Inducibility by SCN (dpi)								
Core motif	Nucleotide sequence	Length (bp)	0	0.5	1	2	3	10	17	24	Wounding
1.1.1	TAAAGT	6	–	+++	+++	+++	+++	++	++	+	–
1.1.2.1	TCTTTA	6	–	+++	+++	+++	+++	++	+	+	–
1.1.2.2	TTAATT	6	+	+	+	++	+++	++	+	+	–
2.2.1	TAATTA	6	–	+	+++	+++	+++	+	+	–	–
2.2.2	AAGTG	5	–	–	++	+	+++	++	++	–	–
2.2.3	AGTGTTA	7	+	–	++	+++	++	+	+	–	–
2.3.1	ATATAA	6	–	–	–	++	+++	++	+	–	–
2.3.2	ATTAAGT	7	+	–	+++	+++	+++	+	+	+	–
3.1.1	AGGGA	5	+	+++	+++	+++	+++	++	++	–	–
4.1.1	AAAG	4	–	–	+	+	++	+	+	–	–
4.1.2.1	GAAAA	5	+	++	+++	+++	+++	++	++	+	–
4.2.2.1	AAAGTG	6	+	+++	+++	+++	++	+	–	–	–
4.2.2.2	TGGTG	5	+	++	+++	++	+	+	–	–	–
4.3.2	GTAAAAAA	8	–	++	+++	+++	+++	++	++	+	–
7.1.1	TTAGTT	6	–	+	+++	++	+++	++	++	+	–
7.1.2	GTTAAC	6	–	–	–	++	++	++	++	–	–
10.2.1	AGAGCG	6	–	+++	+++	+++	+++	++	+	–	–
10.2.3	GCGAGG	6	–	–	+++	+++	+++	++	++	+	–
13.1.2	ATTATT	6	–	+++	+++	+++	+++	+	+	–	–
14.3.1	GCTATTA	7	–	+++	+++	+++	+++	+++	++	+	–
15.3.2	TTTTAA	6	–	–	–	+	+++	++	++	+	–
16.2.3	AACAAA	6	–	–	–	++	+++	++	+	–	–
17.2	GAGTGA	6	–	+++	+++	+++	+++	++	+	–	–

+++, strong induction (see the after staining image of M1.1.1 in Figure 3); ++, median induction (see the after staining image of M1.1.2.2 in Figure 3); +, weak induction (see the after staining image of M2.2.2 in Figure 3); –, no induction (see the after staining image of the negative control promoter -4635S in Figure 3); SCN, soybean cyst nematode.

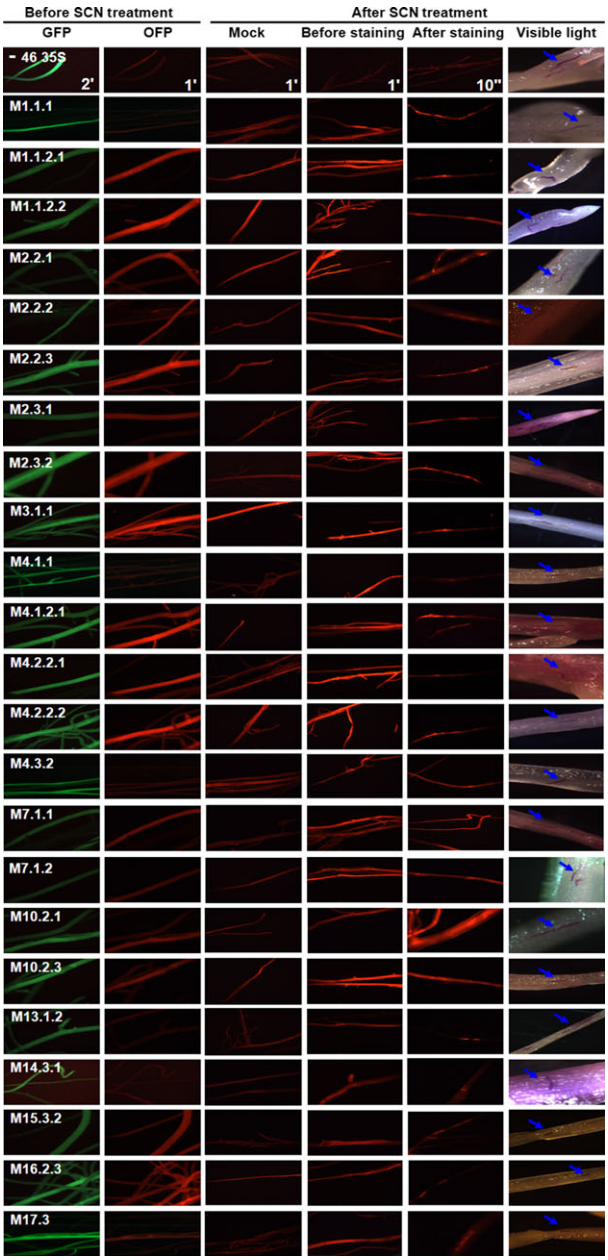

Figure 3 Inducibility of the 23 core motifs by co-localization of gain of function of *pporRFP* and the presence of soybean cyst nematode (SCN) in the transgenic hairy roots at 2 days postinfection (dpi). Before SCN treatment, *GFP* driven by a full-length CaMV 35S promoter was used for transgenic hairy root selection, while *pporRFP* driven by tetramers (i.e. four head-to-tail copies) of motifs was used for the indigenous expression of each motif. Plasmid -4635S *pporRFP*-35S *GFP* was used as a negative control. All the motifs were inserted into the 5′ end of -4635S in the binary vector pZP222 containing *35S::GFP*. Pictures were taken before SCN infection or at 2 dpi. Arrows indicate the presence of the nematodes stained in red. The SCN-infected transgenic hairy roots were cleared with 20% (v/v) bleach and stained by acid fuschin according to Byrd *et al.* (1983) with modifications so that the presence of nematodes and localized gain of function of the *pporRFP* reporter could be observed within the roots. The exposure time of GFP, orange fluorescent protein (OFP), mock, before staining and after staining was 2, 1, 1, 1 min (′) and 10 s (″), respectively.

they are mainly involved in the regulation of gene expression or defence as shown in the above-mentioned databases (Table 5). This further demonstrates the efficiency of our ensemble strategy for *de novo* motif discovery.

We also observed that computational tools from different algorithms have different abilities to predict the true overlapping motif regions and/or core motifs. In predicting the overall 116 OMRs within the 18 promoter regions, the enumerative tools (i.e. MDScan and Weeder) were as good as one of the two representative tools from either deterministic optimization (MEME) or probabilistic optimization (W-AlignACE; Figure 2). However, only one representative from each of the three algorithms showed better ability in predicting the true 23 SCN-inducible core motifs, that is, Weeder from enumerative algorithms (39.1%), MEME from deterministic expectation (17.3%) and W-AlignACE from probabilistic expectation (47.8%). The best tool in predicting the 23 core motifs in this study was SCOPE (56.6%), which uses three different algorithms such as BEAM, PRISM and SPACER. According to Tompa *et al.* (2005), the tool Weeder outperformed other tools such as MEME, AlignACE and MotifSampler. We found that the tool Weeder predicted fewer core motifs than W-AlignACE and SCOPE, which might be due to the ensemble strategy we applied, the algorithms each tool used and/or the default parameters used for bioinformatic analysis in this study. It was interesting to see that a combination of the three best computational tools (Weeder, W-AlignACE and SCOPE) could detect all 23 core motifs (Table S6). In addition, we collected the top 10 motifs detected by each tool. We found that, if only the top five motifs were collected from each tool, the combination of the three best tools (Weeder, W-AlignACE and SCOPE) could detect 21 out of the 23 core motifs (Table S6). Thus, a combination of these three computational tools along with the top five motifs collected from each tool is expected to be very effective in *de novo* motif discovery in the soybean genome. It is worthwhile to test whether these three tools still work efficiently on different promoter regions in other plant species.

Transgenic crop development could be greatly enabled by the ability to computationally design synthetic promoters. To date, all commercial transgenic crops use strong constitutive promoters, which have limited value for multitrait transgenics for which multiple unique promoters are needed. In addition, it would be of tremendous value to be able to computationally design very short (~100 bases) and strong inducible promoters for precise transgene expression (Liu *et al.*, 2013a). Synthetic promoters provide an effective means in testing motif functions (Rushton *et al.*, 2002; Venter, 2007). We previously demonstrated that synthetic promoters consisting of defence signalling-inducible *cis*-motifs fused to a fluorescent reporter could detect bacterial pathogen attacks in a transient phytosensing system and in stable transgenic plants (Fethe *et al.*, 2014; Liu *et al.*, 2011, 2013b,c; Mazarei *et al.*, 2008). In the present study, these novel motifs were assembled to produce strong and short synthetic SCN-inducible promoter (Figure 4).

Soybean is an economically important crop that provides a valuable source of protein and oil worldwide. However, soybean cyst nematode (SCN) is the most devastating pest of the soybean (Wrather and Koenning, 2006), as it feeds on the roots of soybean and limits soybean production. So far, the resistance improvement in soybean to fight against SCN mainly focuses on introducing SCN resistance genes or quantitative trait loci (QTLs)

Table 5 Known motifs in PlantCARE and PLACE databases that the 23 core motifs contain

Core motif	Sequence	Motifs in PlantCARE	Motifs in PLACE	Motif sequence	Motif function
1.1.1	TAAAGT	–	DOF-binding site; NtBBF1-binding site; TAAAG motif	AAAG; ACTTTA; TAAAG	DNA-binding proteins; Tissue-specific expression/auxin induction; Guard cell-specific gene expression
1.1.2.1	TCTTTA	–	DOF-binding site; TAAAG motif	AAAG; TAAAG	Regulate gene expression; Guard cell-specific gene expression
1.1.2.2	TTAATT	–	–	–	–
2.2.1	TAATTA	–	–	–	–
2.2.2	AAGTG	–	CACT motif; MYC-binding site	YACT; CANNTG	CACT motif; Regulates transcription of CBF/DREB1 in the cold
2.2.3	AGTGTTA	–	CACT motif	YACT;	CACT motif
2.3.1	ATATAA	TATA–box	–	–	–
2.3.2	ATTAAGT	–	–	–	–
3.1.1	AGGGA	–	–	–	–
4.1.1	AAAG	–	DOF-binding site	AAAG	Regulate gene expression
4.1.2.1	GAAAA	–	–	–	–
4.2.2.1	AAAGTG	–	CACT; DOF-binding site; MYC-binding site	YACT; AAAG; CANNTG	C4 Mesophyll-specific; Regulate gene expression; Regulates the transcription of CBF/DREB1 in the cold
4.2.2.2	TGGTG	–	–	–	–
4.3.2	GTAAAAAA	–	–	–	–
7.1.1	TTAGTT	–	(CA)n element; Homeodomain protein target site	CNAACAC; CTAATTGTTTA	Embryo- and endosperm-specific transcription of napin; Pathogenesis-related
7.1.2	GTTAAC	–	–	–	–
10.2.1	AGAGCG	–	–	–	–
10.2.3	GCGAGG	–	–	–	–
13.1.2	ATTATT	–	Plant polyA signal	AATAAT	Polyadenylation signal
14.3.1	GCTATTA	–	–	–	–
15.3.2	TTTTAA	–	–	–	–
16.2.3	AACAAA	–	–	–	–
17.2	GAGTGA	–	CACT motif; GTGA motif	YACT; GTGA	C4 Mesophyll-specific; Found in late pollen gene *g10*

from resistant sources into new breeding lines. We are also in the process of testing improvements to soybean–SCN resistance using transgenics containing potential resistance genes (unpublished data). There is a risk that overexpression of transgenic resistance genes might cause side effects on plant growth and metabolism. The discovery of SCN-inducible motifs in the soybean genome provides a better means to study conditional expression of transgenic resistance genes leading to improved SCN resistance in soybean. Thus, our ensemble strategy is a high-throughput approach for *de novo* motif discovery in soybean and offers great potential for novel motif discovery and synthetic promoter engineering for any plant and trait for which sufficient transcriptomics and genomics data exist.

Experimental procedures

Bioinformatics analyses for *de novo* SCN-inducible motif discovery

Our previous study detected 675 genes whose expression was significantly induced in the soybean genome during a susceptible soybean–SCN interaction using the Affymetrix Soybean GeneChip assay (Mazarei *et al.*, 2011). After having compared these induced genes with those published from other microarray datasets studying the susceptible soybean–SCN interaction (Ithal *et al.*, 2007a,b; Klink *et al.*, 2007a,b; Puthoff *et al.*, 2007), we found 49 common genes that existed both in our dataset (Mazarei *et al.*, 2011) and in at least one of the other five datasets (Ithal *et al.*, 2007a,b; Klink *et al.*, 2007a,b; Puthoff *et al.*, 2007; data not shown). Eighteen out of the 49 candidate genes were selected for *de novo* SCN-inducible motif discovery (Table 1). Through the Affymetrix website (https://www.affymetrix.com/site/login/login.affx) and GenBank, we used the probe IDs of these 18 candidate genes to download their nucleotide sequences that were used for probe design for microarray hybridization. Then, through the Phytozome website (http://www.phytozome.net/search.php?show=blast&blastdb=soybean), we downloaded the nucleotide sequences of 1-kb-long promoter regions (right before the translation initiation sites) of the 18 candidate genes for *de novo* motif discovery.

A total of seven widely used and user-friendly bioinformatic tools were chosen for *de novo* motif discovery, that is, MDscan (Liu *et al.*, 2002) and Weeder (Pavesi, 2004; enumeration), MEME (Bailey *et al.*, 2006) and BioProspector (Liu *et al.*, 2001; deterministic optimization), and MotifSampler (Thijs *et al.*, 2002) and W-AlignACE (Chen *et al.*, 2008; probabilistic optimization; Table 2). In addition, an ensemble method, SCOPE (Carlson *et al.*,

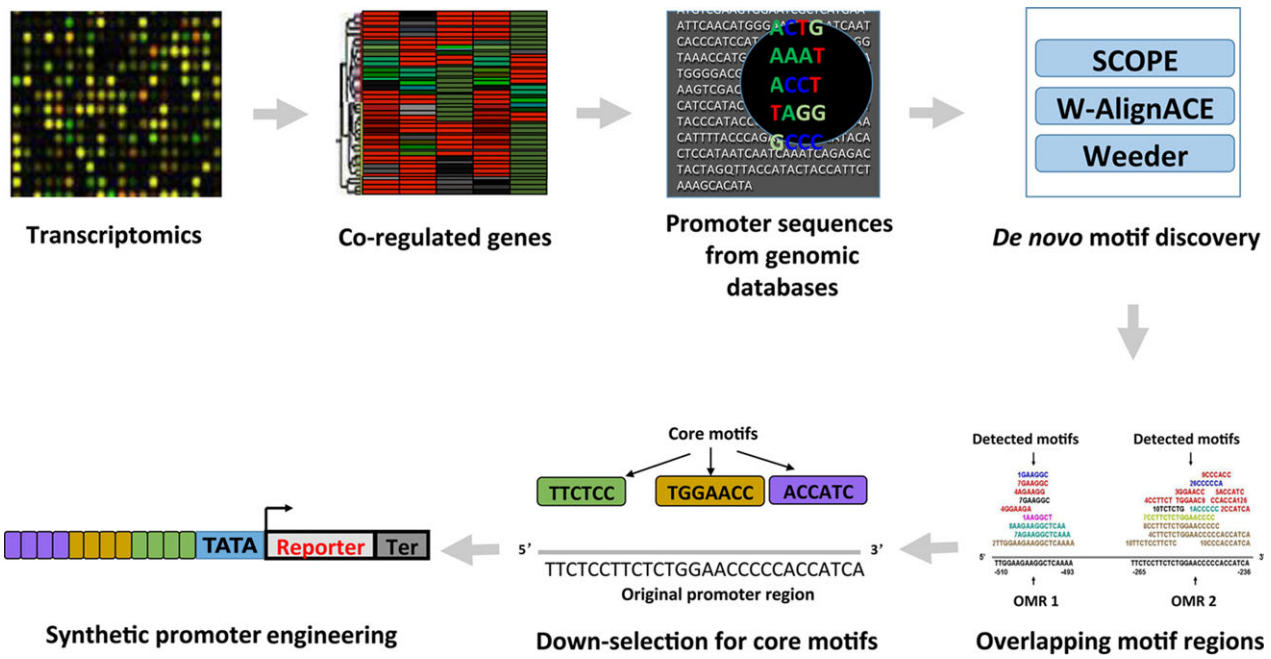

Figure 4 Scheme of an optimized ensemble strategy combining high-throughput transcriptomics data with *de novo* motif discovery for novel motif discovery and synthetic promoter engineering in crop plants. High-throughput gene expression profiling can be conducted using microarray (transcriptomics), and co-regulated genes can be deduced according to their similar mRNA expression profiles that are likely to be co-regulated via the same signal transduction pathways. The promoter sequences of co-regulated genes can be obtained from whole-genome sequence databases (genomics) and used for *de novo* motif discovery utilizing the three best tools (i.e. SCOPE, W-AlignACE and Weeder) as described here. Overlapping motif regions (OMRs) are collected for the down-selection of core motifs of 5–7 bp in length. These core motifs can be used in various orders and motif numbers together with a minimal promoter (i.e. one including TATA box) for synthetic promoter utilization in engineered crops. *pporRFP*, reporter gene; Ter, terminator.

2007; Chakravarty *et al.*, 2007), was also used in this study (Table 2). Each bioinformatic tool was used to search for both strands with default parameters unless otherwise specified in Table 2. Either input sequences or soybean intergenic sequences were used as background sequences. The top 10 ranked motifs from the output of each bioinformatic tool were chosen for this study.

Search for known motifs in the PlantCARE and PLACE databases

The detected core motif regions were used as query sequences to search for known motifs in the PlantCARE (Lescot *et al.*, 2002) and PLACE (Higo *et al.*, 1999) databases, which are the databases of known plant *cis*-acting regulatory elements.

Synthetic promoter construction

A *Spe*I site was inserted into the region between *Sac*I and *Not*I sites in plasmid pZP4×PR1 RFP (Liu *et al.*, 2013a,b,c) by PCR using primers pZP4×PR1-SacI-SpeI-F and pZP-BamHI-R (Figure 1b,c), followed by double restriction enzyme digestion with *Sac*I and *Bam*HI. The new construct was named pZPSpeI4×PR1-4635S RFP. Using plasmid pBIN-m-GFP5-ER (Haseloff *et al.*, 1997) as template, a fragment of the CaMV 35 promoter, m-GFP5-ER and NosT was fused together by five rounds of PCR amplification with primer pairs (i) p35S-HindIII-F and p35S-GFP-R, (ii) p35S-GFP-F and pGFP-Nos-R, (iii) pGFP-Nos-F and pNos-HindIII-R, (iv) p35S-GFP-F and pNos-HindIII-R, and (v) p35S-HindIII-F and pNos-HindIII-R, respectively (Figure 1b,c). At the same time, the restriction sites for *Bam*HI and *Sac*I were removed from that fragment. This PCR fragment was purified and inserted into the *Hind*III site of plasmid pZPSpeI4×PR1-4635S RFP to make a construct, pZPSpeI4×PR1-4635S RFP-35S GFP.

The seven 1-kb-long soybean promoter regions were PCR-amplified from genomic DNA of soybean line TN02-275 (Mazarei *et al.*, 2011; for primer sequences, see Table S2) and used to replace the 4×PR1-4635S fragment in plasmid pZPSpeI4×PR1-4635S RFP-35S GFP with the help of restriction enzymes *Spe*I and *BamH*I (Figure 1b). The primer dimer of each tetramerized motif region, as well as the motif alone, was synthesized via fusion of two primers that were reverse complementary to each other. Then, each primer dimer was digested with *Xba*I + *Spe*I and inserted into *Spe*I-digested pZPSpeI4×PR1-4635S RFP-35S GFP (Figure 1c).

Plants

Soybean line TN02-275, which is susceptible to SCN race 2 (HG type 1.2.5.7; Mazarei *et al.*, 2011), was used in this study. Sterile soybean seeds were germinated in sealed sterile Petri dishes for 3 days. Then, seedlings were transferred into sterilized vermiculite for another 4 days in a growth chamber at 25 °C under fluorescent white light with a 16/8 h light/dark cycle.

Generation of transgenic soybean hairy roots

Transgenic soybean hairy roots were generated as described (Kereszt *et al.*, 2007; Lin *et al.*, 2013) with some modifications. *Agrobacterium rhizogenes* strain K599 was transformed with each individual construct by electroporation. *Agrobacterium rhizogenes* containing the individual constructs was grown on yeast extract peptone [(YEP), 10 g/L yeast extract, 10 g/L peptone, 5 g/L NaCl, 15 g/L agar] solid medium supplemented with spectinomycin (200 mg/L) at 28 °C for 2 days. One single

colony was inoculated into 250 μL YEP liquid medium in duplicate, spread onto two YEP solid medium supplemented with the above-mentioned antibiotics and grown for ~2 days at 28 °C. The cultured bacterial lawn was collected and suspended in 1 mL of sterile distilled water. Bacterial suspensions were injected three times into 1-week-old soybean cotyledonary nodes and upper hypocotyls with unfolded cotyledons using a 3-mL needle. After injection, soybean seedlings were covered with transparent plastic covers, which were sprayed with water, and maintained in a growth chamber for 1 week. Then, the plastic covers were removed and the *A. rhizogenes* wounding sites were covered with sterile vermiculite. Twenty-five biological replicates (i.e. 25 individual plants) were used for each construct.

Three weeks later, the hairy roots grew to approximately 10 cm in length. The tap roots were excised, and transgenic soybean hairy roots harbouring each synthetic promoter were screened for GFP expression with an epifluorescent microscope (Olympus stereo microscope model SZX12; Olympus America, Center Valley, PA) using a GFP filter set: 475/30 nm excitation and 535/40 nm band-pass emission and QCapture 2.56 imaging software, and an Olympus Q-colour 5 camera (Olympus, Center Valley, PA).

Nematode source

Soybean cyst nematode race 2 (HG type 1.2.5.7), which was originally collected from soybean field in Beaufort County, NC, was cultured in the greenhouse under controlled conditions of temperature and light, and maintained on the roots of cv. Pickett-71 (Hartwig *et al.*, 1971) before being used for inoculum preparation (Arelli *et al.*, 2000).

Nematode infection and tissue harvesting

Transgenic soybean hairy roots harbouring each synthetic promoter were loaded horizontally in a 13 × 9 × 2 cm sterilized inoculating tray containing a thin layer of a mixture of sterile sand and top soil (1 : 1). About 10 mL of inoculum, which contained about 66 000 SCN eggs, was added to each inoculating tray. SCN eggs were allowed to hatch and infect soybean roots for 7 days under humid conditions. All the roots were then taken out and washed to remove extra SCN eggs and juvenile nematodes that had not penetrated the root tissues. The infected chimeras were grown in containers with sterile vermiculite in a growth chamber. Infected transgenic hairy roots were cleaned with 20% (v/v) bleach for about 4–7 min and then stained by acid fuschin for the detection of nematodes (Byrd *et al.*, 1983) at time points 0.5, 1, 2, 3, 10, 17 and 24 days postinfection (dpi).

Wounding treatments

Wounding was performed by repeatedly piercing the 4-week-old transgenic soybean hairy roots containing each construct with a needle. The wounded plants were incubated in B&D solution (Byrd *et al.*, 1983) for 24 h. Unwounded transgenic soybean hairy roots were incubated in B&D solution for 24 h as mock control.

Determination of gain of function of *pporRFP* expression

Expression of the orange fluorescence reporter gene, *pporRFP,* was visualized with an epifluorescent microscope (Olympus stereo microscope model SZX12; Olympus America) and QCapture 2.56 imaging software, and an Olympus Q-colour 5 camera (Olympus). A tdTomato filter set (535/30 nm excitation and 600/50 nm band pass emission) was used for the visualization of *pporRFP* expression. Time-course analysis of expression of the *pporRFP* reporter was conducted at time points 0.5, 1, 2, 3, 10, 17 and 24 dpi prior to bleach washing and acid fuschin staining.

Genomic DNA extraction

Total genomic DNA of soybean line TN02-275 was isolated from the leaves of 3-week-old plants using a CTAB method. Purity and concentration of genomic DNA was determined at wavelengths of 260 and 280 nm using a NanoDrop ND-1000 spectrophotometer (Wilmington, DE).

Acknowledgements

We gratefully acknowledge funding by grants from USDA-NIFA. We thank Dr. Priya Ranjan for downloading the whole intergenic sequences of soybean, Dr. Robert Gross for creating the soybean background for SCOPE and Dr. Giulio Pavesi for creating the soybean background for Weeder.

Conflict of interests

The authors declare competing financial interests: W.L., M.M. and C.N.S.Jr. have filed for a patent on SCN-inducible motif regions discovered in this article.

References

Alieva, N.O., Konzen, A.K.A., Field, S.F., Meleshkevitch, E.A., Hunt, M.E., BeltranRamirez, V., Miller, D.J., Wiedenmann, J., Salih, A. and Matz, M.V. (2008) Diversity and evolution of coral fluorescent proteins. *PLoS ONE*, **3**, e2680.

Altarawy, D., Ismail, M.A. and Ghanem, S.M. (2009) MProfiler: a profile-based method for DNA motif discovery. In *Pattern Recognition in Bioinformatics*, Vol. 5780/2009, pp. 13–23. Lecture Notes in Computer Science, Heidelberg: Springer Berlin.

Altman, R.B. and Raychaudhuri, S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* **11**, 340–347.

Ao, W., Gaudet, J., Kent, W.J., Muttumu, S. and Mango, S.E. (2004) Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, **305**, 1743–1746.

Arelli, P.R., Sleper, D.A., Yue, P. and Wilcox, J.A. (2000) Soybean reaction to races 1 and 2 of *Heterodera glycines*. *Crop Sci.* **40**, 824–826.

Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373.

Byrd, D.W. Jr., Kirkpatrick, T. and Barker, K.R. (1983) An improved technique for clearing and staining plant tissue for detection of nematodes. *J. Nematol.* **14**, 142–143.

Carlson, J.M., Chakravarty, A., DeZiel, C.E. and Gross, R.H. (2007) SCOPE: a web server for practical de novo motif discovery. *Nucleic Acid Res.* **35**, W259–W264.

Chakravarty, A., Carlson, J.M., Khetani, R.S. and Gross, R.H. (2007) A novel ensemble learning method for *de novo* computational identification of DNA binding sites. *BMC Bioinformatics*, **8**, 249.

Chen, X., Guo, L., Fan, Z. and Jiang, T. (2008) W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics*, **24**, 1121–1128.

Davis, E.L., Hussey, R.S. and Baum, T.J. (2004) Getting to the roots of parasitism by Nematodes. *Trends Parasitol.* **20**, 134–141.

D'haeseleer, P. (2006) How does DNA sequence motif discovery work? *Nat. Biotechnol.* **24**, 959–961.

Endo, B.Y. (1991) Ultrastructure of initial responses of susceptible and resistant soybean roots to infection by *Heterodera glycines*. *Rev. Nématol.* **14**, 73–94.

Fethe, M.H., Liu, W., Mazarei, M., Burris, J.N., Rudis, M.R., Millwood, R.J., Yeaman, D.G., Dubosquielle, M. and Stewart, C.N. Jr. (2014) The performance of pathogenic bacterial phytosensing transgenic tobacco in the field. *Plant Biotechnol. J.* doi: 10.1111/pbi.12180. (In print).

Gheysen, G. and Mitchum, A.G. (2009) Molecular insights in the susceptible plant response to nematode infection. In *Cell Biology of Plant Nematode Parasitism* (Berg, H.R. and Christopher, G.T., eds), pp. 45–81. Berlin: Springer-Verlag.

Hartwig, E.E., Epps, J.M. and Edwards, C.J. Jr. (1971) Registration of 'Pickett 71' soybean. *Crop Sci.* **11**, 603.

Haseloff, J., Siemering, K.R., Prasher, D.C. and Hodge, S. (1997) Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic Arabidopsis plants brightly. *Proc. Natl Acad. Sci. USA*, **94**, 2122–2127.

van Helden, J., Andre, B. and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842.

van Helden, J., Rios, A.F. and Collado-Vides, J. (2000) Discovering regulatory elements in noncoding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **28**, 1808–1818.

Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* **27**, 297–300.

Hu, J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.* **33**, 4899–4913.

Ithal, N., Recknor, J., Nettleston, D., Hearne, L., Maier, T., Baum, T.J. and Mitchum, M.G. (2007a) Parallel genome-wide expression profiling of host and pathogen during soybean cyst nematode infection of soybean. *Mol. Plant Microbe Interact.* **20**, 293–305.

Ithal, N., Recknor, J., Nettleton, D., Maier, T., Baum, T.J. and Mitchum, M.G. (2007b) Developmental transcript profiling of cyst nematode feeding cells in soybean. *Mol. Plant Microbe Interact.* **20**, 510–525.

Kereszt, A., Li, D., Indrasumunar, A., Nguyen, C.D., Nontachaiyapoom, S., Kinkema, M. and Gresshoff, P.M. (2007) *Agrobacterium rhizogenes*-mediated transformation of soybean to study root biology. *Nat. Protoc.* **2**, 948–952.

Kim, Y.H., Riggs, R.D. and Kim, K.S. (1987) Structural changes associated with resistance of soybean to *Heterodera glycines*. *J. Nematol.* **19**, 177–187.

Klink, V.P., Overall, C.C., Alkharouf, N., MacDonald, M.H. and Matthews, B.F. (2007a) Laser capture microdissection (LCM) and comparative microarray expression analysis of syncytial cells isolated from incompatible and compatible soybean roots infected by soybean cyst nematode (*Heterodera glycines*). *Planta*, **226**, 1389–1409.

Klink, V.P., Overall, C.C., Alkharouf, N., MacDonald, M.H. and Matthews, B.F. (2007b) A comparative microarray analysis of an incompatible and compatible disease response by soybean (*Glycine max*) to soybean cyst nematode (*Heterodera glycines*) infection. *Planta*, **226**, 1423–1447.

Koschmann, J., Machens, F., Becker, M., Niemeyer, J., Schulze, J., Bulow, L., Stahl, D.J. and Hehl, R. (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in Arabiopsis. *Plant Physiol.* **160**, 178–191.

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

Li, N. and Tompa, M. (2006) Analysis of computational approaches for motif discovery. *Algorthims Mol. Biol.* **1**, 8.

Lin, J., Mazarei, M., Zhao, N., Zhu, J., Zhuang, X., Liu, W., Pantalone, R.V., Arelli, P.R., Stewart, C.N. Jr. and Chen, F. (2013) Overexpression of a soybean salicylic acid methyltransferase gene confers resistance to soybean cyst nematode. *Plant Biotechnol. J.* **11**, 1135–1145.

Liu, X., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* **6**, 127–138.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839.

Liu, W., Mazarei, M., Rudis, M.R., Fethe, M.H. and Stewart, C.N. Jr. (2011) Rapid *in vivo* analysis of synthetic promoters for plant pathogen phytosensing. *BMC Biotechnol.* **11**, 108.

Liu, W., Yuan, J.S. and Stewart, C.N. Jr. (2013a) Advanced tools for plant biotechnology. *Nat. Rev. Genet.* **14**, 781–793.

Liu, W., Mazarei, M., Rudis, M.R., Fethe, M.H., Peng, Y., Millwood, R., Shoene, G., Burris, J.N. and Stewart, C.N. Jr. (2013b) Bacterial pathogen phytosensing in transgenic tobacco and *Arabidopsis*. *Plant Biotechnol. J.* **11**, 43–52.

Liu, W., Rudis, M.R., Peng, Y., Mazarei, M., Millwood, R.J., Yang, J.-P., Xu, W., Chesnut, J.D. and Stewart, C.N. Jr. (2013c) Synthetic TAL effectors for targeted enhancement of transgene expression in plants. *Plant Biotechnol. J.* **12**, 436–446.

Mann, D.G.J., Abercrombie, L.L., Rudis, M.R., Millwood, R.J., Dunlap, J.R. and Stewart, C.N. Jr. (2012) Very bright orange fluorescent plants: endoplasmic reticulum targeting of orange fluorescent proteins as visual reporters in transgenic plants. *BMC Biotechnol.* **12**, 17.

Matys, V., Fricke, E., Geffers, R., ßling, E.G., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378.

Mazarei, M., Teplova, I., Hajimorad, M.R. and Stewart, C.N. Jr. (2008) Pathogen phytosensing: plants to report plant pathogens. *Sensors*, **8**, 2628–2641.

Mazarei, M., Liu, W., Al-Ahmad, H., Arelli, P.R., Pantalone, V.R. and Stewart, C.N. Jr. (2011) Gene expression profiling of resistant and susceptible soybean lines infected with soybean cyst nematode. *Theor. Appl. Genet.* **12**, 1193–1206.

Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**, W199–W203.

Puthoff, D.P., Ehrenfried, M.L., Vinyard, B.T. and Tucker, M.L. (2007) GeneChip profiling of transcriptional responses to soybean cyst nematode, *Heterodera glycines*, colonization of soybean roots. *J. Exp. Bot.* **58**, 3407–3418.

Quest, D., Dempsey, K., Shafiullah, M., Bastola, D. and Ali, H. (2008) MTAP: the motif tool assessment platform. *BMC Bioinformatics*, **9**, S6.

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

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays – a technology review. *Nat. Cell Biol.* **3**, E190–E195.

Sinha, S. and Tompa, M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **31**, 3586–3588.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2002) A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* **9**, 447–464.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144.

Venter, M. (2007) Synthetic promoters: genetic control through *cis* engineering. *Trends Plant Sci.* **12**, 118–124.

Wijaya, E., Yiu, S.-M., Son, N.T., Kanagasabai, R. and Sung, W.-K. (2008) MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics*, **24**, 2288–2295.

Workman, C.T. and Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *Pacific Symposium on Biocomputing* (Altman, R., Dunker, A.K., Hunter, L. and Klein, T.E., eds), pp. 467–478. Stanford, CA: Stanford University.

Wrather, J.A. and Koenning, S.R. (2006) Estimates of disease effects on soybean yields in the United States 2003 to 2005. *J. Nematol.* **38**, 173–180.

Yanover, C., Singh, M. and Zaslavsky, E. (2009) M are better than one: an ensemble-based motif finder and its application to regulatory element prediction. *Bioinformatics*, **25**, 868–874.

Yu, H., Luscombe, N., Oian, J. and Gerstein, M. (2003) Genomic analysis of essentiality within protein networks. *Trends Genet.* **19**, 422–427.

Zhou, Q. and Wong, W.H. (2004) CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA*, **101**, 12114–12119.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Inducibility of seven promoter regions by co-localization of the presence of SCN and localized gain of function of the OFP reporter at 3 days postinfection (dpi).

Figure S2 Display of detected overlapping motif regions as well as core motifs whose SCN inducibility was confirmed experimentally within the promoter regions of Glyma20g19200.1 (A) and Glyma13g41160.1 (B).

Table S1 Number of bioinformatic tools that detected each overlapping motif region (OMR).

Table S2 Primer sequences used for PCR amplification of the seven 1-kb-long promoter regions.

Table S3 Time-course analysis of inducibility of the seven promoters by SCN treatment.

Table S4 Capability of the seven bioinformatic tools in discovery of the 16 overlapping motif regions (OMRs) within the two promoter regions of Glyma20g19200.1 and Glyma13g41160.1.

Table S5 Time-course analysis of inducibility of the 31 narrower overlapping motif regions (OMRs) within the promoter regions of Glyma20g19200.1 and Glyma13g41160.1 by SCN treatment.

Table S6 Capability of the seven bioinformatic tools in discovery of the 23 core motifs.