

TECHNICAL ADVANCE

New BAR tools for mining expression data and exploring *Cis*-elements in *Arabidopsis thaliana*

Ryan S. Austin[†], Shu Hiu, Jamie Waese, Matthew Ierullo, Asher Pasha, Ting Ting Wang, Jim Fan, Curtis Foong, Robert Breit, Darrell Desveaux, Alan Moses and Nicholas J. Provart*

Department of Cell & Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada

Received 14 April 2016; revised 23 June 2016; accepted 1 July 2016; published online 5 October 2016.

*For correspondence (e-mail nicholas.provart@utoronto.ca).

[†]Present address: Agriculture & Agri-Food Canada, London Research & Development Center, London, ON N5V 3V3, Canada.

SUMMARY

Identifying sets of genes that are specifically expressed in certain tissues or in response to an environmental stimulus is useful for designing reporter constructs, generating gene expression markers, or for understanding gene regulatory networks. We have developed an easy-to-use online tool for defining a desired expression profile (a modification of our Expression Angler program), which can then be used to identify genes exhibiting patterns of expression that match this profile as closely as possible. Further, we have developed another online tool, Cistome, for predicting or exploring *cis*-elements in the promoters of sets of co-expressed genes identified by such a method, or by other methods. We present two use cases for these tools, which are freely available on the Bio-Analytic Resource at <http://BAR.utoronto.ca>.

Keywords: coexpression analysis, *cis*-element prediction, gene expression markers, reporter constructs, promoter analysis, *Arabidopsis thaliana*, technical advance.

INTRODUCTION

Coexpression analysis, the identification of genes exhibiting similar expression patterns in different tissues or in response to different perturbations, is a powerful method for gene function hypothesis generation in plant biology (Usadel *et al.*, 2009). Several groups have published coexpression networks for *Arabidopsis thaliana*, for example the condition-independent AraNet (Lee *et al.*, 2010), or condition-specific networks like SeedNet (Bassel *et al.*, 2011), FlowerNet (Pearce *et al.*, 2015) and 'BioticStressNet' (Amrine *et al.*, 2015). Often such condition-specific coexpression analyses can provide more insight into how a biological system functions. For instance, most genes identified as being involved in seed biology in SeedNet were not uncovered in the condition-independent AraNet. While condition-dependent networks can provide more precise hypotheses about gene function, generating such coexpression networks requires a fair degree of computational resources and expertise. A simpler coexpression approach might be appropriate in order to identify promoters for use in generating reporter constructs or to be

able to examine whether the promoters of coexpressed genes contain certain *cis*-elements in common. In this case online coexpression tools, such as Expression Angler (Toufighi *et al.*, 2005), ATTED-II (Obayashi *et al.*, 2011) and CressExpress (Srinivasasainagendra *et al.*, 2008), will often provide lists of similarly expressed genes at the click of a mouse. Such tools require a guide or query gene as input, whose expression profile is used to identify other genes with closely related expression profiles as scored by a similarity metric, such as the Pearson correlation coefficient (Expression Angler and CressExpress) or Mutual Rank (ATTED-II). Often, however, a specific pattern may be desired in order to identify genes that might be useful as gene expression biomarkers, either as positive controls in RT-PCR experiments or to identify promoter candidates for driving reporter gene expression. While Genevestigator has a gene search tool (see https://genevestigator.com/gv/file/GENEVESTIGATOR_UserManual.pdf, section 3.2), the selection process for biomarker identification using this tool involves a binary filter, such that the genes

identified with it are expressed above a preset threshold in the 'target' tissues but are not strongly expressed in 'base' tissues (target and base are the terms used by Genevestigator to define tissues exhibiting strong and weak levels of expression for the identified genes, respectively). Unfortunately, no tool currently provides a means to query a user-defined expression pattern, such as a slowly-increasing expression level in a specific tissue, for instance. We present here a flexible method of defining any user-desired expression pattern as a 'custom bait' in the Expression Angler tool for searching the BAR's extensive gene expression databases to identify sets of genes that most closely match the user-defined pattern of expression.

While such sets of coexpressed genes can be useful for gene discovery under the 'guilt by association' paradigm, they are also useful for exploring transcriptional regulation under the assumption that genes with similar patterns of expression are regulated by the same transcription factors (TFs) and thus binding sites for these should be common to their promoters. Recent high-throughput methods have determined or predicted the transcription factor binding specificities captured as position-specific weight matrices (TFBMs) of 745 Arabidopsis TFs (Weirauch *et al.*, 2014). Additionally, there is a relatively large literature of *in vivo* promoter analyses and TF binding site assays. For instance, there are 48 experimentally characterized *A. thaliana* TFBMs in the JASPAR database of Mathelier *et al.* (2013), see http://jaspar.genereg.net/cgi-bin/jaspar_db.pl?select1=Species&selectfield1=3702&rm=select, which can (as with the Weirauch *et al.*, 2014 data) also be used to explore a set of promoters from coexpressed genes for *cis*-elements in common. Finally, there is also a good if somewhat older collection of functionally active promoter sequences available in the PLACE database (Higo *et al.*, 1998).

We have enabled the exploration and analysis of sets of Arabidopsis promoters with a second online tool called Cistome. With this tool it is possible to ask for a set of promoters (potentially identified with the custom bait feature of Expression Angler, described above) if there is enrichment for particular motifs (used hereafter to refer to potential *cis*-regulatory elements) in the Weirauch *et al.* (2014) data set, in the Arabidopsis subset of the JASPAR database, in PLACE, or in our own set of motif predictions from a computational pipeline that integrates five well-cited prediction programs and a novel enumerative strategy in the promoters of coexpressed gene sets we identified using the custom bait approach. We demonstrate *in planta* that one of our predicted motifs directs reporter gene expression in a known manner. Two use cases of the Expression Angler custom bait feature to identify marker genes for genotoxic stress and different kinds of pathogen response are presented.

RESULTS AND DISCUSSION

The 'custom bait' feature of Expression Angler for identifying sets of transcripts with any specified expression profile

Figure 1 shows a screenshot of the Expression Angler interface for designing a custom bait with which to query the BAR's gene expression compendia. Users can opt to select only a subset of the samples to search in, or they can set the expression level to be minimal in all samples. Tissue samples in which genes should exhibit higher expression levels are then chosen, and the desired expression level in a given tissue is set by clicking on the corresponding part of the image – these images are displayed in a manner similar to the BAR's widely used 'electronic fluorescent pictograph' (eFP) browser (Winter *et al.*, 2007) – and setting the level with a slider. Once a desired pattern has been specified, the user clicks 'Search' and the Expression Angler engine (Toufighi *et al.*, 2005) uses the pattern (which effectively is a vector with length equal to the number of samples specified) to search for genes with similar expression profiles (as measured using the Pearson correlation coefficient), as shown in the inset in Figure 1. The results page allows users to download expression data for either the top 25 or 50 best expression pattern matches or for those exceeding a specified *r*-value threshold. These data may also be viewed as heatmaps within the interface.

Use case 1: using 'custom baits' to identify new pathogenesis-induced gene expression markers

The plant defence system can be divided into two major types of immunity, pathogen/microbe-associated molecular pattern (PAMP)-triggered immunity (PTI) and effector-triggered immunity (ETI). PTI is the first level of the plant immune system. It functions by using transmembrane pattern recognition receptors (PRRs) to recognize portions of microbial molecules called PAMPs found on both pathogenic and non-pathogenic microbes (Zipfel and Felix, 2005). For example, the PRR FLS2 is a leucine-rich repeat receptor kinase that recognizes a conserved 22 amino acid segment, Flg22, in bacterial flagellin (Gómez-Gómez and Boller, 2000). This pattern recognition results in a defence cascade that results in PTI (Jones and Dangl, 2006). Successful invaders are able to suppress PTI through the use of effectors that dampen the plant's PTI-inducing defences or increase virulence to ensure successful colonization (Nomura *et al.*, 2005). Gram-negative bacterial pathogens can make use of the type III secretion system to inject various effectors into plant cells (Nomura *et al.*, 2005), leading to effector-triggered susceptibility (ETS). Disease-resistance R genes that encode proteins with nucleotide-binding and leucine-rich repeat domains recognize specific effectors (Dangl and Jones, 2001). The result of this induces ETI, which results in a hypersensitive response

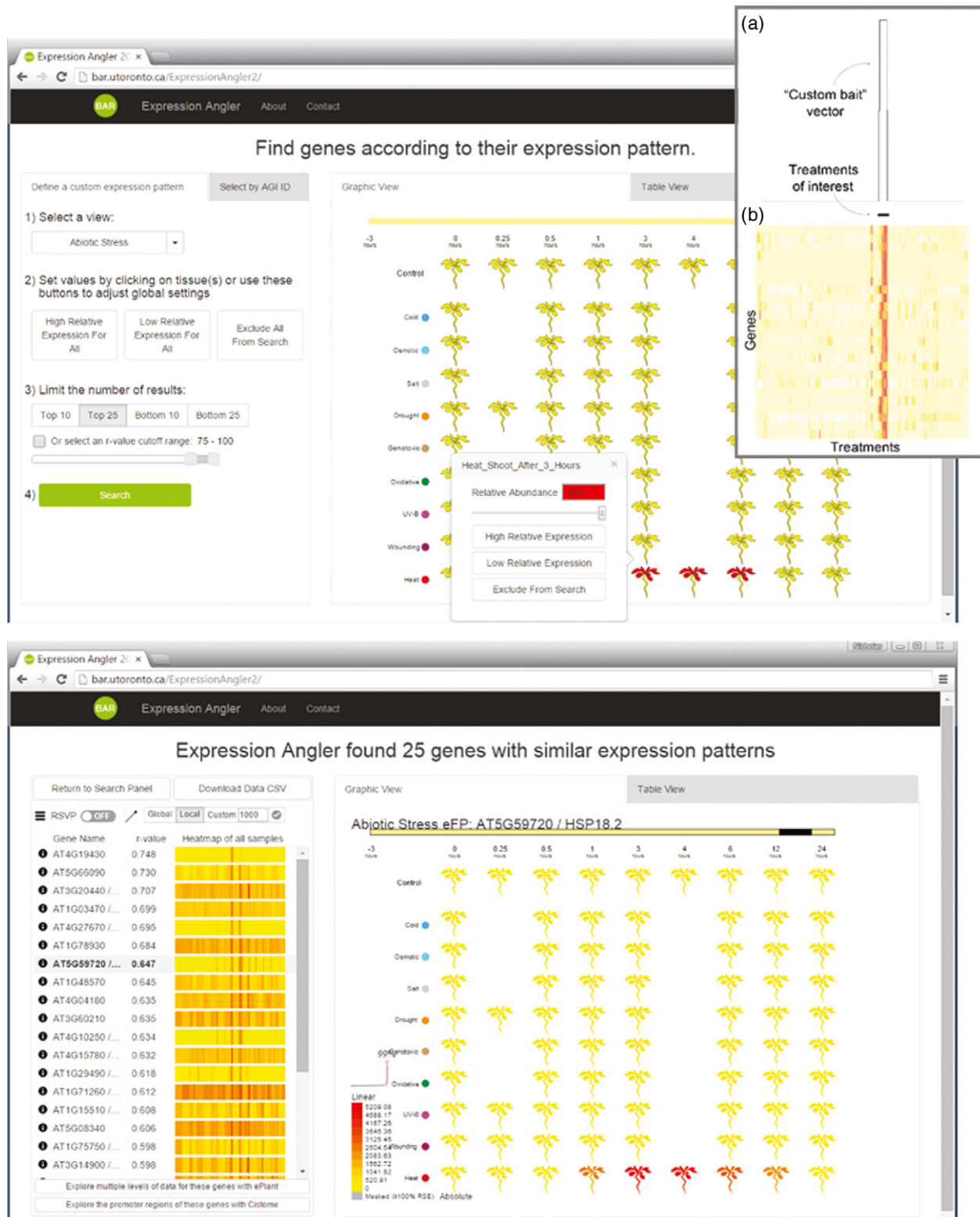


Figure 1. Defining a desired expression pattern ('custom bait') with Expression Angler. Here, a high level of expression (100 expression units; 100-fold above the baseline level, which was set to 1) in shoots of plants 3, 4 and 6 h after an applied heat stress is set to be the desired pattern with which to search the BAR's Abiotic Stress expression compendium containing data from Kilian *et al.* (2007). Inset shows an example custom bait profile/vector generated by the Expression Angler algorithm after query submission (a) in order to identify genes exhibiting such expression patterns (b). Bottom panel shows the results of the heat stress query, with the expression pattern for At5g59720/HSP18.2 (a gene encoding a heatshock protein), highlighted.

(Greenberg and Yao, 2004). An arms race between the pathogen and host ensues as the pathogen tries to avoid the immune system by gaining, losing and/or modifying effectors while the host attempts to maintain effector detection.

Pathogenesis-related (PR) proteins have been used as markers to characterize numerous aspects of plant defence in several plant species, including *Arabidopsis*. PR proteins were originally identified through SDS-PAGE analysis of proteins from infected tobacco plants to identify proteins that are strongly expressed during infection (van Loon and van Kammen, 1970). While the use of PR genes as markers has advanced our understanding of plant-pathogen interactions immensely, it would be desirable to have markers that are reflective of the infection outcome; PTI, ETI or susceptibility/pathogen induction. We thus designed several 'custom baits' in Expression Angler to identify sets of genes exhibiting responses to treatments designed to elicit these outcomes: PAMP-induced genes (PIGs), ETI-induced genes (EIGs) and disease-induced genes (DIGs), representing genes whose expression is induced under PTI-related, ETI-related and disease-related conditions, respectively (see Experimental procedures).

Disease-induced genes would be specifically induced by virulent *Pseudomonas syringae* pv. tomato DC3000 (as a representative of susceptibility), whereas EIGs would be induced by avirulent *P. syringae* pv. tomato DC3000 AvrRpm1, which is recognized by the R protein RPM1 to induce ETI. PIGs would be induced under three PTI inducing conditions: (i) non-virulent *P. syringae* pv. tomato DC3000::hrcC, which contains a defect in the type III secretion system rendering it unable to transfer effectors to suppress PTI; (ii) the non-host *P. syringae* pv. phaseolicola 14 48A, which is not able to infect *A. thaliana* plants because PTI is induced; and (iii) the Flg22 portion of bacterial flagellum, a well-characterized PAMP. Using custom baits and subsequent filtering, we identified approximately 20 genes that could be considered DIGs, EIGs or PIGs (see Experimental procedures). Because we wanted to use these as expression markers in RT-PCR-based assays, we undertook RT-PCR on RNA samples from plants treated in the same

way that plants used to generate the gene expression data for the Biotic Stress – *P. syringae* or – Elicitor views in Expression Angler were. A total of seven genes (see gene-specific primers in Table 1) were confirmed to possess the desired expression patterns of DIGs, EIGs or PIGs suitable for use as RT-PCR gene expression markers based on genes identified using the 'custom bait' feature of Expression Angler, as shown in Figure 2.

Tissue- and abiotic stress response-specific transcripts identified using custom baits

The custom bait approach described above appeared to be a useful way of identifying sets of genes, both for use as gene expression markers and as input for promoter motif analysis programs. In order to leverage this approach further, three large expression compendia from the BAR were used to generate a series of coexpression clusters with profiles specific to distinct tissues or stress responses. These consisted of: (i) 20 487 transcripts across 370 microarray experiments for tissue-specific expression; (ii) 20 172 transcripts across 272 microarray experiments in response to abiotic stress; and (iii) 21 003 transcripts over 230 microarray experiments for hormone treatments of wild-type seedlings. For details regarding the metadata (i.e. perturbation, type/age of tissue, time of sampling) specific to each compendium, see Experimental procedures, and Tables S1 and S2.

Metadata from each of these expression compendia were used to generate a series of coexpression sets that showed increases in expression levels for all genes in the set under a single condition. Briefly, the custom bait approach described above was used to construct pseudo-expression profiles, equal in length to the number of treatments or tissues in the expression compendium. These baits possessed a pattern of increased expression specific to the treatment or tissue of interest (see depiction of one such vector in Figure 1). This 'bait' profile was then used to identify transcripts within each compendium that matched its behaviour with a high degree of correlation ($r > 0.75$; genes with r -values of 0.75 or higher are significant with a P -value of at least 8.27×10^{-43} in the case of

Table 1 Primers used for PCR amplification of cDNA products for pathogenesis-induced gene expression markers

AGI ID	Marker for:	Forward (5'–3')	Reverse (5'–3')
At1g19640	ETS	AGGAGGGTTTAGGGTTCGGT	TCGGAGCTCGCAGCATAGTA
At3g47340	ETS	CTCTTCCTGGACATCTGTCTGT	TGACACCAATCGCATCACGA
At1g77450	ETI	CTCTGTCGTAAATGCGCGTC	TCACAGGCTTTAACCCGTCC
At1g51800	PTI	GATTGCGGTTGCGCTAGAGA	CGGAGATAAAAGGCGTTGCG
At1g51850	PTI	AACGTTGGGAAACCCGGTAG	TGTGTCCATGAGTTGTGGAAGT
At2g44370	PTI	AAGAAAACCGTCGGTGAGGC	TGCGTGCCATAATCACATTCC
At3g46280	PTI	TCTCGCCGCCATCTTTTGAT	TCTCTGGTGGCTTGTCGAC

ETS, effector-triggered susceptibility; ETI, effector-triggered immunity; PTI, pathogen/microbe-associated molecular pattern-triggered immunity.

Treatment	Time (h)	At1g19640	At3g47340	At1g77450	At1g51850	At2g44370	At3g46280	At1g51800	PR1	UBQ10
No treatment										
MgCl2	6	57.6	51.7	124.5	28.7	11.5	198.5	182.2	57.4	2062.3
	24	18.8	47.6	27.2	10.8	5.8	43.2	77.2	236.4	2547.6
DC3000	6	676.1	57.7	191.2	28.3	21.8	185.3	113.3	103.5	1781.7
	24	1744.9	1069.5	210.4	4.4	2.4	52.6	65.1	501.9	2068.7
<i>hrcC</i>	24	124.7	37.6	31.3	353.9	343.8	850.6	863.1	1262	2135.2
Phaseolicola	6	41.9	47.8	90.9	702.1	886	1812.4	1141.8	286	2073.7
	24	30.3	85.7	30.2	298	392.2	1090.5	690.6	2693.4	2177.6
AvrRpm1	6	243.5	188	728.9	11.9	10.6	259.8	268.6	102.6	2022.5
	24	191.2	341.5	108.4	6.6	2.6	158.7	217.5	2226.7	2124.6
No treatment										
H2O	1	244.5	71.7	175.1	46.7	18.8	125.8	304.7	30.7	2690.8
	4	28.9	9.6	44.3	5.3	1.6	24.1	50.7	32.7	2153.5
Flg22	1	204.9	51.4	162	405.3	426.9	756.2	1215.4	22.8	3361.1
	4	35.4	7	76	1406.9	946.9	1721.8	1258.3	34.7	2160.8

Figure 2. Gene expression markers for different types of plant–pathogen interactions identified using the custom bait feature of Expression Angler. Blue: markers for pathogen-associated molecular pattern-triggered immunity (PTI); pink: markers for virulent/susceptible (ETS) interactions; green: marker for effector-triggered immunity (ETI). Numbers with coloured backgrounds represent expression levels from the AtGenExpress Biotic Stress series, with a red background denoting 100% of the maximal column expression and yellow denoting low expression. Data for the PR-1 expression marker commonly used for pathogen studies are also included. Gel images represent RT-PCR results from experiments conducted using the same conditions as for the AtGenExpress series. UBQ10 is a loading control. RT-PCR gels were imaged on a gene by gene basis, with each gene’s PCR reactions across all conditions run on a single gel.

the 230 microarray experiment Hormone compendium, or better in the case of the two larger compendia, as calculated using equation 3 of Usadel *et al.*, 2009). A total of 67 custom baits were used with the Tissue expression map, resulting in a total of 24 coexpression sets, containing at least 10 transcripts each (Table S1). For transcripts specific to the tissue expression compendium, custom baits were constructed in a tissue-specific manner. For transcripts in the Hormone and Abiotic Stress Compendium, we designed bait vectors either specific for a given stress or hormone without differentiating the tissue type being queried, or we generated bait vectors specific for both the hormone/stress and tissue type. Thus, for each condition, three different custom baits were constructed. Further, for the stress and hormone expression experiments that consisted of time series data, a collection of six different bait vectors with profiles meant to depict various possibilities of transcript accumulation across the time-course was used. These baits covered scenarios of rapid and consistent transcript accumulation, delayed transcript accumulation, and transcript responses showing accumulation early

in the time-course followed by a decrease. For instance, to identify genes whose transcripts steadily increase in the shoots of cold stressed plants after 24 h of cold exposure, our custom bait vector looked like [1,1,1,...,1,4,8,25,50,75,100,1,...,1,1,1], where the increasing values were set to correspond with the 0.5, 1, 3, 6, 12 and 24 h shoot samples of cold stressed plants, respectively, with ‘1’ entered for all other samples, including the time-matched untreated (control) shoot samples (see Figure S1 for details of such an ‘up steady’ vector and others). For a list of all conditions examined, the bait vectors and the resulting coexpressed transcripts, see Tables S1–S4.

Cistome: analysing sets of promoters identified with custom baits reveals motifs in expected promoter contexts

We developed a pipeline, called Cistome, consisting of several widely-used *cis*-element prediction programs as well as a novel implementation of our Promomer algorithm (Toufighi *et al.*, 2005; see Supporting Information – Methods S1–S3; Figures S2–S4). We assessed the results

of these predictions with a common objective function using an algorithm called Cismar, which is built into the Cistome pipeline. Cismar is a discriminative objective function that compares the distribution of motif occurrence in the target promoter set against its distribution in sets of promoter sequences randomly sampled from all possible promoters. Discriminative functions have been shown to provide a more reliable significance assessment (Redhead and Bailey, 2007; Fauteux *et al.*, 2008; Huggins *et al.*, 2011; Simcha *et al.*, 2012; Grau *et al.*, 2013; Patel and Stormo, 2014; Yao *et al.*, 2014). Cistome was applied to the promoters of the coexpression gene sets identified with our custom bait method in the previous section. This pipeline successfully identified a collection of known motifs occurring within expected promoter contexts. These motifs are listed in Figure 3. The most prominent of these motifs is the well-characterized abscisic acid response element (ABRE), which was found to be highly significantly over-represented in the promoters of transcripts specific to ABA treatment of seedlings ($P = 1.18 \times 10^{-5}$), late seed development ($P = 0.0$), and guard and mesophyll cells treated with ABA ($P = 2.98 \times 10^{-2}$). The ABRE motif has been studied and characterized as functionally relevant in all three of these contexts (Ezcurra *et al.*, 1999; Leonhardt *et al.*, 2004; Nakashima *et al.*, 2006). Similarly, as would be expected for heat shock inducible transcripts, a heat shock element (HSE) was highly significantly over-represented in promoters of this class. Interestingly, the HSE motif exhibited a strong positional disequilibrium in a region approximately 50 basepairs upstream of the transcriptional start site (TSS) of the promoters of genes identified using our custom bait approach as being heat shock-inducible. Finally, a well-known regulatory motif, the RY motif, capable of conferring seed-specific expression was identified in the promoters of transcripts specific to mid/late seed development.

Cistome identifies motifs in novel promoter contexts

In addition to finding the above known motifs in expected contexts, several known motifs were found to occur in novel contexts in our results (Figure 3). These included the presence of a sugar-responsive element (SRE) in the promoters of pollen microphore-specific transcripts, GAGA repeats in the promoters of genes expressed in the shoot apex, and the coupling of a cold-responsive element (CRE) with the circadian-linked evening element (EE) in transcripts specific to cold response. While these are novel contexts, their presence has biological relevance. Namely, sugar response has been shown to play an important role in early pollen development (Zhang *et al.*, 2010). Moreover, a sugar transporter gene (*At1g07340*, *AtSTP2*), identified as microphore-specific by *GUS* reporter gene expression (Truernit *et al.*, 1999), was present in our 'pollen: microphore' and 'pollen: microphore to bicellular' coexpression

clusters and possesses an SRE in its promoter. Likewise, cold response is known to be tightly linked with circadian control: Mikkelsen and Thomashow (2009) have established that the EE and CRE work in concert to govern cold response in *Arabidopsis*. Finally, the presence of GAGA elements in the promoters of genes expressed in the shoot apex is very interesting as the GAGA element has been implicated in Polycomb repression in flies and *Arabidopsis* (Granok *et al.*, 1995; Winter *et al.*, 2011; Deng *et al.*, 2013).

Combining 'custom baits' and Cistome analysis to identify novel motifs in the promoters of genes expressed in seven conditions and/or tissues

Our analysis on the promoters of sets of genes identified using the custom bait approach also identified a collection of novel regulatory motifs, also shown in Figure 3. These included two slightly different C-box-like motifs present in transcripts specific to osmotic and salt stress response. An A-box-like motif was found in the promoters of genes whose transcripts are specific to many stages of seed development. TATA-like signals that appear distinct from traditional TATA box signals were also found in the promoters of genes expressed in the ovary/stigma and root. And, along with a novel motif potentially involved in epidermis-specific expression, an intriguing poly-A signal specific to genes induced by genotoxic stress was uncovered.

Relevance of the custom bait/Cistome analysis approach for identifying novel motifs: *in planta* validation and use case 2

Several of the *de novo* predicted motifs were tested *in planta* for their ability to direct expression of a reporter gene under the corresponding custom bait conditions/samples used to generate the coexpressed gene sets. In total, eight 'synthetic' (promoters assembled with multiple copies of a motif upstream of a minimal promoter) and 30 native promoter constructs driving expression of an eGFP: *GUS* reporter protein (see Experimental procedures) were generated for cold, genotoxic, heat, osmotic and salt stress motifs; the EE circadian clock motif in combination with the cold motif; a motif for ABA response; and motifs for root and epidermis tissues. Constructs were stably transformed into *A. thaliana* Col-0. The ability of the motifs to drive expression of the reporter gene was evaluated by using *in planta* stress or tissue assays, and then quantifying *GUS* expression levels. Cold (10 synthetic lines, 10 native lines), cold and EE (five synthetic lines), genotoxic (13 synthetic lines, four native lines), and heat (five synthetic lines) stress related-motifs; as well as root (seven synthetic lines) and epidermis (eight synthetic lines, five native lines) tissue related-motifs were investigated for their ability to direct reporter gene expression in stable transformants.

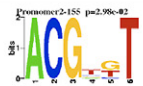

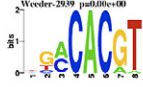

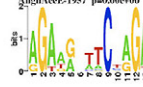

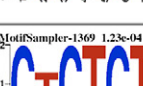
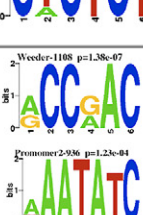
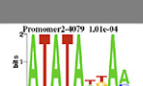
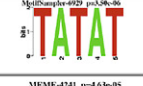
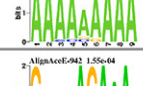

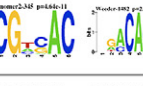


Known Motifs, Known Contexts		
Guard & mesophyll cells		Absciscic acid response element (ABRE) like
Seeds (stgs 6-10)		RY repeat
Seeds (stgs 9-10)		Absciscic acid response element (ABRE)
ABA treatment		Absciscic acid response element (ABRE)
Heat		Heat shock element (HSE)
Known Motifs, Novel Contexts		
Pollen (microphore)		sugar response element (SRE)
Shoot apex		GAGA repeats
Cold		Cold response element (CRE) & Evening element (EE)
Novel Motifs		
Ovary/stigma		TA repeats
Root		TA repeats
Genotoxic		poly-A
Epidermis		novel
Seeds (stgs 3-10)		A-box like
Osmotic		C-box like & ABRE
Salt		C-box like & ABRE

Figure 3. Promoter sets for motif identification were identified using 'custom baits', according to the conditions in the first column. A variety of known motifs were found in promoter contexts in which they are expected to occur (top section). A collection of known motifs were also identified in novel promoter contexts in the promoters of sets of gene transcripts identified with 'custom baits' specific for the condition in the first column (middle section). Novel motifs were identified as significantly enriched in the promoters of genes expressed in seven different conditions and/or tissues, identified with 'custom baits' specific for the condition or tissue shown in the first column (bottom section).

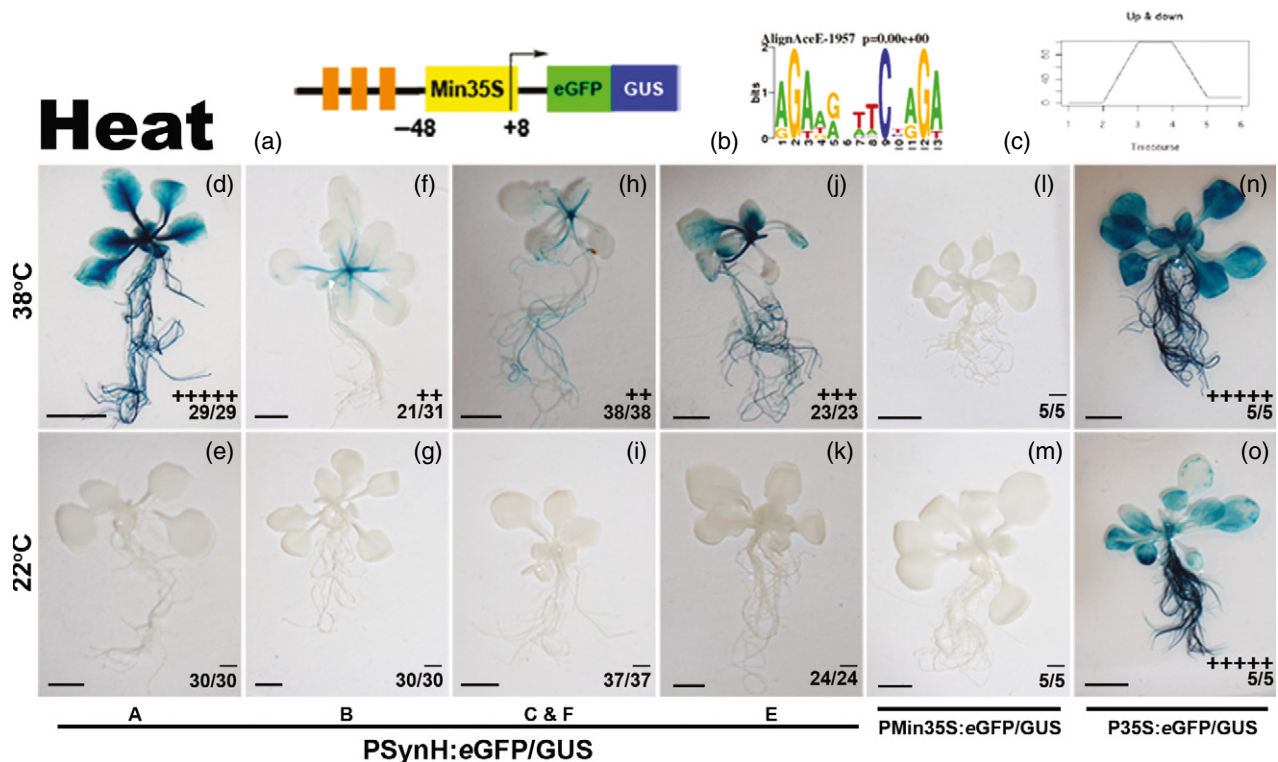


Figure 4. GUS expression of synthetic 'heat' motif lines.

(a) Schematic map of the synthetic heat promoter construct (PSyNH). The orange bars represent the three heat shock elements (HSEs) RGWDRNWHCYRGW in triplet tandem repeats. A minimal 35S promoter and eGFP and GUS reporter genes are also present.

(b) SeqLogos of the HSE identified in this work.

(c) The expected expression profile of genes under control of the RGWDRNWHCYRGW motif. Time points 1–6 represent 0.5, 1, 3, 6, 12 and 24 h.

(d–k) Untreated (22°C) and treated (38°C) seedlings of PSyNH 4.2 lines A (d, e), B (f, g), C and F (h, k) and E (j, k) are shown. Heat-treated seedlings all exhibited GUS expression. Representative seedlings are shown for all lines except line F, which has similar expression patterns to line C, and has been grouped together with line C (h, k). Untreated (m) PMin35S and (o) P35S lines and treated (l) PMin35S and (n) P35S lines. Plus (+) symbols represent the intensity and coverage of the GUS expression, minus (–) symbols on the bottom right represent no GUS expression. Fractions shown in each figure are the number of replicates with that staining pattern/intensity out of the total number stained; transgenic lines with similar staining patterns have been combined. The scale bar at the bottom left of the photographs represents 5 mm.

Positive, heat-inducible GUS staining results constituting our second use case (predicting motifs for driving expression with synthetic or native promoters) were observed for the well-characterized HSE (RGWDRNWHCYRGW) driving reporter expression via a synthetic promoter (Figure 4), indicating that our system for testing motifs in a synthetic promoter system works appropriately, at least in the context of this particular motif. Genotoxic-stress-inducible reporter results from lines driven by promoters containing a novel element (AAMMVRAAA) predicted to confer response to genotoxic stress (bleomycin + mitomycin C) indicate our custom bait pipeline is able to identify promoters that can be used to drive reporter gene expression (Figure 5), although in this case the element used on its own in a synthetic promoter does not appear to be sufficient to drive a genotoxic response (see discussion at the end of this section as to possible reasons why this is).

Positive GUS staining results that recapitulate the expression patterns of the genes whose promoters were

used to predict the motif were also obtained for the novel epidermis (GYDVAGARA) and root (TATAT)-specific elements (Figures S6 and S7). Negative results in the cases where no expected staining patterns were seen (for the 'cold' motif and the 'cold and EE' motifs in combination) could be explained by positional effect variegation. Another explanation might be that the synthetic promoters were designed with the elements 50 bp upstream from the minimal promoter after the recommendation of Gurr and Rushton (2005), but it may be possible that this positioning does not allow the required TF-enhancer complex to correctly interact with the promoter, especially if some type of chromosomal looping is involved. Further, to keep the number of possible constructs manageable, we did not allow for the possibility of the TF complex to form and bind on opposite sides of the DNA strand for the second and/or third instances of the motif in our synthetic promoter constructs – Jacobson *et al.* (1997) have shown that this configuration is required

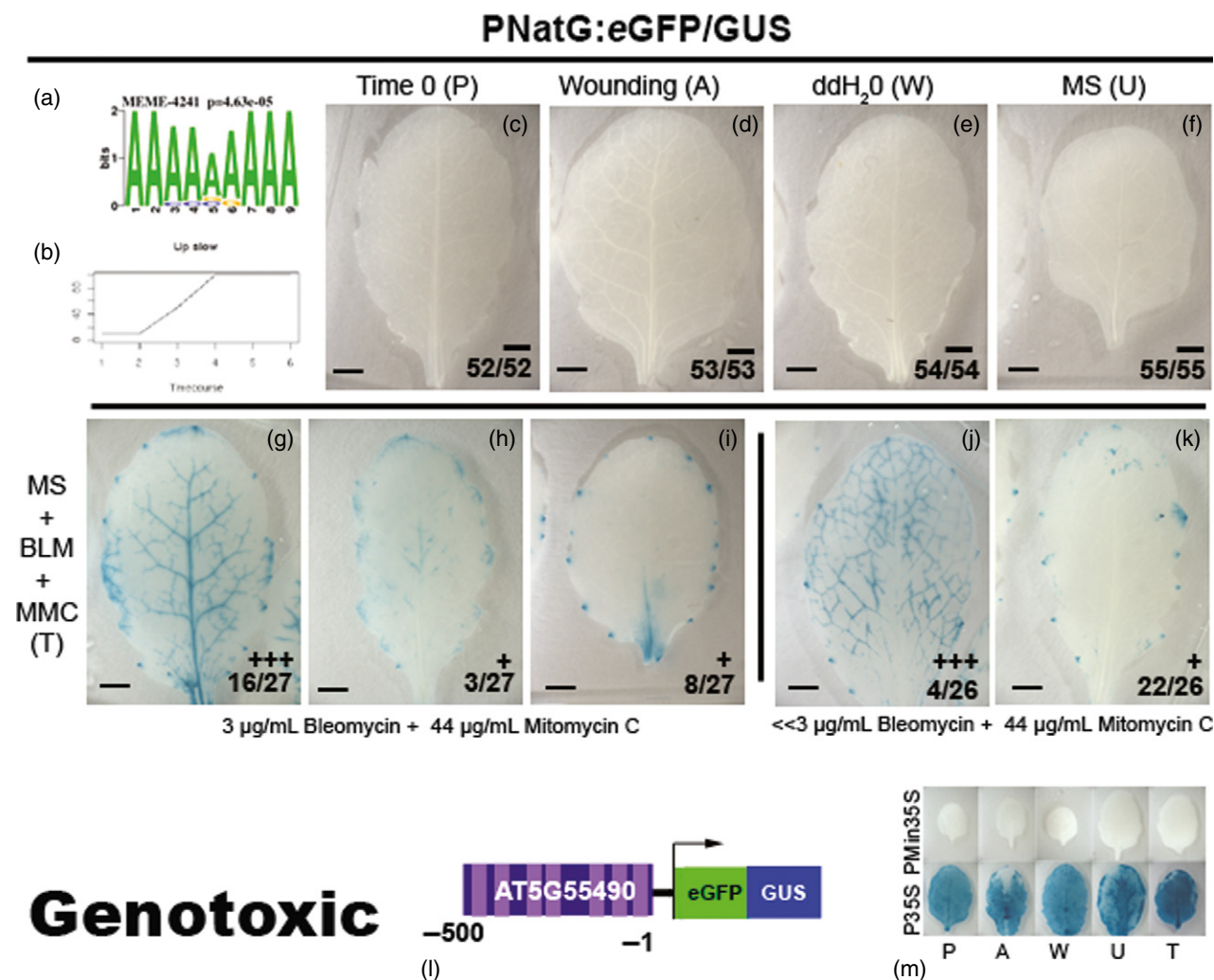


Figure 5. GUS expression of native 'genotoxic' motif promoter line 7.1 C from gene *At5g55490*.

(a) SeqLogos of the predicted genotoxic motif.
 (b) The expected expression profile of genes under control of the genotoxic AAMMVRAAA element. Time points 1–6 represents 0.5, 1, 3, 6, 12 and 24 h.
 (c) Time 0 (P) rosette leaves stained right off the plant at time 0.
 (d) Wounding (A) rosette leaves stained after 24 h in an empty closed container.
 (e) ddH₂O (W) rosette leaves stained 24 h after placement in water.
 (f) MS (U) rosette leaves stained 24 h after placement in sucrose-free MS solution.
 (g–k) MS + BLM + MMC (T) rosette leaves stained 24 h after placement in sucrose-free MS solution containing 3 µg ml⁻¹ bleomycin and 44 µg ml⁻¹ mitomycin C (g–i) or less (j, k).
 (l) Schematic map of the -500 TSS native genotoxic promoter construct. The light purple bars represent the locations of the 'genotoxic' motifs in the *At5g55490* promoter.
 (m) Control PMin35S and P35S leaves. Plus (+) symbols represent the intensity and coverage of the GUS expression, minus (-) symbols represent no GUS expression. Fractions shown in each figure represent the number of replicates with that staining pattern/intensity out of the total number stained. Scale bar at bottom left of each photograph represents 5 mm.

with the rat Pit-1 TF, which binds as a homodimer to sites on opposite sides of the DNA strand. It might also be that we did not see the correct expression pattern because the elements need to be spaced either closer together or further apart for the required interactions or transcriptional complexes to correctly form. Finally, we may have been unaware of the requirement of other 'coupling' elements to direct expression in the expected manner. The data set describing the TF binding

specificities of 745 *A. thaliana* TFs from Weirauch *et al.* (2014) was used in combination with a motif co-occurrence searching program, SpaMo (Whittington *et al.*, 2011), without success to try to identify missing coupling elements. Nevertheless, our intention here is to show that the custom bait feature of Expression Angler is able to identify genes or sets of genes that can be useful as marker genes or for *cis*-element prediction – indeed, this is the case for the several examples shown.

Cistome: a web-based tool for exploring promoters of sets of transcripts identified using the custom bait method of Expression Angler or other approaches

We have developed an online version of the Cistome pipeline used above for exploring the promoter sets generated by our coexpression analysis, and also for visualizing and analysing motifs predicted by this study and those identified by others, such as those collated in the JASPAR (Mathelier *et al.*, 2013) and PLACE (Higo *et al.*, 1999) databases, and those generated by Weirauch *et al.* (2014). A number of functions have been incorporated into the Cistome tool, including the ability to display maps of motif locations in promoter sets, and to cluster and merge related motifs. Further, given the importance of positional bias in 'true' motifs (Ma *et al.*, 2013), we have incorporated an automatic positional distribution function into Cistome's output, along with automatic SeqLogo (Schneider and Stephens, 1990) visualizations of significant motifs (Figure 6). We plan on integrating information on polymorphisms (Korkuć *et al.*, 2014), conserved non-coding regions in the Brassicaceae (Haudry *et al.*, 2013) and DNase I hypersensitive sites (Sullivan *et al.*, 2014), as these data become available through the nascent Arabidopsis Information Portal, Araport.org (2012) to facilitate the exploration of a researcher's own promoter data sets in the context of his/her own e.g. ChIP-Seq data or in the context of known TF binding site data. Cistome is available at <http://bar.utoronto.ca/cistome/>.

CONCLUSION

We present an approach within Expression Angler for designing a custom bait to identify sets of genes exhibiting any desired expression pattern and the results of an attempt to predict the *A. thaliana* 'cis-ome' using publicly available expression data, and a comprehensive prediction and significance testing pipeline. Several predicted motifs direct expression patterns *in planta* in a manner consistent with published results, or exhibit the ability to direct expression in a way predicted by the input expression profiles. We have developed a web-based tool, Cistome, for exploring these and other publicly-available data sets of TF binding specificities, towards a better understanding of transcriptional regulation in *A. thaliana*.

EXPERIMENTAL PROCEDURES

Expression data and analysis

Expression data were acquired from the Bio-Analytic Resource (BAR). The BAR contains *A. thaliana* expression data generated from a wide variety of treatments, tissues and conditions (Toufighi *et al.*, 2005). Three main expression compendia were used: the AtGenExpress developmental series of tissue expression (Schmid *et al.*, 2005); the global abiotic stress expression compendium from AtGenExpress (Kilian *et al.*, 2007); and the AtGenExpress hormone

response series (Goda *et al.*, 2008). The tissue series was extended to include supplemental expression data for gametogenesis, vascular tissues, embryo, stigma/ovaries and epidermis (Honys and Twell, 2003; Suh *et al.*, 2005; Spencer *et al.*, 2007), as well as a collection of some further seed developmental stages (Nakabayashi *et al.*, 2005), and a collection of guard cell and mesophyll cell experiments (Yang *et al.*, 2008). In addition, ATH1 data sets from the AtGenExpress Biotic Stress series have been included in the Expression Angler interface and were used to identify gene expression markers for different kinds of plant responses to pathogens (<http://bar.utoronto.ca/NASCArrays/index.php?ExpID=120>).

Collected expression compendia from all three series were background corrected using MAS5, normalized with scaling (target set to 100) and probe set summarized to expression values with Benj & Wilks summarization from CEL file format using Bioconductor (Gentleman *et al.*, 2004). Affymetrix control probes were removed from each series and pre-processing filters were used to select probes with >25% of their values over 100 fluorescence units and interquartile ranges greater than 0.5, as recommended by Affymetrix.

Identification and validation of pathogenesis-induced gene expression markers

Custom bait Expression Angler queries were generated to look for genes with high expression at early, late or early/late time points of specific kinds of pathogen interaction samples, and low expression across all remaining conditions and times. We defined custom expression patterns using the 'Graphic View' of the 'Biotic Stress – Pseudomonas syringae' or 'Biotic Stress – Elicitor' data sets. The relevant samples used were 2, 6 and 24 h for 10 mM MgCl₂ (control mock inoculation); virulent *P. syringae* pv. tomato DC3000 (to identify DIGs); non-virulent *P. syringae* pv. tomato DC3000::hrcC, and non-host *P. syringae* pv. phaseolicola 1448A (to identify PIGs); and avirulent *P. syringae* pv. tomato DC3000 AvrRpm1 (to identify EIGs). In addition, 1 and 4 h infiltrations for ddH₂O and 1 µM Flg22 in the Elicitor view were additionally used (also to identify PIGs). To identify the PIGs described in this paper, 'high relative expression' (i.e. 1000 expression units) was set for both the early (1 h) and late (4 h) time points in the 'Biotic Stress – Elicitors' with low levels everywhere else (custom baits designed to query combinations of other conditions expected to induce PIGs and/or single time points were also generated but, in the end, it was this combination that proved most useful in terms of RT-PCR markers; Figure 1). To identify the EIG presented in this paper, a 'high relative expression' was set for the early time point (6 h) of the avirulent *P. syringae* pv. tomato AvrRpm1 sample, with low expression in all other samples. To identify the DIGs presented in this paper, a 'high relative expression' was set for the late time point (24 h) of the virulent *P. syringae* pv. tomato DC3000 sample, with low expression in all other samples. The top 50 genes were retrieved for all queries. Candidate genes were confirmed to have reproducible expression levels in all three replicates for a given perturbation, low expression levels in unperturbed conditions, and high levels of expression upon induction (~1000 expression units, about 50-fold above background; the rationale for this was that we were planning on using these expression markers for RT-PCR-based screening) using the BAR's eFP Browser (Winter *et al.*, 2007) and Expression Browser (Toufighi *et al.*, 2005). Those genes where these conditions were not met were excluded from further analysis. About 20 genes were tested by RT-PCR in total, seven of which are shown in this paper.

For confirmation of candidate gene expression markers, the conditions used to grow and infect plants in the AtGenExpress

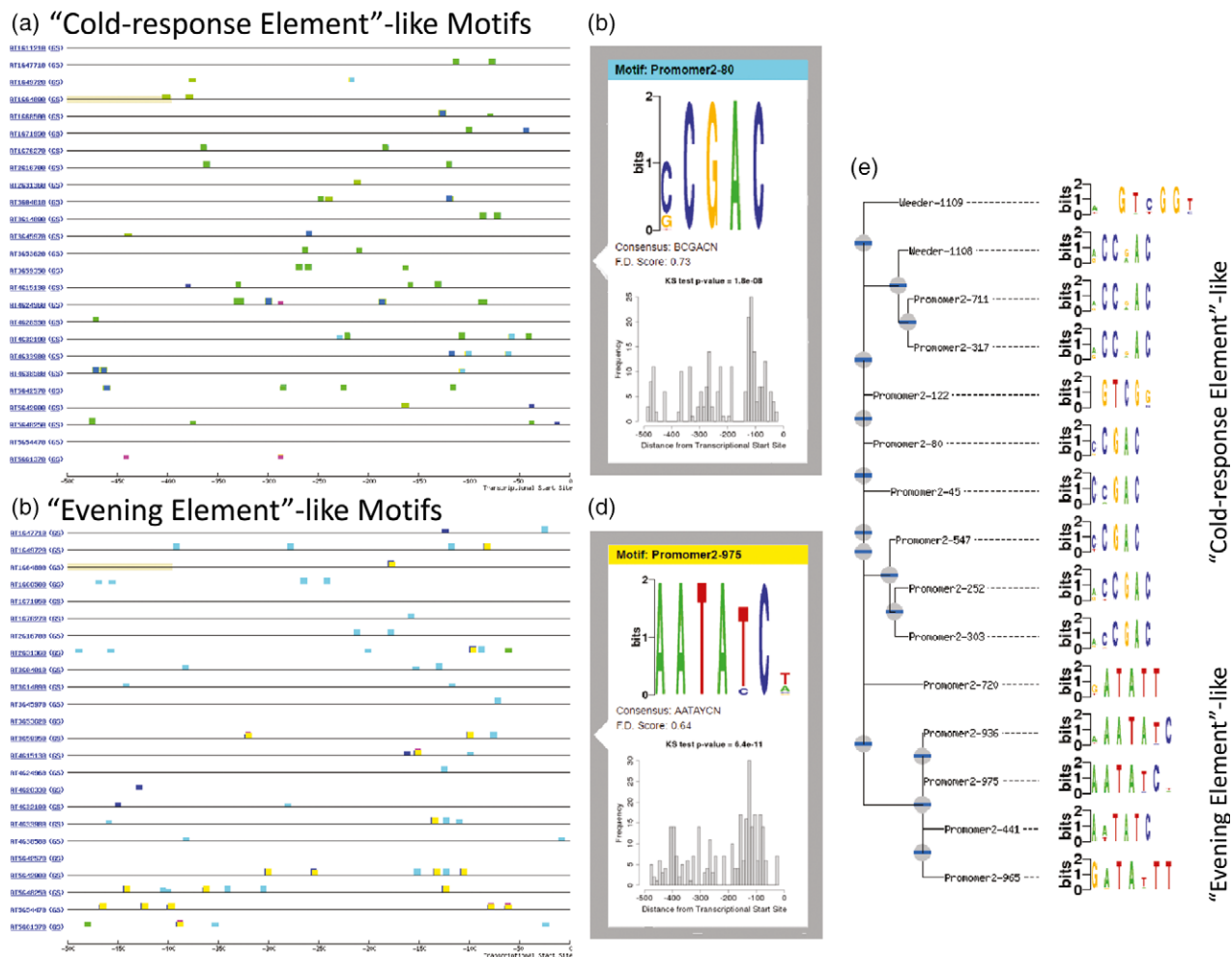


Figure 6. Representative outputs of Cistome.

(a) Related 'cold-response element' (CRE)-like motifs were predicted by several programs used in this study, and these are shown mapped to cold-responsive promoters, also identified in this study (the 'Cold UpLate ShootRoot' set in the Cistome interface, which was created by using a custom bait designed to distinguish such transcripts).
 (b) The positional bias of one of these CREs is shown.
 (c) 'Evening element' (EE)-like motifs were also predicted to be in this same set of promoters.
 (d) The positional bias of one of these 'EE'-like motifs is shown.
 (e) The similarity of the predicted motifs is also viewable within the Cistome application; groups of similar motifs can easily be merged by clicking on a node.

data set were mimicked as much as possible. Arabidopsis Col-0 wild-type seeds were stratified for approximately 72 h at 4°C. Arabidopsis plants were grown in growth chambers set to 22°C, with a light intensity of 100–150 $\mu\text{E m}^{-2} \text{s}^{-1}$ and 9 h of light exposure per day. Plants were infected approximately 4 weeks after stratified seeds were sown on soil.

From *P. syringae* stocks stored at -80°C , bacteria were grown on 1% KB agar (Bauer *et al.*, 1966) plates (with appropriate antibiotic) overnight at 28°C. Bacteria were then restreaked on and grown overnight prior to infection experiments. Bacterial density was adjusted to a concentration of OD_{600} 0.1 in 10 mM MgCl_2 , and these suspensions were pressure infiltrated into the underside of leaves. Leaf tissue was harvested at appropriate 6 or 24 h time points. Similarly, other treatments included 1 μM of Flg22 (in ddH₂O) harvested at 1 or 4 h.

To extract the RNA from leaves, a modified trizol-based protocol was employed based on the tri-reagent protocol

(Chomczynski and Sacchi, 1987). One half of a leaf from three separate plants subject to the same treatment and harvested at the same time were ground to a fine powder, followed by Trizol extraction and ethanol precipitation of the RNA. DNase treatment was used to remove DNA, followed by chloroform extraction and reprecipitation. After quantification, RNA samples were converted to cDNA for RT-PCR using Invitrogen's Superscript II Reverse Transcriptase using the enzyme-specific PCR protocol provided. The quantity of cDNA was assessed using UBQ10 as a loading control to ensure all conditions started with similar amounts of cDNA.

Samples of cDNA for each treatment were diluted in DEPC H₂O by a factor of 1:4. These samples then underwent PCR (using gene-specific primers in Table 1) together using the following settings: 95°C for 60 sec followed by 21 or 24 cycles (of 95°C for 30 sec, 56°C for 30 sec, 72°C for 60 sec), ending with 72°C for 600 sec.

Cistome pipeline for predicting regulatory sequences

In order to identify as comprehensive a regulatory signal as possible within the coexpression clusters, a motif prediction pipeline was devised. This pipeline applied a total of five prediction algorithms from the literature: AlignAce (Hughes *et al.*, 2000); MEME (Bailey *et al.*, 2006); MotifSampler (Thijs *et al.*, 2001); BioProspector (Liu *et al.*, 2004); and Weeder (Pavesi *et al.*, 2001); as well as a custom technique for identifying comprehensive k-mer patterns for a sequence set with allowable mismatches. However, due to the many differences in significance assessment adopted by these programs, all program significance scores were discarded and prediction results from all programs were evaluated using a standardized significance assessment. This consisted of a discriminative scoring technique (encapsulated in the Cismar algorithm) that evaluated the abundance of each pattern within the original promoter set compared with a collection of 1000 randomly generated sets of an equal number of promoters sampled from all possible Arabidopsis promoters. This comparison generated two distributions of overall abundance that were compared with one another by calculating the distance between their means using a Z-statistic in a manner akin to that of the YMF program (Sinha and Tompa, 2003). This collection of significance patterns was further refined in downstream analysis using specific heuristics and filtering strategies. An illustration of the information flow in the project is provided in Figure S2.

As it has been found that all prediction algorithms will identify significant motif patterns even in random sequence sets, a means for assessing an unbiased measure of significance was adopted (Harbison *et al.*, 2004). Briefly, 1000 random promoter sets were sampled from all possible Arabidopsis promoters and run through the prediction pipeline. From each prediction run for each algorithm, the largest Z-score returned by our significance scoring program was identified. The resulting 1000 Z-scores for each algorithm were then used to construct a distribution of null expectancy. Predictions from the biological coexpression sets were then fit against each respective algorithm's distribution to generate a P-value for motif expectancy (Figure S3).

See the Supplemental Results section for a graphical overview of the Cistome pipeline for detecting regulatory sequences, prediction pipeline benchmarking, exhaustive pattern enumeration, and significance assessment.

Synthetic/minimal-promoter construct generation

Consensus sequences for the promoter elements used in the design and construction of synthetic promoters were based on the predictions described in this paper and are summarized in Table S5. Spacer sequences were adapted from the pOp/LhG4 plasmid (Craft *et al.*, 2005), and the minimal 35S promoter sequence was taken from the CaMV35S (−48 to +8) (Cooke and Penon, 1990), such that three copies of the motif to be tested were present at 10 bp spacing to keep them all on the same side of the DNA strand. Before construction, synthetic-promoter and minimal 35S promoter sequences were analysed with PLACE's WebSignal Scan (Higo *et al.*, 1999) to eliminate the presence of unwanted promoter elements. Synthetic promoter (PSyn) promoter-element oligos (Table S6) were annealed to the minimal 35S promoter (PMin35S) oligo (Table S2) and amplified with PCR using Pfu polymerase and primer pairs Synthetic F and R for PSyn, and Negative F and R for PMin35S. Isolated inserts were cloned into pENTR/TOPO (Figure S5 for vector map). Synthetic/minimal-promoter fragments were cloned into pBGWFS7 (Karimi *et al.*, 2002) – see Supplemental Data for vector maps. Positive PSyn:

pBGWFS7 #1–10 lines in *Escherichia coli* were transformed into competent *Agrobacterium tumefaciens* GV3101 and validated with PCR and Eam1104i restriction analysis.

Native promoter construct generation

Sequences used in the construction of the native promoters were taken from TAIR upstream 500 sequences as used in *de novo* predictions. The native promoters selected for constructs (Table S7) were based on promoter analyses performed using TAIR's GBrowse (<http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/>); and the BAR's Expression Browser (Toufighi *et al.*, 2005) and Cistome (<http://bar.utoronto.ca/cistome/>; this work) to select promoters driving strong, specific expression and having multiple copies of the predicted motif if possible. Native promoters were isolated with PCR using Pfu polymerase and promoter-specific primer pairs (Table S7), with wild-type genomic Col-0 DNA as template. Isolated promoters were cloned into pENTR/TOPO with a modified protocol and subsequently transformed into TOP10 *E. coli* cells to create PNat:pENTR constructs. PCR and restriction analyses were performed for validation. Native promoters were then cloned into pBGWFS7 vectors (Karimi *et al.*, 2002) – see Supplemental Data for vector maps, and transformed into TOP10 *E. coli* to create PNat:pBGWFS7s constructs. Validation was performed with PCR and Eam1104i restriction analysis. Positive PNat:pBGWFS7s were transformed into *A. tumefaciens* GV3101.

Constitutive CaMV35S promoter construct generation

A constitutive Cauliflower Mosaic Virus 35S promoter (P35S) was isolated with PCR using Pfu polymerase with P35S primer pairs: forward 5'-AGAGCTCGCATGCCCTTT-3' and reverse 5'-AGAGTCCCCGTGTTCTCTCC-3' with pEGAD (Cutler *et al.*, 2000) as the template. P35S was cloned into pENTR/TOPO and transformed into TOP10 *E. coli*. PCR was performed for validation. PCR cycling conditions and primers pairs used in the amplification of P35S were used, with P35S:pENTR as the template. Positive P35S:pENTR was cloned into pBGWFS7 (Karimi *et al.*, 2002) with LR clonease and transformed into TOP10 *E. coli*. Validation was performed with PCR. Positives were transformed into *A. tumefaciens* GV3101 and validated with PCR.

Growth conditions and plant transformations

Sterilized seeds were plated on 9 × 9 cm² MS-agar plates and stratified for 5 days at 4°C. Sterilized seeds for Basta screening of transgenic lines were directly sowed onto soil and stratified for 7 days at 4°C. Seedlings were grown at 24°C under 110–130 μE m^{−2} s^{−1} of constant light for 1 week, then transferred to soil and placed in a growth chamber at 24°C under 110–130 μE m^{−2} s^{−1} of 16 h light and 8 h darkness. Plants grown for transformation had their first bolts cut and were then allowed to rebolt. Seeds from transformed plants were harvested for selection of positive transformants. Transformation of wild-type Columbia-0 (Col-0) was performed using the floral-dip, *A. tumefaciens*-mediated method as described by Clough and Bent (1998).

In planta motif analysis: stress and tissue assays

The stress and tissue assays were adapted from assays whose microarray expression data were used for the generation of the *de novo in silico* predictions.

Growth conditions of transgenic lines prior to tests. After liquid bleach sterilization, seeds were plated on 9 × 9 cm² plates

containing Basta-MS agar and stratified for 10 days at 4°C. Seedlings were transferred to a growth chamber and grown at 22°C under 110–130 $\mu\text{E m}^{-2} \text{s}^{-1}$ light for 16 h of light and 8 h of darkness. ‘Heat’ motif lines: on the 7th day, positively selected seedlings were transferred to 9 × 9 cm² plates containing MS agar and grown for a further 10 days. ‘Genotoxic’ motif lines: on the 7th day, positively selected seedlings were transferred to soil (Sunshine Mix) and grown at 22°C for another 22 days under 9 h of 110–130 $\mu\text{E m}^{-2} \text{s}^{-1}$ light and 15 h of darkness. Eight–10 seedlings per line were transferred to soil, and four–five seedlings were grown per 10 × 10 cm² pot. After the first set of rosette leaf removal, plants were returned to the growth chamber to recover for 12–15 days for a second set of rosette leaf removal. Control lines: wild-type Col-0 (MS-agar), PMin35S:pBGWFS7 (Basta-MS agar) and P35S:pKGWFS7 (MS agar) were grown under the same conditions as their corresponding stress or tissue transgenic lines. P35S:pKGWFS7 lines were not selected on Kan-MS selection media, and positives were instead identified by eGFP visualization before the assays (Leica MZFLIII stereofluoroscope, GFP2 filter).

Growth conditions during tests. ‘Genotoxic’ motif assay: the genotoxicity assay was adapted from Chen *et al.* (2003). The assays were performed on T2 and T3 leaves. Rosette leaves were taken from plants growing on soil for 22 days, and again (12–15 days after) when the plants had recovered from the first set of leaf removal to allow for a second set to be sampled. Approximately eight–10 plants from each line and five leaves from each plant were assayed under different treatments. Leaf 1 (P) was stained straight off the plant; leaf 2 (A) was placed in an empty Petri dish (5 cm diameter) or 15 ml conical tube; leaf 3 (W) was placed in sterile distilled water; leaf 4 (U) was placed in liquid MS; and leaf 5 (T) was placed in a MS solution containing $\leq 3 \mu\text{g ml}^{-1}$ bleomycin and $\leq 44 \mu\text{g ml}^{-1}$ mitomycin C. All leaves, except for leaf 1, were treated for 24 h under 110–130 $\mu\text{E m}^{-2} \text{s}^{-1}$ of constant light at 22°C. Two different set-ups were used in this assay, and all plants were assayed using both set-ups. The first set-up was performed on the first set of leaves and the second set-up was performed on the second set of leaves. The first set-up used Petri dishes (5 cm diameter) containing 10 ml of the corresponding treatment. Leaves were randomly placed in the treatments with half the leaves placed adaxial side up, and the other half with the adaxial side down. The second set-up used 15-ml conical tubes and 10 ml of the corresponding treatment. All lines were assayed twice. The MS containing approximately $3 \mu\text{g ml}^{-1}$ bleomycin and $44 \mu\text{g ml}^{-1}$ mitomycin C was saved and reused for subsequent assays.

‘Heat’ motif assay: the heat assay was performed on T2 seedlings on the 10th day after transfer to MS agar. The conditions were adapted from Kilian *et al.* (2007). For this assay, the seedlings were transferred to a 38°C incubator for 3 h in constant darkness. The untreated seedlings were kept in the growth chamber in constant darkness. After 3 h, both sets of seedlings were returned to normal growth conditions for 3 h of recovery and were then stained for GUS activity.

Leaf collection for GUS assays, GUS staining and microscopy

Two–three seedlings or leaves were collected from the ‘heat’ and ‘genotoxic’ motif tests, frozen in liquid nitrogen and stored at –80°C for future MUG assays.

GUS staining. Plant tissue was permeabilized at –20°C for 20 min with 90% acetone, then rinsed twice at room temperature

with 0.1 M NaPO₄ pH 7.2–7.4. Tissue were then stained and vacuum-infiltrated with 1 mg ml^{−1} X-gluc dissolved in GUS buffer [0.1 M NaPO₄ pH 7.2–7.4; 0.1 M K₃Fe(CN)₆; 0.1 M K₄Fe(CN)₆] and incubated at 37°C for 24 h. After 24 h, buffer-staining solution was removed and tissue was cleared with 70% ethanol at room temperature. Seven days after clearing, 70% ethanol was removed and replaced.

Tissue preparation and microscopy. Leaf samples were prepared as wet mounts; and stem hand cross-sections were cut by a steel surgical blade under a Leica MZ6 modular stereomicroscope. All stem, leaf and root samples were observed with a Leica MZFLIII stereofluoroscope under bright-field, and images were captured with a Leica DFC300FX camera. All seedling samples were prepared as wet mounts suspended in 70% ethanol, and photographed with a Canon EOS Rebel XSi.

ACKNOWLEDGEMENT

The authors would like to acknowledge the contribution of Hardeep Nahal, who curated the expression data maps from the BAR database.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. The shape of the vectors used to identify coexpressed genes in particular tissues or under particular treatments.

Figure S2. Information flow of the Cistome pipeline.

Figure S3. The distribution of maximum Z-statistic scores for 1000 random promoter sets for each program used in the Cistome pipeline.

Figure S4. The results of benchmarking of the Cistome prediction pipeline against synthetic data sets from yeast and Arabidopsis.

Figure S5. The vector maps of the plasmids used to generate transgenic lines for *in planta* validation.

Figure S6. GUS expression patterns of synthetic and native ‘epidermal’ motif lines.

Figure S7. GUS expression of synthetic ‘root’ motif lines.

Table S1. Samples used for custom bait generation.

Table S2. The number of transcripts found using the ‘custom baits’ from Table S1.

Tables S3 and S4. The actual promoter set lists generated for tissues and abiotic/hormone responses.

Table S5. The putative promoter *cis*-element sequences we discovered with the Cistome pipeline that used for follow-up *in planta* validation.

Tables S6 and S7. The oligonucleotides used to amplify ‘native’ promoters and to generate synthetic promoters containing multiple copies of a predicted *cis*-element.

Methods S1. Describes how we benchmarked the pipeline.

Methods S2. Describes additional experiments not shown in the main text.

Methods S3. File regarding our approach for exhaustive pattern enumeration and significance assessment.

REFERENCES

Amrine, K.C.H., Blanco-Ulate, B. and Cantu, D. (2015) Discovery of core biotic stress responsive genes in Arabidopsis by weighted gene co-expression network analysis. *PLoS ONE*, **10**, e0118731.

- Bailey, T.L., Williams, N., Mischak, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373.
- Bassel, G.W., Lan, H., Glaab, E., Gibbs, D.J., Gerjets, T., Krasnogor, N., Bonner, A.J., Holdsworth, M.J. and Provart, N.J. (2011) Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc. Natl Acad. Sci. USA*, **108**, 9709–9714.
- Bauer, A.W., Kirby, W.M., Sherris, J.C. and Turck, M. (1966) Antibiotic susceptibility testing by a standardized single disk method. *Am. J. Clin. Pathol.* **45**, 493–496.
- Chen, I.-P., Haehnel, U., Altschmied, L., Schubert, I. and Puchta, H. (2003) The transcriptional response of Arabidopsis to genotoxic stress – a high-density colony array study (HDCA). *Plant J.* **35**, 771–786.
- Chomczynski, P. and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159.
- Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743.
- Cooke, R. and Penon, P. (1990) In vitro transcription from cauliflower mosaic virus promoters by a cell-free extract from tobacco cells. *Plant Mol. Biol.* **14**, 391–405.
- Craft, J., Samalova, M., Baroux, C., Townley, H., Martinez, A., Jepson, I., Tsiantis, M. and Moore, I. (2005) New pOp/LhG4 vectors for stringent glucocorticoid-dependent transgene expression in Arabidopsis. *Plant J.* **41**, 899–918.
- Cutler, S.R., Ehrhardt, D.W., Griffiths, J.S. and Somerville, C.R. (2000) Random GFP::DNA fusions enable visualization of subcellular structures in cells of Arabidopsis at a high frequency. *Proc. Natl Acad. Sci. USA*, **97**, 3718–3723.
- Dangl, J.L. and Jones, J.D. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.
- Deng, W., Buzas, D.M., Ying, H., Robertson, M., Taylor, J., Peacock, W.J., Dennis, E.S. and Helliwell, C. (2013) Arabidopsis polycomb repressive complex 2 binding sites contain putative GAGA factor binding motifs within coding regions of genes. *BMC Genomics*, **14**, 593.
- Ezcurra, I., Ellerström, M., Wycliffe, P., Stålberg, K. and Rask, L. (1999) Interaction between composite elements in the napA promoter: both the B-box ABA-responsive complex and the RY/G complex are necessary for seed-specific expression. *Plant Mol. Biol.* **40**, 699–709.
- Fauteux, F., Blanchette, M. and Stromvik, M.V. (2008) Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, **24**, 2303–2307.
- Gentleman, R.C., Carey, V.J., Bates, D.M. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.
- Goda, H., Sasaki, E., Akiyama, K. et al. (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J.* **55**, 526–542.
- Gómez-Gómez, L. and Boller, T. (2000) FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol. Cell*, **5**, 1003–1011.
- Granok, H., Leibovitch, B.A., Shaffer, C.D. and Elgin, S.C.R. (1995) Chromatin: Ga-Ga over GAGA factor. *Curr. Biol.* **5**, 238–241.
- Grau, J., Posch, S., Grosse, I. and Keilwagen, J. (2013) A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.* **41**, e197.
- Greenberg, J.T. and Yao, N. (2004) The role and regulation of programmed cell death in plant-pathogen interactions. *Cell. Microbiol.* **6**, 201–211.
- Gurr, S.J. and Rushton, P.J. (2005) Engineering plants with increased disease resistance: how are we going to express it? *Trends Biotech.* **23**, 283–290.
- Harbison, C.T., Benjamin Gordon, D., Lee, T.I. et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
- Haudry, A., Platts, A.E., Vello, E. et al. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898.
- Higo, K., Ugawa, Y., Iwamoto, M. and Higo, H. (1998) PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res.* **26**, 358–359.
- Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **27**, 297–300.
- Hony, D. and Twell, D. (2003) Comparative analysis of the Arabidopsis pollen transcriptome. *Plant Physiol.* **132**, 640–652.
- Huggins, P., Zhong, S., Shiff, I. et al. (2011) DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, **27**, 2361–2367.
- Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214.
- Jacobson, E.M., Li, P., Leon-del-Rio, A., Rosenfeld, M.G. and Aggarwal, A.K. (1997) Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. *Genes Dev.* **11**, 198–212.
- Jones, J.D.G. and Dangl, J.L. (2006) The plant immune system. *Nature*, **444**, 323–329.
- Karimi, M., Inzé, D. and Depicker, A. (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. *Trends Plant Sci.* **7**, 193–195.
- Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J.* **50**, 347–363.
- Korkuč, P., Schippers, J.H.M. and Walther, D. (2014) Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.* **164**, 181–200.
- Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M. and Rhee, S.Y. (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol.* **28**, 149–156.
- Leonhardt, N., Kwak, J.M., Robert, N., Waner, D., Leonhardt, G. and Schroeder, J.I. (2004) Microarray expression analyses of Arabidopsis guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *Plant Cell*, **16**, 596–615.
- Liu, Y., Wei, L., Batzoglou, S., Brutlag, D.L., Liu, J.S. and Shirley Liu, X. (2004) A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.* **32**, W204–W207.
- van Loon, L.C. and van Kammen, A. (1970) Polyacrylamide disc electrophoresis of the soluble leaf proteins from *Nicotiana tabacum* var. 'Samsun' and 'Samsun NN'. II. Changes in protein constitution after infection with tobacco mosaic virus. *Virology*, **40**, 190–211.
- Ma, S., Shah, S., Bohnert, H.J., Snyder, M. and Dinesh-Kumar, S.P. (2013) Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. *PLoS Genet.* **9**, e1003840.
- Mathelier, A., Zhao, X., Zhang, A.W. et al. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147.
- Mikkelsen, M.D. and Thomashow, M.F. (2009) A role for circadian evening elements in cold-regulated gene expression in Arabidopsis. *Plant J.* **60**, 328–339.
- Nakabayashi, K., Okamoto, M., Koshiba, T., Kamiya, Y. and Nambara, E. (2005) Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. *Plant J.* **41**, 697–709.
- Nakashima, K., Fujita, Y., Katsura, K., Maruyama, K., Narusaka, Y., Seki, M., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2006) Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of Arabidopsis. *Plant Mol. Biol.* **60**, 51–68.
- Nomura, K., Melotto, M. and He, S.-Y. (2005) Suppression of host defense in compatible plant-Pseudomonas syringae interactions. *Curr. Opin. Plant Biol.* **8**, 361–368.
- Obayashi, T., Nishida, K., Kasahara, K. and Kinoshita, K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.* **52**, 213–219.
- Patel, R.Y. and Stormo, G.D. (2014) Discriminative motif optimization based on perceptron training. *Bioinformatics*, **30**, 941–948.
- Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**(Suppl 1), S207–S214.

- Pearce, S., Ferguson, A., King, J. and Wilson, Z.A. (2015) FlowerNet: a gene expression correlation network for anther and pollen development. *Plant Physiol.* **167**, 1717–1730.
- Redhead, E. and Bailey, T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.
- Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.
- Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100.
- Simcha, D., Price, N.D. and Geman, D. (2012) The limits of de novo DNA motif discovery. *PLoS ONE*, **7**, e47836.
- 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.
- Spencer, M.W.B., Casson, S.A. and Lindsey, K. (2007) Transcriptional profiling of the *Arabidopsis* embryo. *Plant Physiol.* **143**, 924–940.
- Srinivasainagendra, V., Page, G.P., Mehta, T., Coulbaly, I. and Loraine, A.E. (2008) CressExpress: a tool for large-scale mining of expression data from *Arabidopsis*. *Plant Physiol.* **147**, 1004–1016.
- Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **23**, 109–113.
- Suh, M.C., Lacey Samuels, A., Jetter, R., Kunst, L., Pollard, M., Ohlrogge, J. and Beisson, F. (2005) Cuticular lipid composition, surface structure, and gene expression in *Arabidopsis* stem epidermis. *Plant Physiol.* **139**, 1649–1665.
- Sullivan, A.M., Arsovski, A.A., Lempe, J. et al. (2014) Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030.
- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P. and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
- Toufighi, K., Brady, S.M., Austin, R., Ly, E. and Provart, N.J. (2005) The botany array resource: E-Northern, Expression Angling, and promoter analyses. *Plant J.* **43**, 153–163.
- Truernit, E., Stadler, R., Baier, K. and Sauer, N. (1999) A male gametophyte-specific monosaccharide transporter in *Arabidopsis*. *Plant J.* **17**, 191–201.
- Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S. and Provart, N.J. (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.* **32**, 1633–1651.
- Weirauch, M.T., Yang, A., Albu, M. et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
- Whittington, T., Frith, M.C., Johnson, J. and Bailey, T.L. (2011) Inferring transcription factor complexes from ChIP-Seq data. *Nucleic Acids Res.* **39**, e98.
- Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J. (2007) An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, **2**, e718.
- Winter, C.M., Austin, R.S., Blanvillain-Baufumé, S. et al. (2011) LEAFY target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Dev. Cell*, **20**, 430–443.
- Yang, Y., Costa, A., Leonhardt, N., Siegel, R.S. and Schroeder, J.I. (2008) Isolation of a strong *Arabidopsis* guard cell promoter and its potential as a research tool. *Plant Methods*, **4**, 6.
- Yao, Z., Macquarrie, K.L., Fong, A.P. et al. (2014) Discriminative motif analysis of high-throughput datasets. *Bioinformatics*, **30**, 775–783.
- Zhang, H., Liang, W., Yang, X., Luo, X., Jiang, N., Ma, H. and Zhang, D. (2010) Carbon starved anther encodes a MYB domain protein that regulates sugar partitioning required for rice pollen development. *Plant Cell Online*, **22**, 672–689.
- Zipfel, C. and Felix, G. (2005) Plants and animals: a different taste for microbes? *Curr. Opin. Plant Biol.* **8**, 353–360.