The Plant Journal (2005) **43**, 153–163 doi: 10.1111/j.1365-313X.2005.02437.x

TECHNICAL ADVANCE

The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses

Kiana Toufighi[1], Siobhan M. Brady[1], Ryan Austin[1], Eugene Ly[2] and Nicholas J. Provart[1,*]
[1]*Department of Botany, University of Toronto, 25 Willcocks Street, Toronto, ON, M5S 3B2 Canada, and*
[2]*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA*

Received 27 January 2005; revised 31 March 2005; accepted 22 April 2005.
[*]For correspondence (fax +1 416 978 5878; e-mail provart@botany.utoronto.ca).

Summary

The Botany Array Resource provides the means for obtaining and archiving microarray data for *Arabidopsis thaliana* as well as biologist-friendly tools for viewing and mining both our own and other's data, for example, from the AtGenExpress Consortium. All the data produced are publicly available through the web interface of the database at http://bbc.botany.utoronto.ca. The database has been designed in accordance with the Minimum Information About a Microarray Experiment [Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365] convention – all expression data are associated with the corresponding experimental details. The database is searchable and it also provides a set of useful and easy-to-use web-based data-mining tools for researchers with sophisticated yet understandable output graphics. These include Expression Browser for performing 'electronic Northerns', Expression Angler for identifying genes that are co-regulated with a gene of interest, and Promomer for identifying potential *cis*-elements in the promoters of individual or co-regulated genes.

Keywords: microarray experiment, expression profile, promoter analysis, electronic Northern, data mining, *Arabidopsis thaliana*.

Introduction

Collections of gene expression data are valuable resources for many aspects of biological research. In recent years, various databases containing Arabidopsis gene expression data have become available. Among these, NASCArrays (Craigon *et al.*, 2004), GEO (Edgar *et al.*, 2002), SMD (Sherlock *et al.*, 2001) and ArrayExpress (Rocca-Serra *et al.*, 2003) are some of the more prominent ones. In addition, a number of portals for analysing microarray data have been developed, including TAIR (Garcia-Hernandez *et al.*, 2002; Rhee *et al.*, 2003), AraCyc (Mueller *et al.*, 2003), MAPMAN (Thimm *et al.*, 2004), and GENEVESTIGATOR (Zimmermann *et al.*, 2004), and for other plant species, e.g. BarleyBase (Shen *et al.*, 2004). The Botany Array Resource (BAR) Expression Browser program differs from these analysis tools in its features, such as hierarchical clustering, automatic averaging, and automatic treatment/control calculation capabilities, and in its ease-of-use. Furthermore, a novel tool, called Expression Angler, allows genes showing similar expression or response profiles to be identified from the selected databases. A final tool, Promomer, allows, among other functions, the promoters of such co-regulated or co-responsive genes to be examined for over-represented motifs. We hereby introduce The BAR, consisting of the aforementioned tools and a database, which currently contains expression data for more than 22 000 genes collected across approximately 150 samples from our own microarray facility. In addition, data from NASCArrays and the AtGenExpress Consortium have been loaded into the system for easy exploration. The expression values in all cases were measured using Affymetrix's GeneChip microarray technology. We validate the use of large-scale gene expression data sets and our programs for functional genomics by using concrete examples from the literature. In the case of the Expression Browser tool to identify potential interaction

partners within or between families, we examine members of the SKIP, CULLEN and F-BOX families that interact to form SCF complexes. The Expression Angler tool can be used to identify genes of unknown function that are co-regulated with one's gene of interest, and we use the example of RGL2 to show how a simple on-line query suggests similar results as recently elucidated genetically in the literature. Finally, we use our promoter analysis program, Promomer, to examine the over-representation of ACGT, the core of the abscisic acid response element (ABRE), both in a single promoter known to be responsive to abscisic acid (ABA) and in a cluster of promoters that cause upregulation of a collection of genes upon treatment with ABA. The graphics shown in this paper are taken directly from the output pages of the aforementioned programs.

Results and discussion

The Botany Array Service

Microarray technology has enabled the measurement of the steady-state mRNA levels of thousands of genes in parallel. In recent years the sequencing of the entire genomes of numerous organisms has been completed, with Arabidopsis (Arabidopsis Genome Initiative, 2000) and rice (Goff *et al.*, 2002; Yu *et al.*, 2002) leading the way in the plant kingdom. The availability of genome sequences allows for the design of microarrays with probes that are specific for genes within gene families, and for whole-genome tiling chips (Yamada *et al.*, 2003) for novel transcription unit identification. The Affymetrix technology allows for highly reproducible results, with a technical error rate of 0.03–0.2% (Redman *et al.*, 2004; Zhu and Wang, 2000, respectively), depending on the criteria used to assess significant differences in gene expression levels.

The data available in our database have been collected using the Affymetrix's GeneChip microarray known as 'ATH1' which represents most of the Arabidopsis genome (roughly 22 810 genes) as annotated by TIGR in their database in December 2001 (Redman *et al.*, 2004). Data from rice and poplar generated using the respective GeneChip microarrays will be added as they become available. Users of the Botany Array Service (departmental researchers and collaborators) perform the necessary steps including growing the plants, conducting the experiment and extracting the RNA before sending the samples to the facility for further microarray processing and final data preparation. Once the data are prepared and normalized (using the standard Affymetrix MAS5.0 algorithm with a target value of 500), they are entered into the BAR database along with the MIAME-compliant experimental details (Brazma *et al.*, 2001). These experimental details are made available to the public as an entire data set as soon as the researcher publishes the project, or after 6 months, whichever comes first. However, the gene expression data themselves are available immediately for use in the data-mining tools described in the following sections. We are currently indexing our samples using the Boyes growth stages (Boyes *et al.*, 2001), and TAIR's Anatomy Ontology (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Ontologies/anatomy.tair.txt), however, in the case of the AtGenExpress data set, we use the Temporal Ontology (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Ontologies/temporal.tair.txt) where possible.

Database features

Font sizes, display colours and site navigation were all designed with the goal of making the website easy to use. The main page of the web interface contains an introduction to the BAR Database and Microarray facility – history of operation, cost summary, and procedures of the service. It also has a menu bar with a search tool and links to three main programs of the BAR: Expression Browser, Expression Angler, and Project Browser. An additional program, called the Data Metaformatter, will reformat the text-based outputs of the first two tools, and append annotations, gene names, functional categories and other information to facilitate interpretation and obviate the need to jump between different websites to gather all information about a gene. A final program, called Promomer, will identify promoter elements based on a word-count method, making it of use in conjunction with the Expression Angler. The menu bar appears on every page thus aiding navigation. The BAR Database is based on MySQL, and the interfaces have been implemented in Perl and C. All functionality is web-based, so no programs or updates must be downloaded to the user's computer, allowing her to focus on the biology.

Project navigation and search

The Project Browser allows users to both view and download all MIAME-compliant attributes of a specific experiment in one place. The Project Browser's front page includes a list of all the projects that have been released by the principal investigator and made available to the public. When an experiment from this list is selected, the project's main page is shown, containing its abstract and research proposal. This general page has links to various 'index cards' containing other aspects of the project including biosource, extraction and labelling, slide, hybridization and scanning protocol information.

Each individual project page within the Project Browser also has a 'Get the Data' link that allows users to download the expression data (including MAS5.0 'Present', 'Marginal', or 'Absent' calls) keyed by probe set ID, and appended with annotation information from TAIR. The search capability conducts an experiment-wide keyword search and consequently retrieves a list of experiments that match the

keyword specified. Project details are only available to external users once the experiment has been published, while the expression data themselves, along with some associated meta-information (plant age, tissue type, control/treatment, and experiment category), are available immediately for use in the tools described below.

Data mining tools – Expression Browser

The Expression Browser allows users to perform so-called 'electronic Northerns'. An electronic Northern is of use to a researcher who is interested in the expression profiles of a particular set of genes. The Expression Browser enables users to input a list of up to 125 genes, which will then be selected across all the experiments in the database and hierarchically clustered and displayed graphically or as plain-text as desired by the user. In addition, we have implemented the ability to query the AtGenExpress Developmental Series data set (Schmid *et al.*, 2005).

The front page of the Expression Browser includes two text boxes for pasting gene AGI codes and corresponding 'My Protein Categories', which the user herself may define. For instance, if working with a gene family, the user may have her own internal nomenclature system, and it is these 'working' codes that may be pasted into this box. Also in this page are three sets of filters that enable users to focus their search of expression profiles. These are experiment research area, sample tissue type, and sample age. The other important user option on this page is the output format: 'Raw', 'Average of replicate treatments', and 'Average of replicate treatments relative to average of appropriate control'. In the case of 'Raw', the expression values for each probe set in each sample are returned to the user. In the case of the 'Average of replicate treatments' option, expression values from two or more samples that were treated in the same fashion are averaged. These averages are then used for clustering or the plain-text display. Finally, the 'Average of replicate treatments relative to average of appropriate control' option takes the ratio of the average of the replicates to the average of the control samples within a project. The final option is useful for looking for similarities in the response of genes across the experiments in the database, while looking at the raw or averaged expression values tends to identify tissue-specific similarities in expression patterns.

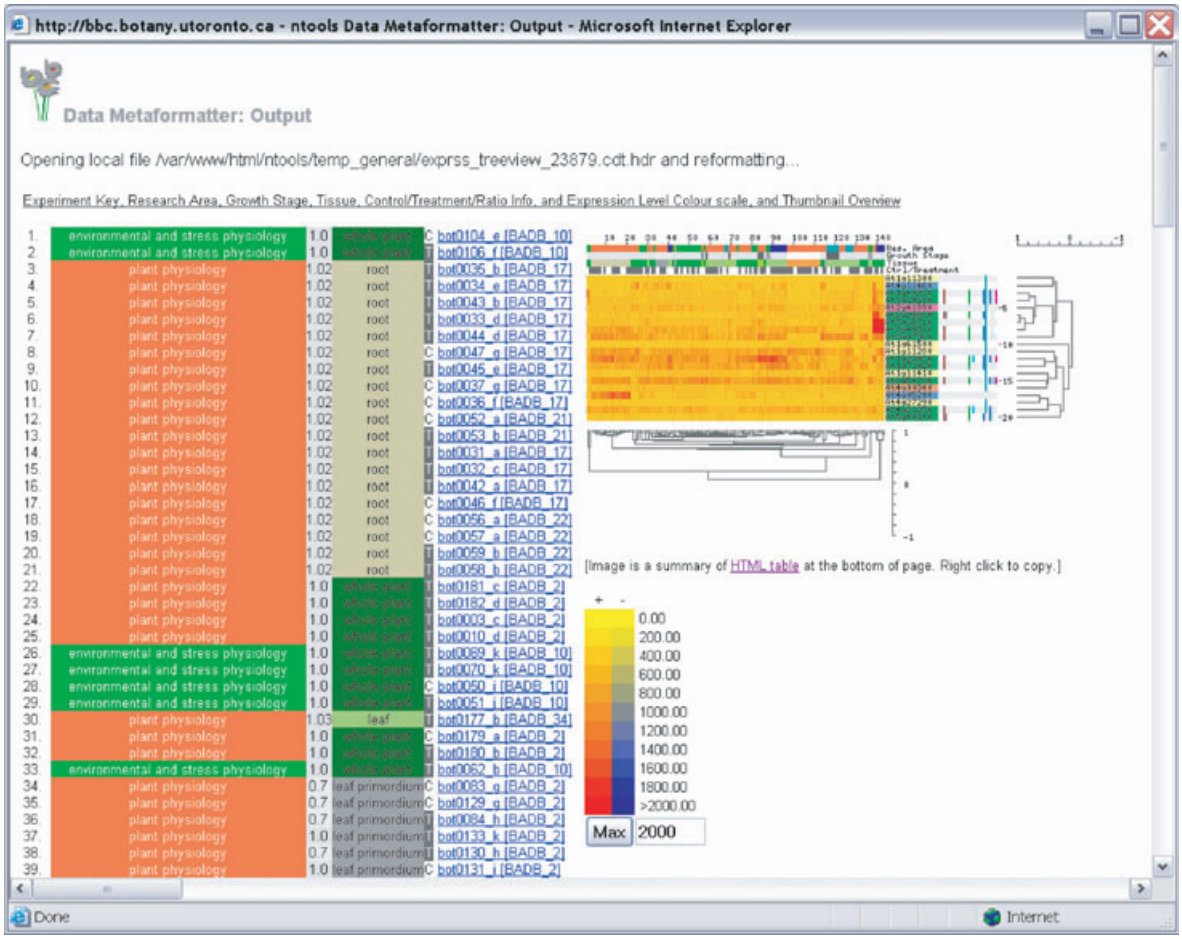
Once the user inputs her desired list of gene AGI codes and specifies the appropriate options, the application selects the genes by looking up the appropriate probe set identifier as defined in a lookup table from TAIR (affy25k_array_elements-2004-04-05.txt), processes them and displays the results as a list of three output options on the results page. The first option is a graphical representation of the results without any hierarchical clustering. The second option is a graphical representation of the probe sets and samples clustered hierarchically. Probe sets and samples are clustered based on their expression profiles or ratios by making use of a Linux-based program called Cluster (de Hoon *et al.*, 2004). The graphical representations of both these output versions have the same features. The only difference between the two is that the former is hierarchically clustered while the latter is not. The third option available on the output page is a file in plain-text format, which may be downloaded to the user's computer, for example, to analyse using other algorithms available in different software, or to import into a spreadsheet program.

In the case of the first two options, the data in plain-text format are passed to the Data Metaformatter program. At the top of the Data Metaformatter output page, all the samples from which the expression profiles were obtained are listed along with their research category, their tissue type, and their growth stage. These are all colour-coded for easy reference. Each sample name has a hyperlink to the project to which it belongs. In the case of the AtGenExpress data set, the hyperlinks are to the appropriate experiment at NASCArrays (see Figure 1).

The expression profiles on the Data Metaformatter output page are displayed in tabular format with each row representing a probe set and each column representing a sample, a group of replicates, or the replicate samples used to calculate the ratio of treatment to control. Again, the sample IDs (or their derivatives) are listed across the top (columns of the table) in the form of links to the project to which they belong. Underneath the sample IDs there are four other fields displaying the research category, age, tissue type, and also whether the sample is a control sample. The gene AGI IDs listed beside the rows of the table are all hyperlinked to the corresponding TAIR gene listing. Beside each gene is a coloured bar-code of its MIPS (Schoof *et al.*, 2002) functional classification(s) along with its corresponding annotation and gene name or alias if available. Because there is not always a one-to-one mapping of the ATH1 probe set identifier to an AGI number (Redman *et al.*, 2004), the rows of the table are actually keyed by the appropriate probe set identifier, and a lookup table from TAIR (affy25k_array_elements-2004-04-05.txt) is used to map to the corresponding AGI number(s).

The last output option of Expression Browser is a plain-text file that contains an unclustered table of probe sets/AGI IDs versus samples. This table is very similar to the graphical one. Each column in this file represents one sample or sample derivative by listing its ID and the corresponding project ID in square brackets, as well as its research category, tissue type, age, and control flag and its expression profile across all the gene identifiers entered by the user. The rows of the table represent probe sets. Each row has an AGI code and annotation for a gene as well as the expression profile across those samples that passed the indicated filters (i.e. those listed across the top row). The data in the text file are generated according to the output

(a)



(b)

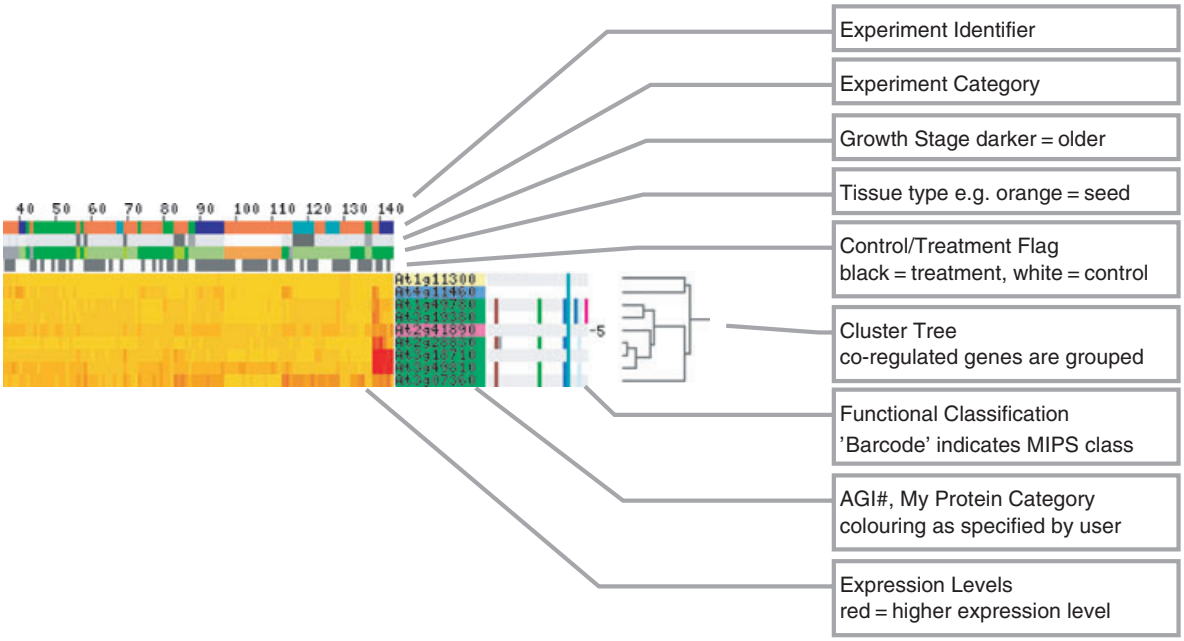


Figure 1. (a) Output of the Expression Browser showing experiment list with experiment categories, plant growth stages, tissues types, treatments, and identifiers, and thumbnail summary of expression levels and cluster results.
A close-up of such a thumbnail graphic with an interpretation guide is shown in (b).

option in the main page of Expression Browser (e.g. raw, average of replicates, or ratio of averages of treatment replicates to designated control replicates).

The Expression Browser tool may be used in at least two different ways by the biologist. One, he can use the tool to gain an idea of where his gene of interest is being expressed, especially by utilizing the AtGenExpress tissue data set. Secondly, he can perform e-Northerns to identify potential redundancies within gene families (e.g. Nakhamchik *et al.*, 2004). Such a methodology can limit the number of double, triple, or higher order knockout mutants that must be generated to address such a question. Alternately, the cluster results can suggest which members across gene families are potential interactors. Fraser *et al.* (2004) and others have shown that there is co-evolution of gene expression among interacting proteins. The Expression Browser tool accepts up to 125 AGI IDs, meaning it can be used even for large gene lists. The results of one particular analysis are shown in Figure 2 for members of three protein families in Arabidopsis, the SKIP, CULLEN and F-Box proteins, members of which can interact to form SCF complexes (Risseeuw *et al.*, 2003) that are involved in

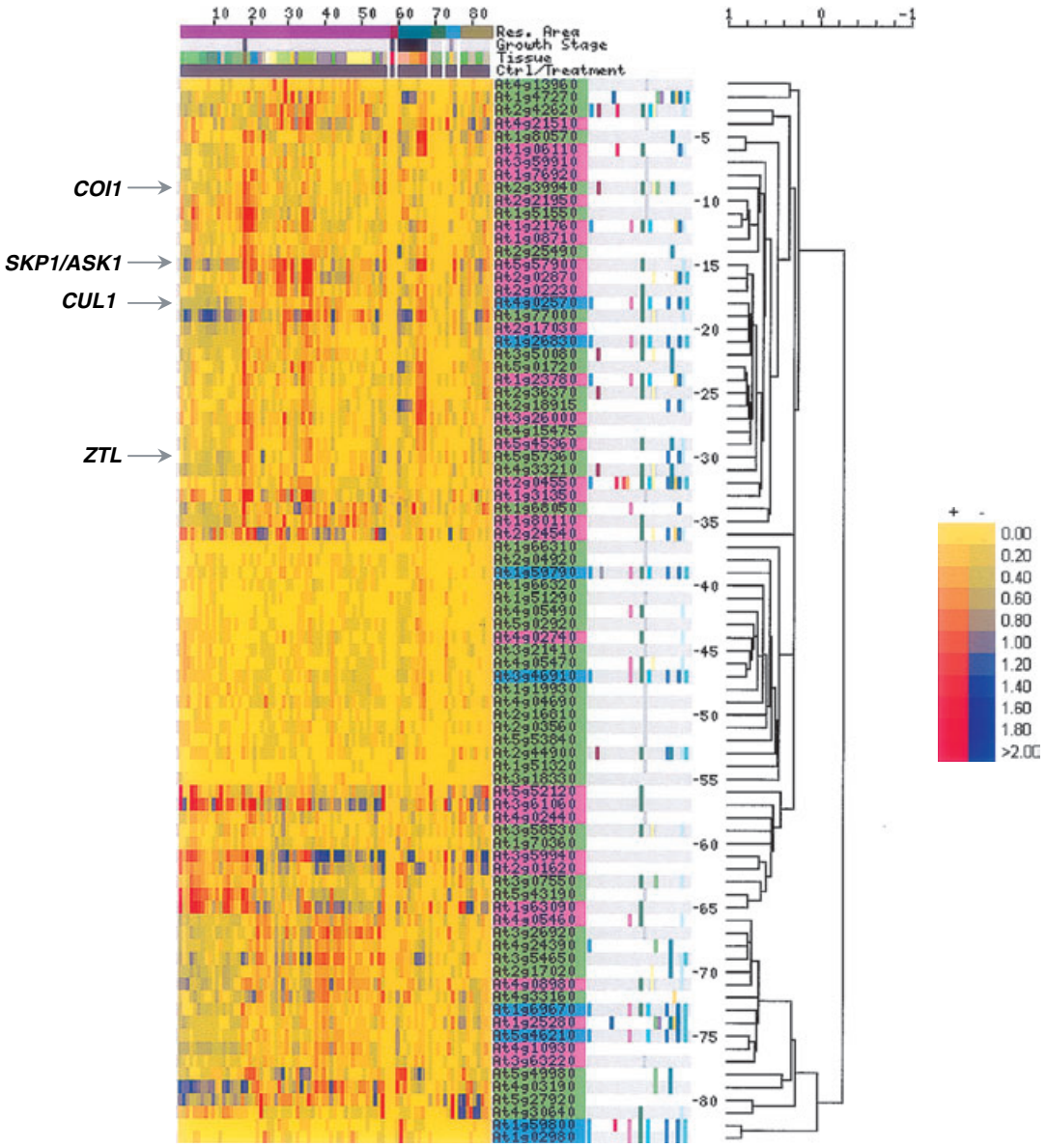


Figure 2. Thumbnail graphic of e-Northern results for members of the SKIP, CULLEN and F-BOX families. The families are denoted by pink, blue and green, respectively, surrounding the AGI numbers. SKP1/ASK1, CUL1 and ZTL can be co-purified by immunoprecipitation *in vivo* (Han *et al.*, 2004), and exhibit more than 70% co-regulation at the level of expression, as indicated by the tree and scale on the right of the image. The colour scale indicates the $\log_2$-level of expression above or below the median. Strong red indicates more than fourfold above the median, while dark blue indicates fourfold below.

protein degradation. It has been shown by co-immunoprecipitation that ZEITLUPE (ZTL), an F-box protein, interacts with both CUL1 and SKP1/ASK1 *in vivo* (Han *et al.*, 2004). The results of our analysis indicate that the genes for these three proteins are co-regulated at the level of gene expression by more than 70%, as measured by the Pearson correlation coefficient and average linkage hierarchical clustering, in the AtGenExpress Development Series data set. Furthermore, the analysis shows that the genes are strongly expressed, relative to the median expression value, in senescent leaves, nodes and internodes, flower parts, and seeds, which corresponds to a promoter-GUS fusion analysis for SKP1/ASK1 performed by Takahashi *et al.* (2004). The analysis by Takahashi *et al.* (2004) and other analyses (Devoto *et al.*, 2002; Xu *et al.*, 2002) have also shown that COI1 (At2g39940) interacts with SKP1/ASK1. We find that COI1 is in a cluster not far removed from SKP1/ASK1. In addition, our Expression Browser analysis suggests other members of these three families that could be potential interaction partners. For instance, the F-BOX protein FBX6 (At5g43190) and a putative phloem protein 2 ortholog At1g63090 (numbers 64 and 65 in Figure 2) are strongly expressed relative to the median expression levels in all leaf stages, and are almost 80% correlated with the level of expression. Experiments using knock-out mutants of these two genes (Alonso *et al.*, 2003), or employing yeast two-hybrid experiments, could thus be designed to explore this suggested interaction and involvement in leaf physiology.

Data mining tools – Expression Angler

The Expression Angler is a program that allows a user to identify genes that respond similarly in terms of their gene expression levels or activation or repression response relative to the appropriate control across all samples in the database. The metric used to identify co-regulated genes is the Pearson correlation coefficient. The Pearson correlation coefficient between two sets, X and Y, of expression values, where $X = \{X_1, X_2,\ldots,X_N\}$ and $Y = \{Y_1, Y_2,\ldots,Y_N\}$, is defined as

$$r = \frac{1}{N}\sum_{i=1,N}\left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right)$$

and ranges from 1, for perfect correlation, to −1, for perfect anti-correlation. A Pearson score of zero means that there is no correlation between the two sets of expression values.

A user enters her AGI number of interest and the program calculates the Pearson correlation between the gene expression vector (set of expression values across all experiments) for that gene and all other genes in the database (a gene must be able to be keyed to a probe set present on the ATH1 GeneChip microarray), in a pairwise fashion. The program then displays those genes exhibiting a degree of expression correlation higher than the selected cut-off. Again the results may be viewed in three ways. As is the case with the Expression Browser, the results in text-only format may be downloaded or viewed. Alternately, they can be passed via a hyperlink to the Data Metaformatter program for easier visualization. Depending on the database in which one 'angles', additional meta-information regarding tissue type, growth stage, experiment category, and control/treatment information is viewable. An example of a thumbnail and part of the corresponding HTML table generated by angling with the *RGL2* gene (At3g03450) in the 392 samples in the NASCArrays database is shown in Figure 3. In the case of the NASCArrays database, no meta-information for tissue type, growth stage, experiment category or treatment/control is available.

The Expression Angling tool may be used in at least three different ways by the biologist. One, she may identify genes co-regulated with a gene of interest and then characterize unknown ones in the list for potential involvement in the biological system that the laboratory is working on. Secondly, the promoters of co-regulated genes may be subject to promoter element discovery, either with our Promomer program, or with other programs such as MotifSampler (Thijs *et al.*, 2002). It has been shown in other systems that highly co-regulated genes often share *cis*-elements – specifically, in the case of human gene Affymetrix expression sets, it was shown that in order for two genes to have a >50% chance of sharing a common *cis*-element, the correlation between their expression profiles across the 611 microarrays used in the study must be >0.84 (Allocco *et al.*, 2004). A third use is to identify genes that are co-regulated with a gene identified in, e.g. a mutant screen to ask the question 'does the gene I've identified 'make sense' biologically?'. The *RGL2* example shown in Figure 3 is illustrative. Yu *et al.* (2004) recently showed that floral homeotic genes are targets of gibberellin signalling. The Expression Angler results shown in Figure 3 corroborate this finding, without performing a wet-lab experiment. While other bioinformatic tools such as the Two Gene Scatterplot program at NASCArrays can provide support, some caveats must be made with the results from Expression Angler. Genes with low average expression levels tend to return large numbers of matches, although the best matches in these lists may still be informative. Figure 4 illustrates this relationship. However, in the case of the data sets in the BAR DB of around 100 samples, 20 732 of 22 810 probe sets (90%) have a call of 'Present' or 'Marginal' in at least one sample, so we have decided not to restrict which probe sets are used by the Expression Angler. Rather, it is ultimately up to the researcher to decide. We strongly advise that gene lists be examined for genes which are known to be involved in some way with the input gene. The presence of these provides support that other genes in the list are not spuriously being identified as co-regulated. A second caveat is that the list of co-regulated genes returned by Expression

(a)

(b)

AGI-ID	Classcode	r-value to Bait|GeneChip-ID|Annotation
1. At3g03450		1.0|BAIT 259042_at|RGL2__gibberellin response modulator, putative / gibberellin-responsi nitrogen metabolism
2. At3g03450		1|259042_at|RGL2__gibberellin response modulator, putative / gibberellin-responsive modu metabolism
3. At3g02310		0.767|259124_at|SEP2__developmental protein SEPALLATA2 / floral homeotic protein (A type transcription factors (DNA-binding and dimerization domain)
4. At1g24260		0.762|264872_at|SEP3__MADS-box protein (AGL9) strongly similar to GB:O22456, MADS
5. At5g44620		0.747|249012_at|CYP706A3__cytochrome P450 family protein similar to cytochrome P45
6. At4g18960		0.745|254595_at|AG__floral homeotic protein AGAMOUS (AG) contains an ACG start cod
7. At5g20240		0.744|246072_at|PI__floral homeotic protein PISTILLATA (PI) contains Pfam profiles PF01
8. At5g22430		0.724|249939_at|expressed protein
9. At3g54340		0.723|251898_at|AP3__floral homeotic protein APETALA3 (AP3)
10. At1g66350		0.72|260141_at|RGL1__gibberellin regulatory protein (RGL1) similar to GB:CAA75492 fron [Arabidopsis thaliana] GI:15777857
11. At3g10570		0.719|258962_at|CYP77A6__cytochrome P450, putative similar to cytochrome P450 77A
12. At2g20870		0.715|265441_at|cell wall protein precursor, putative identical to Putative cell wall protein p
13. At1g58430		0.715|245842_at|RXF26__GDSL-motif lipase/hydrolase family protein similar to family II lip Lipase/Acylhydrolase
14. At5g33370		0.715|246687_at|GDSL-motif lipase/hydrolase family protein similar to family II lipase EXL motif
15. At5g45960		0.712|248924_at|GDSL-motif lipase/hydrolase family protein
16. At1g69120		0.704|259372_at|AP1__floral homeotic protein APETALA1 (AP1) / agamous-like MADS b
17. At2g38110		0.704|267144_at|phospholipid/glycerol acyltransferase family protein low similarity to SP|(

Figure 3. (a) Thumbnail graphic of Expression Angler output (median-centred and normalized), showing response of genes exhibiting similar expression profiles to *RGL2*, At3g03450, at an *r*-value of 0.7 or higher, across 392 samples present in the NASCArrays database.
(b) Close-up of HTML table summarized by the thumbnail, with functional classification barcode, gene aliases and annotations, and link-outs to TAIR.

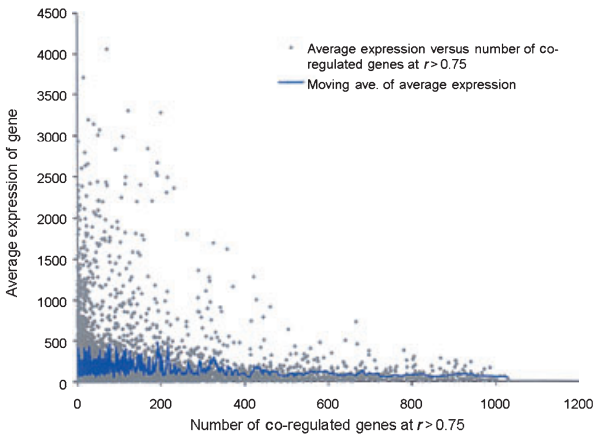


Figure 4. Average expression level across an expression vector versus number of matches to that vector identified by the Expression Angler at $r = 0.75$ or higher.

Angler depends on the samples in the database. If one angles with *RGL2* in the BAR database, a large number of seed-specific genes show up as being co-regulated, due to the presence of seed samples. The involvement of *RGL2* in seeds is well known. It is possible using the subselect page of the Expression Angler to exclude certain samples for the calculation. For instance, if the seed samples are excluded from the BAR data set, then the results returned are similar to those returned when angling in the NASCArrays data set. Uniquely, we also enable angling among the 'response' profiles, i.e. the ratio of the treatment to control gene expression level averages across all genes and samples in the database. To perform this sort of analysis manually would have typically taken several months, and it can now be done in a few minutes with the click of a mouse. As an aside, the Data Metaformatter tool may also be used for visualizing a user's own data set.

Promomer: a novel program for determining statistically over-represented cis*-elements*

With the advent of sequenced genomes and the wealth of existing microarray data, the need for tools to determine modulators of transcriptional networks is apparent. There are two methods that can be used to identify potential *cis*-elements that are the targets of such modulators. The first is the alignment method, and the second, the enumerative method (reviewed by Ohler and Niemann, 2001). The alignment method is exemplified by MotifSampler, which uses a Gibbs sampling method (Thijs *et al.*, 2002). In the enumerative method, the frequency of oligomers of a certain length is examined and is determined to be over-represented when this frequency is higher compared with a background model (van Helden *et al.*, 1998). In this method, the sequence composition of the background model is very important, and should take into account the uneven oligonucleotide representation within each set of sequences or genome. This method avoids finding motifs that are common to all promoters such as the TATA box. There are currently no available web-based enumerative methods specialized for Arabidopsis.

Both the alignment and enumerative methods have been applied to analyse *cis*-elements in plant systems, using various background models. The co-regulated sequences were usually selected by hierarchical clustering methods applied to microarray data or to tissue-specific transcripts (Chen *et al.*, 2002; Harmer *et al.*, 2000; Hudson and Quail, 2003; Hulzink *et al.*, 2003) in order to show biological significance.

The Promomer program aims at providing a user friendly web-based interface to accomplish two goals: (i) To identify statistically over-represented elements in a gene or a group of genes in Arabidopsis using the enumerative method. (ii) To find the number and position of occurrences of an element in genes across the Arabidopsis genome or in a subgroup of genes. When searching for over-represented elements the user is able to analyse 4-mers to 10-mers with a minimum occurrence of three in the case of a single gene promoter, or in 50–100% of all genes in the case of a cluster of promoters. Promomer uses as its reference set the 1 kB 5′ UTR or the 1 kB 3′ UTR data set publicly available from TAIR. The transcription start site of genes is not always well annotated, and therefore the beginning of the ORF is the downstream limit of the putative regulatory region (van Helden *et al.*, 1998). In plants, the 5′ UTRs of several pollen transcripts have been shown to alter gene expression at the transcriptional level (Curie and McCormick, 1997). Promomer's enumerative method is meant to be complementary to alignment methods such as that used by MotifSampler, which allows for degeneracy within an element. As such, a link to MotifSampler is also provided on the output page of the Expression Angler.

The Arabidopsis genome encodes over 1500 transcription factors (Riechmann *et al.*, 2000). However, the change in expression of these genes across various conditions can be too small to be detected by microarray analysis. Clustering methods used to find co-regulated genes may also not be sensitive enough to detect such transcription factors (Chen *et al.*, 2002). Therefore, this program allows for the analysis of the promoters of single genes. The observed frequency of the element in the gene's sequence is compared with the frequency of the element across all promoters in the genome, thus taking into account the sequence composition of promoters of genes in the Arabidopsis genome. The element is ranked in terms of its percentile occurrence.

As a biological test case, we have examined the ACGT-containing ABRE, which has been shown to activate transcription in response to abscisic acid (ABA). In multiple copies it can activate transcription from a minimal promoter, and in single copies must act together with a coupling element (Hobo *et al.*, 1999) to do so. The ACGT box has also been identified as a *cis*-element in promoters of genes regulated by other signals, including auxin and light (Kao *et al.*, 1996; Ulmasov *et al.*, 1995). Late embryogenesis abundant (LEA) genes are highly expressed during late seed development concomitant with an increase in ABA signalling. The Em6 gene is a group 1 LEA gene (Vicient *et al.*, 2000) whose expression is highly responsive to ABA and dependent on ABA biosynthesis (Butler and Cuming, 1993; Gaubier *et al.*, 1993). The ABI5 transcription factor has been shown to strongly bind an ABRE in the *Arabidopsis thaliana* Em6 promoter by an electrophoretic mobility shift assay (Carles *et al.*, 2002). As a functional interaction has been shown for the *cis*-element of the Em6 gene we used the Em6 gene as an example to search for a statistically over-represented 4 bp element in this gene. The ACGT element was the most significant element found, with a *P*-value of 0.002. The output of Promomer for this element is shown in Figure 5(a).

Continuing with ABA as an example of an external stimulus to initiate transcription from an ABRE, a cluster of genes was identified from a microarray that looked at transcript profiling in various cells in response to ABA (Leonhardt *et al.*, 2004). When the same search criteria as used in the search of the Em6 promoter were applied to a cluster of 120 genes that respond to ABA in mesophyll cells, the ACGT box was again determined to be a statistically over-represented element, as illustrated in Figure 5(b). Two distributions are created from 1000 bootstrapped sets of equal number. The first is obtained by sampling for the frequency of occurrence of an element from the given gene cluster promoter set bootstrapped 1000 times, and the second is from 1000 whole genome promoter data sets of equal number to the cluster set. The distribution of occurrence of a given element in both data sets is then obtained

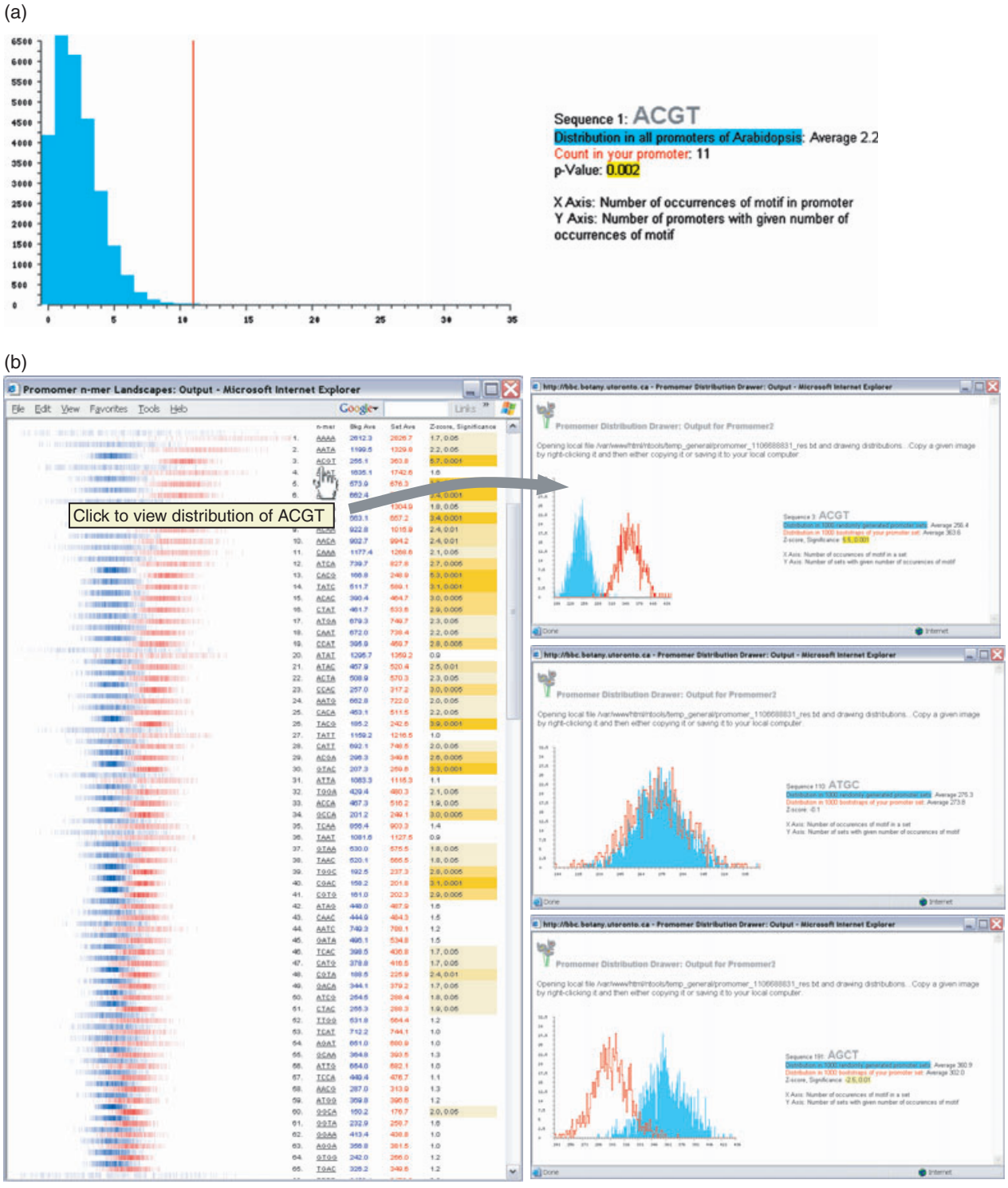


Figure 5. (a) Example Promomer output showing the occurrence of the core ACGT motif of the ABA-responsive element in the promoter of *AtEm6*, At2g40170, a known ABA-responsive gene.
(b) Example Promomer Output showing the distribution of occurrence of ACGT and two variations, ATGC and AGCT, in 1000 bootstrapped sets of 120 promoters of genes upregulated by abscisic acid (ABA) in mesophyll cells (Leonhardt *et al.*, 2004) versus its distribution of occurrence in 1000 sets of 120 promoters randomly selected from the Arabidopsis genome.

and plotted (Chen *et al.*, 2002), and significant differences are highlighted as shown in Figure 5(b). For contrast, the distributions for two variations of the ACGT box, ATGC and AGCT, are also shown. These 4-mers are not significantly over-represented in the cluster in question. We have also enabled a direct link from the Expression Angler so that lists of co-regulated genes may be passed directly to the Promomer program for analysis. Links to PlantCARE (Rombauts *et al.*, 1999), PLACE (Higo *et al.*, 1999), AGRIS (Davuluri *et al.*, 2003) and Athena (http://wyrick.sbs.wsu.edu/Athena/; T.R. O'Connor, J. Wyrick, Washington State University, Pullman, WA, USA, unpublished data) have also been provided to allow the user to cross-reference his list of over-represented elements with databases of known *cis*-elements. In addition to demonstrating the ability to find known *cis*-elements, Promomer also finds novel statistically over-represented elements that can be tested for biological significance.

Finally Promomer is able to find genes that contain an element (for example, a statistically over-represented element found via Promomer analysis) across the genome, or in a subset of genes. Promomer's approach to this is novel, in that it will give the user the offset of each element found, and the average distance between elements. Promomer uses the Boyer–Moore algorithm for matching and counting *n*-mers in a set of sequences (Boyer and Moore, 1977).

Conclusion

In conclusion, the BAR is a multi-purpose, biologist-friendly, web-based collection of programs. Our own Affymetrix data are archived in it according to MIAME conventions. Two tools, the Expression Browser and the Expression Angler, provide powerful means to query the data, while the Data Metaformatter appends many layers of useful information from other public databases, such as TAIR and MIPS, to their outputs and generates easy-to-interpret graphics. In addition, we have loaded data sets into these tools from other sources, including a general NASCArrays data set of 392 samples, and the AtGenExpress data set of 79 tissues in Arabidopsis. By making such data easily accessible to wet-lab researchers, we hope to enable functional genomics to proceed at an accelerated pace. The easy availability of knock-out lines (Alonso *et al.*, 2003) means experiments can be readily conducted based on predictions from 'anonymous' microarray data generated *in silico*. Furthermore, the tools allow researchers to browse microarray data as easily as performing a BLAST search, thereby providing an additional level of information for a particular gene or genes of interest. Finally, the Promomer program allows for the identification of putative *cis*-elements within the promoters of single genes or groups of co-regulated genes, which themselves can then be tested in the laboratory.

Acknowledgements

This project was funded by grants from Genome Canada and NSERC. We thank Jeremy Koch for assistance in programming Promomer, Daphne Goring, John Coleman, and Peter McCourt for critical reading of the manuscript, and George Bassel for helpful discussions.

References

Allocco, D.J., Kohane, I.S. and Butte, A.J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.

Alonso, J.M., Stepanova, A.N., Leisse, T.J. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Boyer, R.S. and Moore, J.S. (1977) A fast string searching algorithm. *Commun. ACM*, **20**, 762–772.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R. and Gorlach, J. (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell*, **13**, 1499–1510.

Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

Butler, W.M. and Cuming, A.C. (1993) Differential molecular responses to abscisic acid and osmotic stress in viviparous maize embryos. *Planta*, **189**, 47–54.

Carles, C., Bies-Etheve, N., Aspart, L., Leon-Kloosterziel, K.M., Koornneef, M., Echeverria, M. and Delseny, M. (2002) Regulation of *Arabidopsis thaliana* Em genes: role of ABI5. *Plant J.* **30**, 373.

Chen, W., Provart, N.J., Glazebrook, J. *et al.* (2002) Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. *Plant Cell*, **14**, 559–574.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.* **32**, D575–D577 (Database issue).

Curie, C. and McCormick, S. (1997) A strong inhibitor of gene expression in the 5′ untranslated region of the pollen-specific LAT59 gene to tomato. *Plant Cell*, **9**, 2025–2036.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M. and Grotewold, E. (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

Devoto, A., Nieto-Rostro, M., Xie, D., Ellis, C., Harmston, R., Patrick, E., Davis, J., Sherratt, L., Coleman, M. and Turner, J.G. (2002) COI1 links jasmonate signalling and fertility to the SCF ubiquitin-ligase complex in Arabidopsis. *Plant J.* **32**, 457–466.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210.

Fraser, H.B., Hirsh, A.E., Wall, D.P. and Eisen, M.B. (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl Acad. Sci. USA*, **101**, 9033–9038.

Garcia-Hernandez, M., Berardini, T.Z., Chen, G. *et al.* (2002) TAIR: a resource for integrated Arabidopsis data. *Funct. Integr. Genomics*, **2**, 239–253.

Gaubier, P., Raynal, M., Hull, G., Huestis, G.M., Grellet, F., Arenas, C., Pages, M. and Delseny, M. (1993) Two different Em-like genes are expressed in *Arabidopsis thaliana* seeds during maturation. *Mol. Gen. Genet.* **238**, 409–418.

Goff, S.A., Ricke, D., Lan, T.H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, **296**, 92–100.

Han, L., Mason, M., Risseeuw, E.P., Crosby, W.L. and Somers, D.E. (2004) Formation of an SCF complex is required for proper regulation of circadian timing. *Plant J.* **40**, 291–301.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A. and Kay, S.A. (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science*, **290**, 2110–2113.

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.

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.

Hobo, T., Asada, M., Kowyama, Y. and Hattori, T. (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *Plant J.* **19**, 679–689.

de Hoon, M.J., Imoto, S., Nolan, J. and Miyano, S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.

Hudson, M.E. and Quail, P.H. (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol.* **133**, 1605–1616.

Hulzink, R.J., Weerdesteyn, H., Croes, A.F., Gerats, T., van Herpen, M.M. and van Helden, J. (2003) In silico identification of putative regulatory sequence elements in the 5′-untranslated region of genes that are expressed during male gametogenesis. *Plant Physiol.* **132**, 75–83.

Kao, C.Y., Cocciolone, S.M., Vasil, I.K. and McCarty, D.R. (1996) Localization and interaction of the cis-acting elements for abscisic acid, VIVIPAROUS1, and light activation of the C1 gene of maize. *Plant Cell*, **8**, 1171–1179.

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.

Mueller, L.A., Zhang, P. and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* **132**, 453–460.

Nakhamchik, A., Zhao, Z., Provart, N.J., Shiu, S.-H., Keatley, S.K., Cameron, R.K. and Goring, D.R. (2004) Expression analysis of the *Arabidopsis* proline-rich extensin-like receptor kinase gene family. *Plant Cell Physiol.* **45**, 1875–1881.

Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.* **17**, 56–60.

Redman, J.C., Haas, B.J., Tanimoto, G. and Town, C.D. (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J.* **38**, 545–561.

Rhee, S.Y., Beavis, W., Berardini, T.Z. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228.

Riechmann, J.L., Heard, J., Martin, G. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Risseeuw, E.P., Daskalchuk, T.E., Banks, T.W., Liu, E., Cotelesage, J., Hellmann, H., Estelle, M., Somers, D.E. and Crosby, W.L. (2003) Protein interaction analysis of SCF ubiquitin E3 ligase subunits from Arabidopsis. *Plant J.* **34**, 753–767.

Rocca-Serra, P., Brazma, A., Parkinson, H. *et al.* (2003) ArrayExpress: a public database of gene expression data at EBI. *C. R. Biol.* **326**, 1075–1078.

Rombauts, S., Dehais, P., Van Montagu, M. and Rouze, P. (1999) PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res.* **27**, 295–296.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J. (2005) A gene expression map of Arabidopsis development. *Nat. Genet.* **37**(5), 501–506.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W. and Mayer, K.F. (2002) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.* **30**, 91–93.

Shen, L., Gong, J., Caldo, R.A., Nettleton, D., Cook, D., Wise, R.P. and Dickerson, J.A. (2004) Barleybase – an expression profiling database for plant genomics. *Nucleic Acids Res.* **33**(Suppl. 1), D614–D618.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A. *et al.* (2001) The Stanford Microarray Database. *Nucleic Acids Res.* **29**, 152–155.

Takahashi, N., Kuroda, H., Kuromori, T., Hirayama, T., Seki, M., Shinozaki, K., Shimada, H. and Matsui, M. (2004) Expression and interaction analysis of Arabidopsis Skp1-related genes. *Plant Cell Physiol.* **45**, 83–91.

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.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939.

Ulmasov, T., Liu, Z.B., Hagen, G. and Guilfoyle, T.J. (1995) Composite structure of auxin response elements. *Plant Cell*, **7**, 1611–1623.

Vicient, C.M., Hull, G., Guilleminot, J., Devic, M. and Delseny, M. (2000) Differential expression of the Arabidopsis genes coding for Em-like proteins. *J. Exp. Bot.* **51**, 1211.

Xu, L., Liu, F., Lechner, E., Genschik, P., Crosby, W.L., Ma, H., Peng, W., Huang, D. and Xie, D. (2002) The SCF(COI1) ubiquitin-ligase complexes are required for jasmonate response in Arabidopsis. *Plant Cell*, **14**, 1919–1935.

Yamada, K., Lim, J., Dale, J.M. *et al.* (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.

Yu, J., Hu, S., Wang, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.

Yu, H., Ito, T., Zhao, Y., Peng, J., Kumar, P. and Meyerowitz, E.M. (2004) Floral homeotic genes are targets of gibberellin signaling in flower development. *Proc. Natl Acad. Sci. USA*, **101**, 7827–7832.

Zhu, T. and Wang, X. (2000) Large-scale profiling of the Arabidopsis transcriptome. *Plant Physiol.* **124**, 1472–1476.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.* **136**, 2621–2632.