Genevestigator. Facilitating Web-Based Gene-Expression Analysis

Microarray data contains a wealth of information. With many journals requiring this data to be deposited in public databases as a condition of publishing, a good deal of information is now publicly available. Data generated from a specific experiment could be of interest to other researchers investigating very different questions and information gleaned from multiple experiments can be combined, increasing the surety of the results. Public microarray databases can be used to determine what happens to your gene(s) of interest during a specific growth stage or under stress conditions. Conversely, the databases can be mined to look at genome-wide changes in gene expression. Using a database is also effective in terms of time and money to pare down large candidate gene lists before validating function. Many tools are available for analyzing publicly available databases. One such tool, Genevestigator, was presented in our September 2004 issue in the article ''GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox'' by Zimmermann et al. As of July 2006, it had been cited 144 times according to Thompson ISI (Thompson ISI Web of Science, http://www.isinet.com).

BACKGROUND

Genevestigator (https://www.genevestigator.ethz.ch/) is a publicly available microarray database coupled with expression-data analysis tools. The online analysis tools allow a large range of questions to be asked about gene expression during developmental stages, stress conditions, or by tissue/organ specificity for either specific genes or for exploring more global expression patterns.

The program was validated with several genes with known expression patterns. In all instances the expected gene-expression pattern was obtained using the expression-data analysis tools, demonstrating that the tools produced accurate, reproducible results. As more microarray data becomes available and the size of the dataset increases, the quality of the information from these programs will continue to improve. Currently, there are 2,620 Arabidopsis microarray chips available for query on the Genevestigator Web site.

The database was initially set up to focus on a single organism (Arabidopsis) and to utilize data generated on the same platform (Affymetrix GeneChip) to ensure obtaining high-quality results using the analysis tools, which the authors believe will allow the ''identification of biologically meaningful expression patterns of individual genes'' (pp. 2621–2622). Thus, at present, comparing array data from other species with those from Arabidopsis is not possible with Genevestigator, but the database is being extended to other organisms such as mouse, for which 3,110 arrays are already available (Laule et al., 2006).

When array analysis tools rely on information from public databases, care must be exercised when selecting which data to include in the database, as data from microarrays potentially can be of low technical quality either due to microarray itself or sample quality. Low-quality data will have a negative impact on results by giving erroneous associations and can also lead to problems with reproducibility (Rensink and Buell, 2005). Thus, the quality of data, along with annotation, needs to be assessed before inclusion. To overcome this potential problem, all of the array data included in the Genevestigator database have been manually assessed.

Another important consideration is the signal intensity of a given gene. Genes that are weakly expressed will have higher background and can give false reports. When selecting which chips to use for analysis, the numbers available is included and should be taken into consideration when interpreting results from any of the tools in the toolbox.

The original Genevestigator toolbox contained six analysis tools enabling users to make queries about signal intensity values for individual genes or to take a more ''genome-centric'' approach for chosen criteria and get a list of genes expressed under those conditions (Zimmermann et al., 2004). With the 2005 update, the database was expanded, the existing tools upgraded, and a new tool (Mutant Surveyor) was added. Documentation and FAQ sections were also updated (Zimmermann et al., 2005). The Documentation section contains additional information about the tools and the data, as well as a section on tips, pitfalls, and precautions to help avoid common misconceptions about results from the tools.

Currently, the toolbox consists of eight analysis tools, briefly outlined below. Since the available arrays include data using both wild-type and mutant plants, users have the option to select all available arrays, wild-type only, or Columbia wild-type only, with the obvious exception of the Mutant Surveyor tool.

The tool Digital Northern can answer either of the following questions: ''How is my gene of interest (or a set of genes) expressed throughout selected experiments?'' or "In which arrays is my gene of interest most strongly expressed?" The user selects GeneChip experiments from a menu that fits their criteria of developmental stage, organ type, or environmental factors, and inputs

www.plantphysiol.org/cgi/doi/10.1104/pp.104.900198.

up to 10 gene identifiers. The resulting signal intensity data is returned in either graph or tabular form.

To investigate how two genes are coexpressed over selected arrays, the signal intensity values of two genes are compared in Gene Correlator within the user-selected experiments.

Gene Atlas answers the questions ''How strongly is my gene of interest expressed in different organs or tissues?'' and ''Which genes are expressed preferentially in a selection of organs or tissues?'' In this tool, the organs are organized into groups each containing subgroups of specific organs. Selecting the main organ group would include all the chips of the subgroup in addition to all those from whole organ extractions. The subgroups would be of the specific subcategory only. An example would be the group ''seedlings.'' By selecting ''seedlings,'' chips representing whole seedlings, as well as those from the subgroups of cotyledons, hypocotyls, and radicles, would be included. In contrast, if the subcategory ''hypocotyl'' is selected, only those chips containing hypocotyl material would be include.

The Gene Chronologer tool addresses queries about the expression of a gene of interest at a specific growth stage or, more globally, which genes are expressed during a growth stage. Growth stages are grouped into 10 subcategories from seed germination to plant senescence. Each growth stage has the number of chips available, and users are cautioned to exercise care in the interpretation of genes that are not heavily represented.

Response Viewer is used for making queries based on single genes or global queries on which genes respond to a specific or combined stresses. The corresponding control for the stress exposure is also available. Experiments where multiple treatments were used are not included.

Gene Atlas, Gene Chronologer, and Response Viewer do not allow multiple gene queries; only a single gene identifier can be entered. Meta-Analyzer is similar to the other three tools in that the expression profiles of genes from organ type, stress response, or growth stage can be investigated, but it also allows multiple genes to be queried simultaneously.

Mutant Surveyor was added in 2005 and demonstrates how a mutation can alter expression of a gene of interest affected (Zimmermann et al., 2005).

The Gene Annotator tool provides ontologies and annotations for genes. The gene ontology annotations are from The Arabidopsis Information Resource (TAIR), and the user can select from biological process, molecular function, or cellular component.

THE IMPACT

This program was conceived with the goal of helping researchers put their gene-expression data into context, allowing them to validate hypotheses and generate new ones, enabling further, more directed research into gene function. This technical validation of gene expression with the microarray databases, although it does not validate gene function, confirms gene expression and can identify candidate genes for further studies of gene function (Clarke and Zhu, 2006).

Information from Genevestigator has been used for just such instances, to support experimental findings on gene expression, as well as to demonstrate where a gene is expressed or confirm gene expression in a particular tissue type. It has also been used to determine the expression of a gene in mutant backgrounds (McGrath et al., 2005).

Another goal of Zimmermann et al. was that Genevestigator would allow the building of hypotheses about gene expression. A study on folate transport into chloroplasts by Bedhomme et al. (2005) used Genevestigator in tandem with quantitative RT-PCR analysis to determine that their gene of interest is constitutively expressed at all growth stages, allowing them to hypothesize when the phenotype of the null mutant could be detected.

Expression data from Genevestigator have been used with expression data obtained from Massively Parallel Signature Sequencing (MPSS), a quantitative measure of gene expression from a particular tissue. McCormack et al. (2005) used primary sequences available for known calmodulin and calmodulin-like genes, and compared expression data from the MPSS database (Meyers et al., 2004; http://mpss.udel.edu/at/) with that from the microarray database available through Genevestigator to determine expression patterns during development, as well as organ specificity and stimulus response. Although there were some discrepancies, both techniques yielded similar findings.

Genevestigator was also used in parallel with MPSS and whole-genome arrays to demonstrate the expression of galacturonosyltransferase (GalAT) superfamily members to add support to a hypothesis about pectin synthesis (Sterling et al., 2006). Galacturonic acid is a main component of pectin and is present in all three types of pectin. Although the activity of GalATs has been detected, no gene had been identified for a GalAT that was enzymatically verified. Genevestigator was one of the bioinformatics tools that aided in the functional identification of a GalAT involved in the biosynthesis of the pectin homogalacturonan.

CONCLUDING REMARKS

There are other online tools available for analyzing Arabidopsis microarray data, such as TAIR (Rhee et al., 2003), MAPMAN (Thimm et al., 2004), and The Botany Array Resource (Toufighi et al., 2005). As more array data become available for other plants, so are online analysis options such as BarleyBase (Shen et al., 2005) and Sol Genomics Network for members of the Solanaceae family (Mueller et al., 2005). Each of these online analysis suites and Genevestigator offer different tools and variations in the databases. As more

gene-expression data become available and statistical methods for comparing data originated from different technologies advances, the accuracy of ''virtual laboratories'' will continue to improve, enabling further advancement of functional genomics.

LITERATURE CITED

Bedhomme M, Hoffmann M, McCarthy EA, Gambonnet B, Moran RG, Rebeille F, Ravanel S (2005) Folate metabolism in plants: an Arabidopsis homolog of the mammalian mitochondrial folate transporter mediates folate import into chloroplasts. J Biol Chem **280:** 34823–34831

Clarke JD, Zhu T (2006) Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives. Plant J **45:** 630–650

Laule O, Hirsch-Hoffmann M, Hruz T, Gruissem W, Zimmermann P (2006) Web-based analysis of the mouse transcriptome using Genevestigator. BMC Bioinformatics **7:** 311

McCormack E, Tsai YC, Braam J (2005) Handling calcium signaling: Arabidopsis CaMs and CMLs. Trends Plant Sci **10:** 383–389

McGrath KC, Dombrecht B, Manners JM, Schenk PM, Edgar CI, Maclean DJ, Scheible WR, Udvardi MK, Kazan K (2005) Repressor- and activator-type ethylene response factors functioning in jasmonate signaling and disease resistance identified via a genome-wide screen of Arabidopsis transcription factor gene expression. Plant Physiol **139:** 949–959

Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, Ning JC, Haudenschild CD (2004) Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. Nat Biotechnol **22:** 1006–1011

Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, et al (2005) The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. Plant Physiol **138:** 1310–1317

Rensink WA, Buell CR (2005) Microarray expression profiling resources for plant genomics. Trends Plant Sci **10:** 603–609

Rhee SY, Beavis W, Berardini TZ, Chen GH, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, 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

Shen L, Gong J, Caldo RA, Nettleton D, Cook D, Wise RP, Dickerson JA (2005) BarleyBase—an expression profiling database for plant genomics. Nucleic Acids Res (Database issue) **33:** D614–D618

Sterling JD, Atmodjo MA, Inwood SE, Kolli VSK, Quigley HF, Hahn MG, Mohnen D (2006) Functional identification of an Arabidopsis pectin biosynthetic homogalacturonan galacturonosyltransferase. Proc Natl Acad Sci USA **103:** 5236–5241

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, 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

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analysis. Plant J **43:** 153–163

Zimmermann P, Hennig L, Gruissem W (2005) Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci **10:** 407–409

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

Aleel K. Grennan
University of Illinois
Urbana, IL