CressExpress: A Tool For Large-Scale Mining of Expression Data from Arabidopsis1[W][OA]

Vinodh Srinivasasainagendra, Grier P. Page, Tapan Mehta,issa Coulibaly, and Ann E. Loraine*

Section on Statistical Genetics, Department of Biostatistics, University of Alabama, Birmingham, Alabama 35294 (V.S., G.P.P., T.M, I.C.); and University of North Carolina at Charlotte, Bioinformatics Research Center, North Carolina Research Campus, Kannapolis, North Carolina 28081 (A.E.L.)

CressExpress is a user-friendly, online, coexpression analysis tool for Arabidopsis (Arabidopsis thaliana) microarray expression data that computes patterns of correlated expression between user-entered query genes and the rest of the genes in the genome. Unlike other coexpression tools, CressExpress allows characterization of tissue-specific coexpression networks through user-driven filtering of input data based on sample tissue type. CressExpress also performs pathway-level coexpression analysis on each set of query genes, identifying and ranking genes based on their common connections with two or more query genes. This allows identification of novel candidates for involvement in common processes and functions represented by the query group. Users launch experiments using an easy-to-use Web-based interface and then receive the full complement of results, along with a record of tool settings and parameters, via an e-mail link to the CressExpress Web site. Data sets featured in CressExpress are strictly versioned and include expression data from MAS5, GCRMA, and RMA array processing algorithms. To demonstrate applications for CressExpress, we present coexpression analyses of cellulose synthase genes, indolic glucosinolate biosynthesis, and flowering. We show that subselecting sample types produces a richer network for genes involved in flowering in Arabidopsis. CressExpress provides direct access to expression values via an easy-to-use URL-based Web service, allowing users to determine quickly if their query genes are coexpressed with each other and likely to yield informative pathway-level coexpression results. The tool is available at http://www.cressexpress.org.

Availability of abundant, high-quality data sets from microarray expression experiments has stimulated rapid progress in gene networks analysis for a variety of plant and animal species (Stuart et al., 2003; Craig et al., 2004; Wille et al., 2004; Wei et al., 2006; Zhong and Sternberg, 2006). These data are making it possible to explore correlated expression patterns for the entire genome, as well as answer focused questions regarding specific pathways and processes. By examining correlated expression patterns between genes, investigators can infer new functions for previously uncharacterized genes or identify potential causal relationships between regulators and their targets. Although the details of individual analyses and applications vary, most are based on the idea that correlated expression, or coexpression, implies biologically relevant relationships between gene products.

Many applications of this idea utilize variations of Pearson’s correlation coefficient and linear regression to quantify coexpression relationships. Figure 1 presents an example scatter plot that illustrates the idea. Each point on the plot represents data from one array; x and y coordinates represent expression values for genes indicated on the horizontal and vertical axes, respectively. In this case, there is a strong positive relationship between the two genes’ expression values; when one gene’s expression is high, the other gene’s expression is also high. Computing a linear regression between expression values for the two genes quantifies the strength of this relationship. This yields an r² value, equivalent to the square of Pearson’s correlation coefficient, and a P value that expresses the probability of obtaining the observed r² value (or larger) assuming a random relationship between the two variables. The regression also yields a slope, which indicates the direction of the coexpression relationship. Larger r² and smaller P values signal higher confidence in the coexpression relationship.

Altogether, these numbers summarize how closely two genes are coexpressed across multiple array experiments and provide a way for experimenters to identify and quantify relationships between genes. When these numbers are available for a large number of genes, they can be used to build coexpression graphs, or networks, in which highly correlated genes (nodes) are linked and less well-correlated genes are not. Studies analyzing coexpression networks have demonstrated that genes connected in the coexpression network...
network often perform related functions, thus demonstrating biological relevance of the approach (for review, see Aoki et al., 2007; Saito et al., 2008).

A number of groups have established Web-based interfaces for mining precomputed coexpression results and coexpression networks in Arabidopsis (*Arabidopsis thaliana*). To our knowledge, none offers ways to recompute the coexpression networks using subsets of experiments and arrays, perhaps because of technical challenges involved. Recomputing correlation between a query gene and all the genes in the genome is computationally intensive and cannot easily be done in real-time during a user’s visit to a Web site. However, the coexpression networks arising from different inputs may vary greatly, depending on sample or tissue type. We address this problem by developing an easy-to-use Web tool (CressExpress) that allows users to select distinct tissue types and experiments to include in an analysis, which then executes the calculations offline. When the analysis finishes, users receive an e-mail linking to a zip file on the CressExpress site that contains a complete package of results, along with a record of all parameters and samples used as inputs to the experiment. By providing a complete report of results and inputs, CressExpress makes it convenient for users to integrate coexpression analysis into their research workflow.

**RESULTS**

CressExpress is an easy-to-use Web-based tool that allows researchers to set up and run coexpression analysis experiments using a variety of different data sets and sample types. To set up and run an experiment, users enter query identifiers and analysis parameters on the CressExpress Web site located at http://www.cressexpress.org. To begin an analysis, users click the “Run the Tool” link and then proceed through a series of screens (Fig. 2) that offer users the opportunity to vary quality control (QC) parameters, specify data release and array platforms, and select subsets of sample types to include in analysis. This latter feature can be particularly important for query genes that exert their effects in a tissue- or developmentally restricted fashion. At each step, a “help” icon links to a page describing the various options and how they would likely affect the analysis results. CressExpress also provides reasonable default choices so that users can easily perform pilot studies and quickly learn how the tool operates.

In step one, users choose a data release of expression values that will be used in the analysis. Currently, there are four data release options, each one representing different array collections and array processing methods (Table I). Each release contains expression data harvested from the Nottingham Arabidopsis Stock Center (NASC) AffyWatch subscription service (Craigon et al., 2004) and includes samples from two Affymetrix Arabidopsis array designs: the ATH1 array (22,810 probe sets) and the AG array (approximately 8,000 probe sets). Release 2.0 provides the same data used in a previously published analysis of metabolic pathways; we provide this data set as a courtesy for users interested in investigating the prior study’s results (Wei et al., 2006). Releases 3.0, 3.1, and 3.2 are the same set of arrays, but the expression values in each were generated using different processing methods. We provide data from different array processing methods mainly for users who want to compare results with other online tools or investigate how these methods may affect downstream coexpression analysis results. However, we generally recommend using releases 2.0 or 3.0, which were generated using the RMA algorithm (Irizarry et al., 2003). We recommend these releases mainly because we have observed good separation between correlation coefficients for probe sets expected to be correlated (e.g. redundant probe set pairs; Cui and Loraine, 2006) versus probe set pairs selected at random from the genome (C. Sherrill and A.E. Loraine, unpublished data).

Step one also features an option to configure a QC setting for individual arrays. This QC setting is based on a Kolmogorov-Smirnov (KS) test of deleted residuals, which is described in detail elsewhere (Persson et al., 2005; Trivedi et al., 2005), but we also describe it briefly here. The KS-D test statistic ranges from 0 to 1 and quantifies how much a given array’s expression values deviate from other arrays in the same group.

![Figure 1](Image 60x480 to 275x711) Two highly coexpressed genes from Arabidopsis. RMA-normalized expression values for *AtST5a* (AT1G74100; x axis) and *CYP83B1* (y axis) are plotted relative to each other. Each point indicates expression values from a single array. Regressing expression values for *CYP83B1* (y axis) against expression values for *AtST5a* (x axis) yields an $r^2$ value of approximately 0.7, equivalent to Pearson’s correlation coefficient of 0.85. The regression line appears as a dashed line on the plot.
Arrays sharing the same NASC experiment identifier are considered as belonging to a single group. Arrays with larger KS-D test statistics are of lower quality, at least with respect to how well they resemble other arrays in the same experiment. Decreasing the KS-D threshold excludes outlier arrays that are more variable with respect to the other arrays in the same experimental grouping. Eliminating these lower-quality, outlier arrays can therefore increase observed correlation between genes that are coexpressed. We recommend that when running the tool with a relatively small number of arrays (e.g., <50), users should utilize the default KS-D value of 0.15, because computing coexpression with a smaller number of arrays makes

Figure 2. CressExpress operation. Screen captures showing step-by-step operation of the coexpression tool are shown.
the regression more vulnerable to outliers that can skew results. Coexpression analysis experiments involving more arrays will be less vulnerable to outlier arrays, and eliminating the outlier arrays will have less effect. Figure 3 presents the distribution of KS-D statistics computed for each data release. Note that the release 2.0 arrays were prescreened to exclude arrays with KS-D \( < 0.15 \) as described by Wei et al. (2006).

In step two, users enter a list of up to 50 queries, using Arabidopsis Genome Initiative (AGI) gene names or probe set names to identify genes. To map AGI gene names onto probe set names, CressExpress uses the probe set-to-gene id annotations provided by The Arabidopsis Information Resource (TAIR). However, because these mappings are problematic in some cases (Cui and Loraine, 2006), CressExpress always reports results using both probe set identifiers and AGI codes. Users also use this screen to specify the array type; options include the ATH1 (22,810 probe sets; Redman et al., 2004) or the older AG (approximately 8,000 probe sets) array, both from Affymetrix. By default, the ATH1 array is selected, because more recent and greater amounts of data are available for the ATH1 relative to the AG array. For each query gene, the CressExpress server will perform a large-scale linear regression experiment, comparing each query to all genes represented on the selected array, using all or just some expression data stored in the CressExpress database, depending on the sample and experiments selected in subsequent steps.

In steps three and four, users choose sample types (step three) and experiments (step four) to include in the regression analysis for their query genes. In step three, CressExpress builds a list of sample types for arrays that met the QC and array type criteria specified in previous steps. Users can select all sample types (the default) or subsets of sample types from a menu listing, where the wording for each item on the list comes from the text provided by NASC. Because NASC obtains the textual description of sample types from experimenters who generated and submitted the original data, the sample type descriptions may use a variety of different terms meaning the same thing. Therefore, we advise users to read the entire list when attempting to limit their analyses to specific tissue types, because different text may have similar meanings. For example, users wishing to examine coexpression networks of flowering organs might select sample types labeled “flowers” and “flower buds” as well as “inflorescence.”

The default option for steps three and four is to use all available sample types. This default setting is mainly for the convenience of users wishing to run quick pilot experiments and become familiar with the tool and how it operates. To take full advantage of the CressExpress tool, we recommend that users choose sample types where the query gene products are expected to be expressed or active. For example, users interested in investigating the coexpression network surrounding genes involved in flowering should choose sample types that include flowers. Similarly, users interested in investigating pathways involved in

| Data Release | Data Source | No. of Arrays (Slides) | Processing Methods | Published Example Analyses |
|--------------|-------------|------------------------|--------------------|----------------------------|
| 2.0          | Affywatch I, II | 486 ATH1, 80 AG | RMA               | Cui and Loraine (2006); Wei et al. (2006) |
| 3.0          | Affywatch I, II, III | 1,769 ATH1, 80 AG | RMA | – |
| 3.1          | Affywatch I, II, III | 1,769 ATH1 | GCRMA | – |
| 3.2          | Affywatch I, II, III | 1,769 ATH1 | MAS5; log2 transformation, divide by average | – |

Figure 3. Distribution of KS-D test statistics for CressExpress data releases. KS-D statistics for release 2.0 (A), 3.0 (B), 3.1 (C), and 3.2 (D) are shown.
photosynthesis should choose sample types derived from green shoots and leaves.

To demonstrate the effects of sample type filtering, Figure 4 presents a diagram showing output from a coexpression experiment in which 185 flowering-related genes were compared to each other. Each connection in the network represents an r^2 value ≥ 0.64, such that each pair of connected genes exhibit expression correlation >0.8 or <−0.8. Figure 4A shows the network as computed using all 1,771 arrays from CressExpress data release 3.0, while Figure 4B shows the network computed using a subset of 129 arrays that were labeled as being from flower- or pollen-related samples. The flower-based network shown in Figure 4B contains many more connections than in Figure 4A, demonstrating that, in this case, restricting inputs to biologically relevant sample types yields a richer, more informative network. For example, one of the best-connected genes in Figure 4B is AT3G18990 (VRN1), which encodes a DNA-binding protein involved in vernalization, the process by which exposure to cold temperatures helps to trigger flowering in Arabidopsis (Chandler et al., 1996; Levy et al., 2002). This gene is absent from the network generated from all available samples. This means that if one were entirely unaware that VRN1 plays a role in flowering, an analysis of coexpression would reveal its role in flowering thanks to the many connections VRN1
shares with flowering-related genes, but an analysis of the network generated from all available samples would not. However, we have also observed that in many instances, reducing the number of samples can inflate the number of connections, even among groups of genes that one would not normally consider to be coexpressed. To control for this, we recomputed the flowering network 100 times, using different randomly selected subsets of 129 arrays. Not once did we observe a network as richly connected as the one computed from flowering samples alone, suggesting that, in this case, using flowering-related samples exposes coexpression relationships that would otherwise be masked when all available data are used.

In step five, users may configure a pathway-level coexpression (PLC) analysis, which identifies genes that are coexpressed with multiple query genes. Seen from the point of view of the larger coexpression network, PLC analysis identifies query genes’ common neighbors, such that each neighbor is connected to two or more members of the query group. We previously used PLC analysis to identify candidate genes involved in metabolic pathways and cell wall biosynthesis, and, in general, we have found that genes identified as coexpressed not just with a single gene but with ensembles of genes acting together are often the best and more successful candidates to test (Persson et al., 2005; Wei et al., 2006). Running a PLC analysis requires the user to have entered two or more query genes in step two and also requires the user to specify a linear regression $r^2$ cutoff parameter for designating coexpression between gene pairs, where the $r^2$ value is taken from the linear regression performed during the initial analysis. However, a default value of 0.36, equivalent to a correlation coefficient (Pearson’s $r$) of 0.6, is provided for convenience.

Interpreting and using the PLC functionality and its results requires understanding how the PLC algorithm operates, and so we describe it in detail here. The PLC method as implemented in CressExpress operates as described previously, with some differences in how results are ranked (Wei et al., 2006). The PLC algorithm examines the coexpression results for each of the user-supplied query genes and then builds a list of candidate genes that are coexpressed with two or more members of the query group, where coexpression is defined as an $r^2$ regression result greater than the user-specified threshold. Genes coexpressed with two or more members of the query gene group are ranked first according to the number of genes within the query group with which they are coexpressed, and second by the average $r^2$ value. Thus, genes that are coexpressed with many members of the query group have a higher rank than genes coexpressed with fewer members of the query group.

The PLC method is most useful and relevant when at least two of the query genes are coexpressed with each other at the given PLC $r^2$ threshold. One way to determine the best cutoff for determining coexpression in PLC is to examine the correlations between the query genes that are expected ahead of time to be coexpressed and then choose an $r^2$ threshold that is lower than the smallest pairwise $r^2$ between them. If this is done, then the CressExpress PLC will recover all genes that are coexpressed with the queries at least as well as any two query genes are coexpressed with each other. To find out the threshold $r^2$ between queries, users can run the tool once, examine the list of coexpressed genes for each query individually to find out the cross-query $r^2$ values, and then rerun the tool with a new PLC $r^2$ threshold that is smaller than the lowest cross-query pair. Another method for finding the correlations between queries would be to use the CressExpress expression data direct access method in conjunction with a statistical analysis environment like R (R Development Core Team, 2008) or TableView (Johnson et al., 2003), described in detail below. Regardless of the precise $r^2$ threshold chosen, the PLC analysis has the most meaning when at least two or more of the query gene products are expected to be coexpressed, because they act in concert either in the same pathway or as constituents of a protein complex.

The final step (step six) asks the user to enter a comma-separated list of one or more e-mail addresses that will receive e-mails reporting on the status of the analysis. Upon successful completion of the analysis, each address receives a “Job Completion” message containing a link to a compressed, archive folder (a zip file) stored temporarily on the CressExpress server. The zip file contains the full complement of analysis results as well as a record of all parameters used in the experiment (Table II). CressExpress generates results files (in csv format) for each query gene, in which regression results comparing the query gene to all probe sets on the selected array are reported. The spreadsheet files are named after the query gene’s matching probe set and can be loaded directly into Excel or any other program capable of reading comma-separated format files. Each row of data represents the results from a single linear regression comparing the query gene against another gene on the designated array and includes the linear regression $P$ value, $r^2$ value, and slope, along with the brief description of the target gene. These descriptions come from TAIR and are provided for the convenience of users as they scan results searching for interesting patterns in the types of genes that are most highly coexpressed with their queries.

The PLC results files include csv files reporting coexpressed neighbor genes and corresponding Web pages with hyperlinks to TAIR, allowing for rapid review of results. In addition, the PLC analysis generates a network file (coexp.sif) together with companion node and edge attribute files suitable for loading and visualization in Cytoscape, a popular network analysis and visualization tool (Shannon et al., 2003). Users can configure Cytoscape to utilize a custom styles file packaged with the results files. Directions on
Table II. Data output files

When an analysis completes, users receive an e-mail message reporting a URL where they can download a zipped folder containing several files with results and descriptions of the analysis. Several of these are listed.

| File Name | Description and Format |
|-----------|------------------------|
| [probeset_id].csv (e.g. 246425_at.csv) | Probe set level result file listing regression results for a single probe set. This file records the P value, slope, r^2, and AGI-to-probe set id mappings for each regression result, where the query probe set is compared to every probe set on the selected array. |
| ExperimentMicroarrayMap.csv | Lists experiment identifiers (NASC), CEL file names, KS-D QC test statistics, and slide names for all GeneChips included in the analysis. |
| Coexp.sif | A simple interaction file that can be loaded into Cytoscape. This file includes the user-supplied queries along with other probe sets/genes identified in the PLC analysis. |
| *.noa, *.eda | Node and edge attribute files that can be loaded into Cytoscape, allowing users to modify the network appearance. For example, users can use node attributes to change node labels from probe sets to gene identifiers or use the edge attributes file to color-code edges by coexpression strength. |
| CoexpCompleteSummaryReport.html | This file contains links and explanatory text to the above files and some others. It also records analysis parameters, such as the KS-D QC cutoff parameter, the PLC r^2 cutoff, etc. |
| All_[array name].csv | A consolidated result file with output statistics recorded for all queries. |

CressExpress Direct Access to Expression Values

CressExpress offers direct access to precomputed expression data via a simple URL-based method in which users access expression values for specific probe sets by encoding the requests as Web addresses. Using the direct access approach, users can retrieve expression values for individual genes and arrays, save the values to local files, or import them into directly into Web-enabled desktop programs like R or TableView (Johnson et al., 2003). This feature of CressExpress is useful for advanced users who wish to perform their own custom analyses and therefore need direct access to the underlying raw data used by the CressExpress tool in the large-scale coexpression analysis. On the CressExpress Web site, the tabbed panel labeled “visualization” links to a tutorial explaining how to use the direct access method to import expression data into the TableView program, a user-friendly, freely-available desktop visualization tool implemented in Java. The tutorial describes how users can identify “outlier” arrays that signal potentially meaningful deviations from the coexpression patterns between genes, identify sets of samples that yield unusually high or low expression values for a given gene, or compare the relative variability of similar sample types from different experiments. For more advanced users, the “web services” link provides an example of a short R script showing how to import the data into the R statistical programming environment and compute Pearson’s correlation coefficient between query genes from the glucosinolates biosynthesis pathway. We also provide several BioMoby services for programmers to incorporate CressExpress functionality into their applications (Wilkinson et al., 2005). We do not describe these here but instead refer readers to the CressExpress “Web Services” section, which documents the BioMoby services and their uses.

Biological Application of the Coexpression Tool: Example Analyses

Example Analysis: Cellulose Synthase Enzymes and Cell Wall Biosynthesis

The Arabidopsis genome contains several CESA genes encoding putative and known components of multisubunit cellulose synthase complexes responsible for primary and secondary wall formation. Previ-
ously, we described a coexpression-based analysis that identified genes that were coexpressed with all or some of the CESA genes, and subsequent analysis of mutant phenotypes for some of these genes confirmed their role in cell wall formation and/or stability (Persson et al., 2005). Here, we demonstrate how researchers can use the coexpression tool to recapitulate and extend the analysis, using new releases of expression data featured as part of the CressExpress tool.

Following the procedures described above, we instigated a CressExpress experiment using the six primary and secondary cell wall genes (Table III) as queries and a PLC \( r^2 \) cutoff 0.36. Figure 5 shows a screen capture from the Cytoscape network visualization tool depicting the six query genes and their PLC-identified neighbors. We find that the secondary and primary cell wall CESA genes are coexpressed with different, nonoverlapping groups. Of the genes linked with secondary cell wall CESA genes (Fig. 5B), at least 17 have been investigated experimentally and found to exhibit secondary cell wall-related phenotypes and functions (Table IV), while many more are annotated as having predicted functions related to carbohydrate synthesis or cell wall functions (Supplemental Data S1). This example illustrates how one might use CressExpress to tease apart the different functions of genes that share considerable similarity at the sequence level but which may play distinct roles in the plant body. In this case, it was already known that
CESA4, CESA7, and CESA8 perform a different role from CESA1, CESA3, and CESA6, and we found that the coexpression analysis tends to confirm this view, because the two groups are coexpressed with non-overlapping groups of genes, as determined by the PLC analysis. The same argument may be made for other closely related genes, potentially yielding new hypotheses regarding gene function even for members of closely related gene families.

**Example Analysis: Glucosinolate Biosynthesis from Trp**

Glucosinolates are nitrogen- and sulfur-containing secondary metabolites that are derived from several different amino acids in plants, including Arabidopsis and many agriculturally important Brassicaceae species (Grubb and Abel, 2006; Halkier and Gershenzon, 2006). Glucosinolates undergo conversion to toxic or otherwise bioactive breakdown products through the action of β-thioglucosidase enzymes called myrosinases that only come into contact with their glucosinolate substrates when cells are damaged and cellular contents mix. This mechanism is termed the “mustard oil bomb” and contributes to the plant’s ability to resist pathogen attack and herbivory. Glucosinolate breakdown

---

**Table III.** Probe set-to-target gene mappings for genes encoding cellulose synthase enzymes involved in secondary and primary cell wall formation

| Gene Symbol | AGI    | Probeset (ATH1) | Role        |
|-------------|--------|-----------------|-------------|
| CESA4       | AT5G44030 | 249070_at       | Secondary   |
| CESA7       | AT5G17420 | 246425_at       | Secondary   |
| CESA8       | AT4G18780 | 254618_at       | Secondary   |
| CESA1       | AT4G32410 | 253428_at       | Primary     |
| CESA3       | AT5G05170 | 250827_at       | Primary     |
| CESA6       | AT5G64740 | 247251_at       | Primary     |

---

**Table IV.** PLC analysis results for secondary cell wall biosynthesis genes CESA4, CESA7, and CESA8

Probe set identifiers and their corresponding target genes (AGI identifiers) that are highly correlated with two or more genes involved in secondary cell wall formation are listed. Column heading \( r^2 \) indicates the mean of the \( r^2 \) values obtained from regressing the gene in column 1 against each of the three CESA genes. Column 5 (Phenotype) indicates whether the studies referenced in column six (Refs.) determined that the gene in column 1 possessed a secondary cell wall-related phenotype. The terms “mild” and “severe” indicate degree of phenotypic severity, n/a indicates no published studies were found, and no/yes indicates conflicting reports. Numbers in the column six refer to the following publications: 1, Brown et al. (2005); 2, Ko et al. (2006); 3, Sawa et al. (2005); and 4, Ubeda-Tomas et al. (2007).

| AGI            | Gene Description                              | ATH1 Probeset | \( r^2 \) | Phenotype | Refs. |
|----------------|-----------------------------------------------|---------------|----------|-----------|-------|
| AT1G09610      | Similar to AT1G38000.1                         | 264559_at     | 0.388    | No        | 1     |
| AT1G24030      | Protein kinase family protein                  | 263028_at     | 0.460    | Yes       | 2     |
| AT1G27380      | RIC2                                          | 264495_at     | 0.459    | No        | 1     |
| AT1G27440      | GUT2                                          | 264493_at     | 0.531    | No/yes    | 1, 4  |
| AT1G32100      | Pinoresinol-lariciresinol reductase, putative  | 245792_at     | 0.496    | n/a       |       |
| AT1G62990      | KNAT7                                         | 261106_at     | 0.404    | No        | 1     |
| AT1G73640      | AtRABA6a                                      | 260074_at     | 0.388    | n/a       |       |
| AT2G28760      | UX56                                          | 266221_at     | 0.442    | n/a       |       |
| AT2G29130      | LAC2                                          | 266783_at     | 0.472    | No        | 1     |
| AT2G37090      | IRX9                                          | 265463_at     | 0.484    | Severe    | 1     |
| AT2G38080      | IRX12/LAC4                                    | 267094_at     | 0.614    | Mild      | 1     |
| AT2G40120      | Protein kinase family protein                  | 263383_at     | 0.400    | n/a       |       |
| AT2G41610      | Similar to Os10g0200700                       | 245105_at     | 0.458    | Yes       | 2     |
| AT3G16920      | Chitinase                                      | 257896_at     | 0.672    | No        | 1     |
| AT3G50220      | Nucleic acid binding/pancreatic ribonuclease   | 252211_at     | 0.446    | Yes       | 2     |
| AT3G59690      | IQD13                                         | 251478_at     | 0.413    | No        | 1     |
| AT3G63205      | GLP10                                         | 251297_at     | 0.476    | Yes       | 2     |
| AT4G08160      | Glycosyl hydrolase family 10 protein           | 255150_at     | 0.416    | Yes       | 3     |
| AT4G18780      | CES4A                                         | 254618_at     | 0.796    | Severe    | 1     |
| AT4G27435      | Similar to AT1G52910                          | 253877_at     | 0.585    | No/yes    | 1, 4  |
| AT4G28500      | ANAC073                                       | 253798_at     | 0.363    | Yes       | 2     |
| AT5G01360      | Similar to AT3G55990                          | 251093_at     | 0.508    | Yes       | 2     |
| AT5G03170      | FLA11                                         | 250933_at     | 0.647    | No        | 1     |
| AT5G03260      | LAC11                                         | 250958_at     | 0.479    | n/a       |       |
| AT5G15630      | COBL4/IRX6                                    | 246512_at     | 0.719    | moderate  | 1     |
| AT5G16600      | MYB43                                         | 250103_at     | 0.446    | No        | 1     |
| AT5G17420      | IRX3                                          | 246425_at     | 0.779    | Severe    | 1     |
| AT5G40020      | Pathogenesis-related thaumatin family protein | 249439_at     | 0.428    | Yes       | 2     |
| AT5G44030      | CES4A                                         | 249070_at     | 0.794    | Severe    | 1     |
| AT5G47530      | Auxin-responsive protein, putative            | 248750_at     | 0.369    | n/a       |       |
| AT5G54690      | AT4G18780/IRX8/BGL76                         | 248121_at     | 0.726    | Severe    | 1     |
| AT5G60020      | LAC17                                         | 247648_at     | 0.685    | n/a       |       |
| AT5G60490      | FLA12                                         | 247638_at     | 0.444    | n/a       |       |
| AT5G60720      | Similar to AT4G37080                         | 247590_at     | 0.622    | No        | 1     |
DISCUSSION

We developed CressExpress, an easy-to-use, freely available Web-based tool that allows users to run coexpression analysis experiments using biologically relevant subsets of expression data harvested from public domain Affymetrix Arabidopsis expression microarrays. For each experiment, the tool recomputes coexpression relationships by performing a linear regression comparing each user-entered query gene to all other genes represented in the expression data, using a user-selected subset of experiments in the database. Computing the linear regressions typically requires several minutes, and so individual analyses are performed offline as distinct jobs. All analysis software executes on a server remote from the user’s desktop, and users set up experiments using their Web browser by proceeding through a series of steps in which they enter query genes, choose QC parameters, and specify sample types to include in the analysis. At each point where a choice must be made, CressExpress provides links to help pages describing how the different parameters may affect results and also supplies reasonable defaults for users wanting to run preliminary pilot experiments. Thus, to operate the tool, users need only be able to operate a Web browser and have access to an e-mail account. When the experiment job completes, users receive an e-mail containing a link to a compressed zip file on the server containing results and output files.

Table V. Genes involved in glucosinolate biosynthesis from Trp as reported in AraCyc version 3.5. Values in the column labeled “Other Names” are from TAIR. Names marked with * are from a review of glucosinolate biosynthesis (Grubb and Abel, 2006). Probe set to gene mappings are from TAIR.

| AGI Code   | Other Names                                                                 | Probe Set (ATH1) | Enzyme                    |
|------------|------------------------------------------------------------------------------|------------------|---------------------------|
| AT2G22330  | CYP79B3, CYTOCHROME P450, FAMILY 79, SUBFAMILY B, POLYPEPTIDE 3, T26C19.1     | 264052_at        | Cytochrome P450            |
| AT4G39950  | CYP79B2, CYTOCHROME P450, FAMILY 79, SUBFAMILY B, POLYPEPTIDE 2, T5J17.120, T5J17_120 | 252827_at        | Cytochrome P450            |
| AT4G31500  | CYP83B1, ATR4, CYTOCHROME P450 MONOOXYGENASE 83B1, F3J17.70, F3J17_70, RED1, RNT1, SUR2 | 253534_at        | Cytochrome P450            |
| AT2G20610  | SUR1, ALF1, F23N11.7, F23N11.7, HLS3, RTY, RTY1, SUPERROOT 1               | 263714_at        | Transaminase activity      |
| AT1G24100  | F3J6.2, F3J6_2, UDP-GLUCOSYL TRANSFERASE 74B1, UGT74B1                      | 264873_at        | UDP-glucosyl transferase   |
| AT1G74100  | F2P9.3, F2P9_3, AtST5a*                                                    | 260387_at        | Sulfotransferase           |
Several groups have developed online, Web-based analysis tools that offer a variety of methods and approaches for analyzing and visualizing publicly available Arabidopsis expression data (Zimmermann et al., 2004; Manfield et al., 2006; Obayashi et al., 2007). However, these tools are based on precomputed correlations stored in a database and do not allow users to specify experiments or sample types to include in the calculations. This makes it possible to compare how the observed coexpression network changes when different arrays are included in an analysis. In addition, CressExpress explicitly tracks data releases, allowing users to experiment with different parameter settings, such as different array types, without concern that the underlying database has changed between experiments. To our knowledge, none of the other coexpression data-mining systems provides either the ability to specify sample types to include in analysis or provides detailed information about analysis inputs and parameters.

Another distinctive feature of CressExpress is that it provides comprehensive analysis results in formats aimed at facilitating downstream data-mining and analysis. After a run of the tool, users receive a set of simple, comma-separated plain text file for each query gene. Each of these files lists the $r^2$, slope, and $P$ values for each linear regression between the query and all possible target genes, along with a short textual description of each target to aid users as they explore the results. In general, CressExpress aims to provide data in ways that allow researchers to visualize and explore results using desktop visualization programs and analysis tools, such as Excel, TableView (Johnson et al., 2003), and Cytoscape (Shannon et al., 2003). CressExpress also provides a direct access method allowing users to retrieve raw expressed data directly from the CressExpress database using a simple, URL-based scheme. Using the direct access method, users can build spreadsheets with expression values for every array in a designated data release or import expression values directly into desktop visualization and statistical analysis programs like R and TableView. By providing this relatively straightforward method of accessing data along with tutorials explaining how to

### Table VI. Results from PLC analysis of genes involved in biosynthesis of glucosinolates from Trp

Several genes with unknown functions or functions related to sulfur metabolism and Trp biosynthesis exhibit strong coexpression with all six genes from the glucosinolate biosynthesis pathway, as annotated in AraCyc 3.5. The full list of genes identified via PLC appears in Supplemental Data S2. References cited were collected using LitRep, a BioMoby literature aggregator search tool that queries the Aramemnon, ATIDB, TAIR, and MPIMP databases (http://mips.gsf.de/proj/planet/araaws/litRepSearch.html).

| Gene and Probe Set ID | GO Annotation |
|----------------------|---------------|
| AT4G39940, AKN2      | GO:0016301:kinase activity |
| 252870_at            | GO:0016772:transferase activity, transferring phosphorus-containing groups |
| Lillig et al. (2001) | GO:0005524:ATP binding |
| AT4G30530            | GO:0000103:sulfate assimilation |
| 253606_at            | GO:0012505:endomembrane system |
| Loeffler et al. (2005)| GO:0003824:catalytic activity |
| AT2G14750, APK/APS kinase | GO:0016301:kinase activity |
| 267112_at            | GO:0016772:transferase activity, transferring phosphorus-containing groups |
| Lillig et al. (2001) | GO:0005524:ATP binding |
| AT1G18590, AtST5c    | GO:000103:sulfate assimilation |
| 255773_at            | GO:0008146:sulfotransferase activity |
| Piotrowski et al. (2004); Grubb and Abel (2006) | |
| AT2G30870, ATGSTF10  | GO:0004364:glutathione transferase activity |
| 267154_at            | GO:004049:anthranilate synthase activity |
| Kiyosue et al. (1993); Wagner et al. (2002); Loeffler et al. (2005) | GO:000162:Trp biosynthetic process |
| AT5G05730, ASA1     | GO:0009617:response to wounding |
| 250738_at            | GO:0009507:chloroplast |
| Niyogi and Fink (1992); Bender and Fink (1998); Schuhegger et al. (2007) | |
| AT5G07460, PMSR2    | GO:0008113:protein-Met-S-oxide reductase activity |
| 250633_at            | GO:0006464:protein modification |
| Ekman et al. (2003); Bechtold et al. (2004) | |
use it with R and Table View, CressExpress aims to give users of varying computational experience the opportunity to experiment with computational methods in their research and extract new value from published, publicly available expression microarray data.

In the future, we plan to add several new features to CressExpress, focusing on three major goals: facilitating comparative coexpression analysis across species, making sample selection easier, and repackaging the software to allow easy deployment at other sites. To streamline array selection (step three), we plan to add a feature that will let users select samples based on their Plant Ontology term annotations as they become available (Avraham et al., 2008). We also would like to support comparisons and candidate gene prediction across species and array types. Toward this end, we developed a prototype tool (see www.ssg.uab.edu/ccpmt) that matches probes and probe sets from different arrays based on target gene similarity. Currently, the tool supports six Arabidopsis arrays and a poplar (Populus spp.) genome array from Affymetrix. We would also like to make it easy for other groups to mirror the CressExpress software on their own sites, thus allowing them to reuse the software with their own custom data sets. In hopes of recruiting community interest in this latter effort, we created a developer’s source code release of the software on the CressExpress Web site under the “Data Mining Code” link. However, it is important to note that CressExpress uses a commercial Java statistics library from Visual Numerics, which charges a licensing fee. Although we feel that the convenience of having a robust library of statistical routines is well worth the price, we may ultimately replace this non-free component with a version from the open source community. Another modification we are considering involves increasing the current limit of 50 queries per CressExpress analysis, depending on community interest. Readers who would like to suggest other new features or changes to the tool behavior are welcome to contribute ideas and requests.

MATERIALS AND METHODS

Array Informatics

CD media containing expression data were obtained from the NASC AffyWatch subscription service. Each CD contained CEL files with “raw” expression data, grouped into folders named for the investigator who contributed the data.

Upon receipt of each CD, XML files describing each experiment were harvested from the NASC site, using the “passsthru” parameter as described on the Web page located at http://arabidopsis.info/bioinformatics/narrrayxml/index.html. Slide names associated with each experiment id were harvested from the XML files by extracting the content of “NASC: Name” tags. For about one-half of the CEL files, we were able to use CEL file names and slide names to connect slides with their corresponding CEL files and thus capture in our database the experimental group affiliation for each CEL file. For the remainder, NASC supplied (via e-mail) Excel spreadsheets that report the correspondence between CEL file names and NASC slide names. The mapping between slide name and CEL file names were also manually checked; cases where the names seemed to disagree or contradict each other were resolved via e-mail correspondence with NASC. The mapping is supplied as part of the direct access method.

Array Processing

CEL file processing steps, including background correction, probe set summarization, and normalization procedures, were done using functions in the Bioconductor “affy” package suite of tools (Gautier et al., 2004). For release 2.0, arrays were processed using the RMA algorithm as described previously (Wei et al., 2006). Each of the 3.x releases included the same set of arrays, including arrays from Affywatch series I, II, and III, but were processed using different affy package methods and algorithms. For release 3.0, arrays were processed using the justRMA algorithm with default settings. For release 3.1, arrays were processed using the justGCRMA algorithm, again with default parameters. Releases 3.0 and 3.1 were generated using an Altix computer at the Alabama Supercomputing Center. Release 3.2 was generated using the MASS algorithm, followed by a scaling step in which expression values for each probe set on the same array were divided by the average probe set value for that array. Following the scaling step, the log (base 2) was taken. Processing for release 3.2 was done on a computer equipped with two 2-GHz AMD Opteron 246 CPUs and 16 Gb of RAM running the Linux operating system.

Coexpression Analysis and QC Procedures

The coexpression tool is based on large-scale linear regression analysis of expression values between genes of interest and the rest of the genes on a selected array using the methodology described previously (Persson et al., 2005; Wei et al., 2006). For each data release, we computed the KS-D statistics for each array, except in cases where there were fewer than three arrays per experiment. Detailed descriptions of the KS-D computation have been previously reported (Persson et al., 2005; Trivedi et al., 2005).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Data S1. Genes coexpressed with CES44, CES7, and CES48.

Supplemental Data S2. Web page reporting all genes identified by PLC ($\rho$ threshold 0.35) as coexpressed with the glucosinolate biosynthesis pathway.

ACKNOWLEDGMENTS

The authors thank the Alabama Supercomputer Center for providing computational assistance with array processing steps, Jim Johnson for help with the TableView software, the School of Public Health MITS team for computer systems support, and Mikako Kawai for Web page and graphic design. We are particularly grateful for the staff of the NASC for their generosity and help with matching slide and CEL file names. Last, we thank the anonymous reviewers for their excellent comments on the manuscript.

Received January 18, 2008; accepted April 29, 2008; published May 8, 2008.

LITERATURE CITED

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48: 381–390

Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A, Reiser I, Rhee SY, Sachs MM, et al (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res 36: D449–D454

Brestel UJ, Murphy DJ, Mullineaux PM (2004) Arabidopsis peptide methionine sulfoxide reductase2 prevents cellular oxidative damage in long nights. Plant Cell 16: 908–919

Bender J, Fink GR (1998) A Myb homologue, ATR1, activates tryptophan gene expression in Arabidopsis. Proc Natl Acad Sci USA 95: 5655–5660
Srinivasasainagendra et al.

Brown DM, Zeef LA, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell 17: 2281–2292

Chandler J, Wilson A, Dean C (1996) Arabidopsis mutants showing an altered response to vernalization. Plant J 10: 637–644

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArray: a repository for microarray data generated by NASC’s transcriptomics service. Nucleic Acids Res 32: D575–D577

Cui X, Loraine AE (2006) Global correlation analysis between redundant probe sets using a large collection of Arabidopsis ATH1 expression profiling data. In LSS Computational Systems Bioinformatics. World Scientific, Stanford, CA, pp 223–227

Ekman DR, Lorenz WW, Przybyla AE, Wolfe NL, Dean JF (2004) Exploration, normalization, and summaries of high dimensional biology data. BMC Bioinformatics 5: 307–315

Gachon CM, Langlois-Meurinne M, Henry Y, Saindrenan P (2005) Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. Plant Mol Biol 58: 229–245

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy: Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315

Grubb CD, Abel S (2006) Glucosinolate metabolism and its control. Trends Plant Sci 11: 89–100

Halkier BA, Gershenzon J

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, et al. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci USA 104: 6478–6483

Irizarry RA, Hobbs B, Collin E, Beazer-Barclay YD, Antonellis KJ, Suerf O, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264

Johnson JE, Stromvik MV, Silverstein KA, Crow JA, Shoop E, Retzel EF (2003) Tableview: a portable genomic data visualization. Bioinformatics 19: 1292–1293

Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K (1993) Characterization of two cDNAs (ERD11 and ERD13) for dehydration-inducible genes that encode putative glutathione S-transferases in Arabidopsis thaliana L. FEBS Lett 335: 189–192

Ko JH, Beers EP, Han KH (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in Arabidopsis thaliana. Mol Genet Genomics 276: 517–531

Levy YY, Mesnage S, Myline JS, Gendall AR, Dean C (2002) Multiple roles of Arabidopsis VRN1 in vernalization and flowering time control. Science 297: 243–246

Lillig CH, Schiffmann S, Berndt C, Berken A, Tischka R, Schwenn JD (2001) Molecular and catalytic properties of Arabidopsis thaliana ade-nyl sulphate (APS)-kinase. Arch Biochem Biophys 392: 303–310

Loeffler C, Berger S, Guy A, Durand T, Bringmann G, Dreyer M, von Rad U, Durner J, Mueller MJ (2001) B5-phytoprostanes trigger plant defense and detoxification responses. Plant Physiol 127: 328–340

Manfield IW, Jen CH, Pinney JW, Michelopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Res 34: W504–W509

Niyogi KK, Fink GR (1992) Two anthranilate synthase genes in Arabidopsis: defense-related regulation of the tryptophan pathway. Plant Cell 4: 721–733

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res 35: D863–D869

Persson W, Wei H, Milene J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA 102: 8633–8638

Piotrowski M, Schemenawitz W, Lopukhina A, Muller A, Janowitz T, Weber EW, Oechting C (2004) Desulfitoglucosinolate sulfo transferases from Arabidopsis thaliana catalyze the final step in the biosynthesis of the glucosinolate core structure. J Biol Chem 279: 50717–50725

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Redman JC, Haas BJ, Tanimoto G, Town CD (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array.

Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with coexpression networks and metabolomics: ‘majority report by precogs’. Trends Plant Sci 13: 36–43

Sawa S, Demura T, Horiguchi G, Kubo M, Fukuda H (2005) The ATE genes are responsible for repression of transdifferentiation into xylem cells in Arabidopsis. Plant Physiol 137: 141–148

Schuhegger R, Rauhut T, Glawischning E (2007) Regulatory variabilty of camalexin biosynthesis. J Plant Physiol 164: 636–644

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ieder K (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for work global discovery of conserved genetic modules. Science 302: 249–255

Trivedi P, Edwards JW, Wang J, Gadbury GL, Srinivasasainagendra V, Zakharkin SO, Kim K, Mehta T, Brand JP, Patki A, et al. (2005) HDiBBst! a platform-independent software suite for statistical analysis of high dimensional biology data. BMC Bioinformatics 6: 86

Ubeda-Tomas S, Edvardsson E, Eland C, Singh SK, Zadik D, Aspeborg H, Gorzsas A, Teeri TT, Sundberg B, Persson P, et al. (2007) Genomic-assisted identification of genes involved in secondary growth in Arabidopsis utilizing transcript profiling of poplar wood-forming tissues. Physiol Plant 129: 415–428

Wagner U, Edwards R, Dixon DP, Mauch F (2002) Probing the diversity of the Arabidopsis glutathione S-transferase gene family. Plant Mol Biol 49: 515–532

Westhead DR, Gilmartin PM, Town CD (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant biochemistry and molecular biology. Nucleic Acids Res 33: D389–D392

Wilkinson M, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlateInet exemplar case. Plant Physiol 138: S1–S7

Wille A, Zimmermann P, Vanova E, Furholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, et al. (2004) Graphical Kaiser modeling of the isoprenoid network in Arabidopsis thaliana. Genome Biol 5: R92

Zhang L, Liao Y, Liu K, Liu X (2004) g2map: a JAVA-based tool for functional analysis of microarray data. Bioinformatics 20: 1481–1482

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