Genevestigator
arabidopsis microarray database and analysis toolbox

Report

Author(s):
Zimmermann, Philip; Hirsch-Hoffmann, Matthias; Henning, Lars; Gruissem, Wilhelm

Publication date:
2004

Permanent link:
https://doi.org/10.3929/ethz-a-005223225

Rights / license:
In Copyright - Non-Commercial Use Permitted
High-throughput gene expression analysis has become a frequent and powerful research tool in biology. At present, however, few software applications have been developed for biologists to query large microarray gene expression databases using a Web-browser interface. We present GENEVESTIGATOR, a database and Web-browser data mining interface for Affymetrix GeneChip data. Users can query the database to retrieve the expression patterns of individual genes throughout chosen environmental conditions, growth stages, or organs. Reversely, mining tools allow users to identify genes specifically expressed during selected stresses, growth stages, or in particular organs. Using GENEVESTIGATOR, the gene expression profiles of more than 22,000 Arabidopsis genes can be obtained, including those of 10,600 currently uncharacterized genes. The objective of this software application is to direct gene functional discovery and design of new experiments by providing plant biologists with contextual information on the expression of genes. The database and analysis toolbox is available as a community resource at https://www.genevestigator.ethz.ch.

A major challenge in biology today is the large-scale determination of gene function (Boyce et al., 2001). First, the establishment of standards and controlled vocabularies facilitates the integration of experimental data into a computational framework, thereby allowing structured and systematic processing of information (Ashburner et al., 2000; Brazma et al., 2001). Second, structured databases and data querying tools provide the means to assign putative functional information to genes.

The complete sequencing of the Arabidopsis genome achieved in the year 2000 (The Arabidopsis Genome Initiative, 2000) enables us to monitor gene expression of this flowering plant on a genome-scale using microarrays. In situ synthesis of high-density oligonucleotides on glass slides (Lockhart et al., 1996) has become a powerful tool to rapidly integrate the sequence knowledge into expression profiling platforms, such as the ATH1 full genome array developed by Affymetrix and The Institute for Genomic Research (TIGR), which represents approximately 23,750 genes from Arabidopsis (Redman et al., 2004). The availability of a full-genome array and the complete technical environment provided by the Affymetrix system led to a wide use of the GeneChip technology in the plant community. Thousands of arrays have since been processed, of which a significant number are publicly available through services and repositories such as Nottingham Arabidopsis Stock Centre Transcriptomics Service (NASCArrays; Craigon et al., 2004), ArrayExpress at the European Bioinformatics Institute (EBI; Brazma et al., 2003), or Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI; Edgar et al., 2002).

The exploitation of large-scale gene expression datasets, mainly from Saccharomyces cerevisiae and Escherichia coli, has already led to the discovery of global structures governing metabolic and regulatory networks (Lee et al., 2002; Ravasz et al., 2002; Stelling et al., 2002; Ihmels et al., 2004). Multiple-genome comparisons have also yielded interesting observations on the modularity and connectivity distributions of gene expression data (Bergmann et al., 2004). Nevertheless, the combination of multiple datasets still raises a number of questions concerning their compatibility, in particular when comparing data from different platforms and organisms. While analyses revealing global properties of networks or modules may not necessarily require full compatibility of expression datasets, the details are often noisy (Friedman, 2004) and the comparative search for the function of individual genes requires a more stringent selection.

The Affymetrix platform provides a standardized system with a high degree of reproducibility (Hennig et al., 2003; Redman et al., 2004). Although data from different experiments may not be pooled for a rigorous expression profiling analysis, one can assume that the large-scale combination and analysis of expression data from a single organism using a single platform like the Affymetrix system allows the identification of biologically meaningful expression patterns of...
individual genes. To date, few tools have been developed for biologists to query large gene expression databases. The Yeast Microarray Global Viewer (yMGV) is a database providing online tools for the analysis of transcriptional expression profiles of yeast genes among 82 different datasets (Lelandais et al., 2004). In the plant community, NASCArrays (Craigon et al., 2004) provides a repository for Arabidopsis microarray data and some simple “gene-centric” data mining tools.

Here, we describe a novel online tool called GENEVESTIGATOR comprising a gene expression database and a number of querying and analysis functionalities developed to facilitate gene functional discovery. GENEVESTIGATOR allows the data to be presented in the context of plant development, plant organ, and environmental conditions, both for individual genes or for families of genes, thereby answering questions such as “in which growth stage is my gene of interest expressed?” or “which genes are specifically expressed in roots?” The main objective of the software is to assign contextual information to gene expression data, directing the design of new experiments and gene functional discovery.

RESULTS

Database Concept and Software Design

GENEVESTIGATOR was conceived as a user-friendly online tool for large-scale expression data analysis. It consists of a MySQL relational database and a Web server application programmed in the PHP (PHP Hypertext Preprocessor) scripting language. The database works as a “data warehouse” containing experimental and annotation data, preprocessed data, as well as diverse tables for control of workflow and analysis (Fig. 1).

Raw experimental data from users is processed using Affymetrix MAS 5.0 software to a target value (TGT) of 1,000 (Liu et al., 2002). Signal intensities and P values are collected for each hybridized Affymetrix GeneChip array. Alternatively, data and annotation can be imported from public repositories such as ArrayExpress (Brazma et al., 2003) and GEO (Edgar et al., 2002). The assignment of array elements (probe sets) to Arabidopsis locus identifiers (AGI codes) and their annotations is based on regularly updated data sets obtained from the Arabidopsis Information Resource (TAIR) ftp server (ftp://ftp.arabidopsis.org/home/tair/Microarrays/Affymetrix/; currently as of April 5, 2004, based on the final Arabidopsis genome annotation release from TIGR [version 5.0, January 2004]). In addition to probe sets representing unique genes (ending “_at”), the ATH1 and AG GeneChip arrays include nonunique probe sets representing two or more closely related genes (ending “_s_at”) or multiple cross-hybridizing probe sets (ending “_x_at”; for details, see Redman et al., 2004). Although these probe set types represent two or more genes, only one locus identifier is displayed per probe set. These ambiguous probe sets are highlighted in GENEVESTIGATOR to draw the attention of the user to this issue.

The experiment annotation is curated, entered, and structured in either hierarchical (e.g. plant organs), unique (e.g. growth stage), or multi-select form (e.g. environmental condition). The software has been designed for easy additions of new annotations in any of these formats and for rapid creation of the corresponding tools to analyze and visualize the data. The annotation of arrays was based on the information provided by users or public repositories. Missing information does not impact the results, as the corresponding arrays are not included into the respective calculations. Ambiguous or unsuitable annotations were further ignored. For example, arrays from RNA extracted from whole adult plants (including roots, rosette leaves, and inflorescence) are unsuitable for tools relating to plant organ specificity (Gene Atlas) and are therefore not included into the corresponding calculations, but may be proper for use in other tools such as Gene Chronologer. Each tool therefore accesses the best respective available sources of data for processing, while unsuitable data is ignored.

Data from the ATH1 and AG arrays are processed separately. Different sets of oligonucleotide sequences are used to probe identical target genes on the two array types, and thus different efficiencies of target to probe hybridization and nontarget to probe cross-hybridization makes a direct comparison of signal intensities impossible. Although a high degree of reproducibility was found for most target genes probed by both the ATH1 and the AG arrays, 300 pairs of probe set for identical target genes yielded strongly differing results (Hennig et al., 2003).
As of July 2004, the database contained publicly available data from 750 ATH1 and 121 AG arrays covering 81 public experiments from the Gruissem Laboratory (http://www.pb.ethz.ch; Menges et al., 2003; Hennig et al., 2004; Kleffmann et al., 2004), the Functional Genomics Center Zurich (http://www.fgcz.ethz.ch), NASCArrays (http://ssbdjc2.nottingham.ac.uk/narrays/experimentbrowse.pl; Craigon et al., 2004), ArrayExpress at EBI (http://www.ebi.ac.uk/arrayexpress/; Brazma et al., 2003), and from GEO at NCBI (http://www.ncbi.nlm.nih.gov/geo/; Edgar et al., 2002).

GENEVESTIGATOR is freely accessible to all academic institutions. Since the database contains at present both publicly available as well as confidential data, we have implemented a dual user profile management system for public and private users. All users are therefore asked to register once and to login for each session. We limit the collection and use of personal information to what is necessary to administer the database and improve the utility of GENEVESTIGATOR. Personal information is not shared with third parties.

Analysis Tools

The GENEVESTIGATOR tools generally contain two types of queries: a gene-centric approach reporting signal intensity values for individual genes, and a genome-centric approach providing lists of genes fulfilling chosen criteria. The results obtained from any tool are based on all available signal intensity values and the corresponding annotations. In some cases, present/absent call information as defined by the MAS5.0 algorithm is indicated (see below).

The first tool, Digital Northern, will retrieve the signal intensity values of input genes for a chosen selection of GeneChip experiments. An elaborate selection tool (Fig. 2A) allows the user to choose exactly those experiments that fit single or multiple criteria such as anatomy, growth stage, or environmental factors. Up to 10 probe sets can be processed simultaneously, displayed in several colors, shapes, and filling, revealing both signal intensity values and present call (closed symbols) and absent call (open symbols) information (Fig. 2B).

The Gene Correlator allows comparing the signal intensity values of two genes throughout all chosen experiments (Fig. 2C; identical selection tool as for Digital Northern). Each spot represents a GeneChip and can be identified by mouse-over or by linking to the annotation database. The Pearson's correlation coefficient is given as a measure for the relationship between expression signals of two genes. Present call information is visualized by a color coding (Fig. 2C).

Because the objective of the software was to provide contextual information for the expression of genes, we additionally focused on relating gene expression to three main annotation groups: plant organ, developmental stage, and environmental stress.

The Gene Atlas tool similarly provides the average signal intensity values of a gene of interest in all organs or tissues annotated in the database (Fig. 2D). Reversely, GENEVESTIGATOR can output lists of genes for which signal intensities exceed a chosen threshold in selected organs versus a baseline choice of organs (Fig. 2E). This allows users to find genes expressed preferentially in certain organs or tissues, such as roots, young leaves or stamina. The anatomy annotation was based on standard anatomy terms as defined by the Plant Ontology Consortium (http://www.plantontology.org/) that we classified into six main groups (callus, cell suspension, seedling, inflorescence, rosette, and roots) and the corresponding subgroups. These categories cover all tissues that can currently be isolated for expression analysis, but can easily be extended as tissue and cell separation techniques become more precise (Birnbaum et al., 2003).

The Gene Chronologer tool, based on the Boyes growth stage ontology (Boyes et al., 2001), possesses two main features. First, it outputs the average signal intensities (or expression levels) and SES of a gene of interest for 10 representative sections of the life cycle of Arabidopsis (Fig. 2F). Second, users can query the database to output all genes expressed above a given threshold at chosen growth stages. For example, all genes can be selected for which the signal intensity at the seedling stage exceeds 90% of the sum of all average signal intensity values for each category, measured for this gene throughout the life cycle of the plant (Fig. 2G).

The Response Viewer tool provides the same functionalities as Gene Atlas and Gene Chronologer, based on stress response annotations (Fig. 2, H and I). For each condition, one or several representative experiments were chosen. Each stress factor is given with the corresponding control from these experiments, allowing direct comparison.

The Meta-Analyzer utility has been designed to study the gene expression profiles of several genes simultaneously in the context of environmental stresses, organs, and growth stages (Fig. 2, J–L). Lists of genes can be entered in diverse formats (comma-, semi-colon-, or space-separated, CRLF [carriage return, line feed], or directly copied from a spreadsheet). The output is a heat map of normalized signal intensity values (see Documentation section on our Web page) clustered by either single, average, or complete linkage hierarchical clustering. This tool is especially useful to compare members of gene families and to identify clusters of similarly expressed genes.

Finally, the Database and Documentation sections provide users with annotation information about experiments in the database, as well as technical information (Fig. 2, M and N). Since GENEVESTIGATOR was conceived to be an analysis tool and not a data repository, a reduced set of annotations is stored locally. The full MIAME (Minimum Information About a Microarray Experiment) compliant annotations (Brazma et al., 2001) are available by linking to the
Figure 2. Screenshots of some of the features of GENEVESTIGATOR. Top left, Logo and available tools. A, Chip Selection tool; B, Digital Northern; C, Gene Correlator; D and E, Gene Atlas (relates to plant anatomy); F and G, Gene Chronologer (relates to the plant growth stages); H and I, Response Viewer (relates to environmental factors); J to L, Meta-Analyzer (multiple gene analysis with respect to anatomy, growth stage, and environmental factors); M and N, Database tool for viewing experiment and array annotation, and Documentation section for user information.
corresponding repository sites from which the experiments were downloaded.

General Approach and Validation

The database contains expression data from a high diversity of experiments covering different tissues, ages, and treatments (Table I). The general hypothesis in our approach is that as the number of experiments per category (e.g. growth stage 5.10) increases, individual effects are averaged out and global trends become visible. As a measure of confidence for the expression of genes in different categories, we indicate the respective number of GeneChips and the SE of the mean for each category.

To validate our hypothesis, we checked whether strongly populated categories yield results that are consistent with the literature. In a first step, we selected a number of marker genes with preferential expression in particular organs, at specific growth stages, or in response to certain stresses and then analyzed their expression patterns generated by GENEVESTIGATOR. Marker genes were chosen from the literature.

First, using Gene Atlas, three AGAMOUS-like genes known to be preferentially expressed in roots as measured by reverse transcription-PCR (AGL12 [At1g71692], AGL14 [At4g11880], and AGL17 [At2g22630]; Parenicova et al., 2003) in fact showed strong expression in roots and radicle, but weaker signals in all other organs (Fig. 3, A–C). Two genes associated with pollen tube growth (At1g55570, Albani et al., 1992; and At2g25600, Mouline et al., 2002) were also identified as being specific to stamina (and by extension to the categories “flower” and “inflorescence”) in our expression database (Fig. 3, D and E). Furthermore, two genes involved in photosynthesis (chlorophyll a/b binding proteins, At1g19150 and At3g08940) were found to be abundantly expressed in green plant tissues (rosette, cauline leaf, stem, node, flower, cotyledon, and hypocotyl), but lowly expressed in photosynthetically inactive tissues (roots, stamen, and seeds; Fig. 3, F and G). This pattern was observed for all genes from the chlorophyll a/b binding family except for one gene (TAIR; http://www.arabidopsis.org/info/genefamily/Chloroplast.html; see Supplemental Table II, available at www.plantphysiol.org).

Second, to verify the reliability of the Gene Chronologer tool, we looked for genes annotated as being developmentally regulated. Two genes involved in seed germination and seedling development (encoding the embryonic abundant protein ATEM1 [AT3G51810,
Vicient et al., 2000] and a gene involved in apical hook development [At4g37580, Lehman et al., 1996]) showed highest expression during mature seed and germination stages (Fig. 3, H and I), but lower levels in all other stages. In contrast, two genes involved in flowering (APETALA1 [At1g69120, Pelaz et al., 2001] and FLOWERING LOCUS T [At1g65480, Ruiz-Garcia et al., 1997]) were shown to be most abundantly expressed in the flowering stages (Fig. 3, J and K).

Third, the Response Viewer tool was used for several genes known to be responsive to particular stresses (Fig. 3, L–Q). GENEVESTIGATOR correctly showed the expression pattern of a light-induced gene encoding a light-harvesting chlorophyll a/b binding protein (AT4G14690, Jansson et al., 2000) and of the light-repressed protochlorophyllide reductase A gene (At5g54190, Runge et al., 1996; Fig. 3, L and M, respectively). Similarly, four genes reported to be
Table IIA. Representative samples of genes expressed in specific tissues or at particular growth stages

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Table continues on following page</td>
</tr>
</tbody>
</table>

- 2 seedling
- 21 cotyledons
- 22 hypocotyl
- 23 radicle
- 3 inflorescence
- 31 flower
- 311 carpel
- 312 petal
- 313 sepal
- 314 stamen
- 315 pedicel
- 32 siliqua
- 33 seed
- 34 stem
- 35 node
- 36 shoot apex
- 37 cauline leaf
- 4 rosette
- 41 juvenile leaf
- 42 adult leaf
- 43 petals
- 44 senescent leaf
- 5 roots
- 52 lateral root
- 55 elongation zone

- The table continues on the following page.
Table IIB.

| Probeset/Gene | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
| 284735_at | At1g20600 | 255915_at | At1g25510 | 260718_at | At1g49180 | 258224_at | At3g35090 | 264450_at | At3g11010 | 244440_at | At1g21020 | 256031_at | At3g22410 | 255041_at | At1g38000 | 263490_at | At3g28560 | 250999_at | At1g36180 | 255614_at | At3g21370 | 250949_at | At1g36910 | 263699_at | At1g49440 | 252577_at | At1g17010 | 252219_at | At3g20340 | 249915_at | At5g55690 | 251900_at | At3g39450 | 250070_at | At1g65510 | 249009_at | At5g43410 | 265211_at | At3g36640 | 257584_at | At3g18620 | 259594_at | At1g32560 | 260839_at | At1g73160 | 255007_at | At4g10330 | 235494_at | At4g16150 | 255961_at | At3g15510 | 260718_at | At1g49180 | 249125_at | At5g47440 | 250394_at | At1g30880 | 249827_at | At3g20540 | 260763_at | At3g28490 | 253630_at | At1g26740 | 256442_at | At1g127360 | 251130_at | At1g60590 | 259240_at | At3g27060 | 265095_at | At1g20900 | 259304_at | At1g27140 | 254056_at | At4g25150 | 253902_at | At1g27170 | 253859_at | At1g27160 | 253894_at | At1g27150 | 262431_at | At1g47540 | 249353_at | At1g50420 | 259167_at | At3g38570 | 253767_at | At1g28530 | 249002_at | At1g54410 | 264745_s_at | At1g62060 | 264755_at | At5g68100 | 251202_at | At3g30340 | 264273_at | At1g39700 | 261699_at | At3g39900 | 260050_at | At3g20710 | 249611_at | At5g29120 | 249754_at | At1g57160 | 264605_at | At1g04960 | 254467_at | At5g38160 | 262758_at | At3g36960 | 258850_at | At5g32460 | 243468_at | At5g07050 | 254740_at | At1g16270 | 261940_at | At1g11990 | 257944_at | At3g21850 | 251125_at | At1g23590 | 256498_at | At1g32910 | 254461_at | At1g17270 | 252090_at | At1g61810 | 262530_at | At1g18610 |

*Annotatio*
responsive to ethylene (ERF1 [At3g23240]; AtERF1 [At4g17500]; AtERF2 [At5g47220]; and AtERF13 [At2g44840]) were correctly found by the software to be responsive to ethylene and to the pathogen Pseudomonas syringae, as reported by the authors (Onate-Sanchez and Singh, 2002; Fig. 3, N–Q).

This first validation step confirms that global trends can be detected in the expression profiles of individual genes by combining numerous normalized expression data sets using the same technical platform, i.e. the Affymetrix system. Based on this information, we performed a second validation step, in which we tested whether GENEVESTIGATOR can identify genes with known expression profiles. Using Gene Atlas, 72 genes were identified to be expressed in pollen. Of these, 9 had been identified by Honys and Twell (2003) as well as Becker et al. (2003) to be pollen-specific using 8K Arabidopsis Genome Arrays (see Table IIA; Supplemental Table II). Of the remaining genes, several could be functionally associated with pollen based on annotations such as “self-incompatibility protein,” “pollen coat protein-related,” or “allergen.” Further, 14 genes were annotated as “expressed protein,” revealing the potential of GENEVESTIGATOR to identify novel genes related to particular organs. A similar analysis was performed to identify genes expressed specifically in siliques (Table IIB, compare with Hennig et al., 2004), roots, photosynthetic active tissues, leaves, senescent leaves, stem and node, carpel, petal, sepal, and shoot apex (see Supplemental Table II) and at specific developmental stages such as seedling stage (Table IIC) or early flowering stage (Table IID; Supplemental Table II). We conclude that with the current set of data, GENEVESTIGATOR generates high quality results. Moreover, we expect that this quality will continue to rise as the size of the dataset increases.

**DISCUSSION**

Public repositories such as GEO and ArrayExpress provide tools for submission, storage, and retrieval of heterogeneous data sets. In contrast, GENEVESTIGATOR contains a coherent data set from a single organism generated on a common hybridization platform. Despite the high diversity of experiments represented in the database, the validation steps we carried out demonstrate that the underlying hypothesis is valid and that biologically meaningful results can be obtained.
using GENEVESTIGATOR. The software generally performs primary level analysis and displays results either as graphs or as numeric data, which can easily be combined, exported, or further analyzed with other data analysis and visualization tools.

The complexity of multicellular life requires the proper context-dependent expression of genes, which is achieved by highly interconnected transcriptional networks. The inference of such module networks may require the use of many data types such as gene expression, protein abundance, protein interaction, metabolite abundance, affinity precipitation, synthetic lethality, etc. (Troyanskaya et al., 2003). Nevertheless, the analysis of gene expression data can reveal significant patterns of such networks (Segal et al., 2003). In contrast to many other tools, GENEVESTIGATOR uses experiment annotation to yield contextual information that can be brought into understanding gene networks. The identification of genes exhibiting similar tissue localization and stress response attributes facilitates modeling of gene networks using network inference tools (Wille et al., 2004) by reducing the number of testable candidates. Thus, the combined gene-centric and genome-centric approaches make it a powerful tool for targeted functional genomics efforts.

Critical issues in using the GENEVESTIGATOR tools are (1) the questions being addressed by queries and (2) the interpretation of output data. First, GENEVESTIGATOR allows queries at a high level of detail and in a large variety of combinations specifying organ, developmental stage, or treatment. Although GENEVESTIGATOR currently contains information from more than 750 publicly available full genome arrays, some combinations at very detailed level may not yet have sufficient data support to yield robust results. The quality of the results therefore depends strongly on the level of granularity the user chooses and the number and types of underlying experiments. Second, care must be taken not to over-interpret

Table IID.

<table>
<thead>
<tr>
<th>Probeset/AGI</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>26297_at/A1g075040</td>
<td>glycosyl hydrolase family 1 protein/a...</td>
</tr>
<tr>
<td>25722_at/A1g27010</td>
<td>myb family transcription factor (MYB3)...</td>
</tr>
<tr>
<td>26388_at/A1g68850</td>
<td>protease inhibitor/seed storage protein...</td>
</tr>
<tr>
<td>26003_at/A1g66975</td>
<td>expressed protein</td>
</tr>
<tr>
<td>26265_at/A1g715930</td>
<td>family II extracellular lipase 6 (EXL6)...</td>
</tr>
<tr>
<td>245622_at/A4g14080</td>
<td>glycosyl hydrolase family 17 protein f/...</td>
</tr>
<tr>
<td>264430_at/A1g16180</td>
<td>terpene synthase/kyrase family protein...</td>
</tr>
<tr>
<td>255101_at/A4g08670</td>
<td>protease inhibitor/seed storage protein1...</td>
</tr>
<tr>
<td>249048_at/A5g44300</td>
<td>dormancy/auxin associated family protein...</td>
</tr>
<tr>
<td>261532_at/A1g71680</td>
<td>lyase and histidine specific transport...</td>
</tr>
<tr>
<td>259265_at/A3g01250</td>
<td>expressed protein</td>
</tr>
<tr>
<td>249538_at/A5g38760</td>
<td>expressed protein similar to ABA-induc...</td>
</tr>
<tr>
<td>200308_at/A1g70540</td>
<td>invertaseependent methyltransferase inhibit...</td>
</tr>
<tr>
<td>245232_at/A4g25590</td>
<td>actin-depolymerizing factor</td>
</tr>
<tr>
<td>248367_at/A5g23260</td>
<td>actin-depolymerizing factor</td>
</tr>
<tr>
<td>22156_at/A1g25260</td>
<td>late embryogenesis abundant protein-rel...</td>
</tr>
<tr>
<td>261015_at/A1g28480</td>
<td>14-3-3 protein GF14 hto (GRF12) identi...</td>
</tr>
<tr>
<td>262742_at/A1g28550</td>
<td>Ras-related GTP-binding protein</td>
</tr>
<tr>
<td>267449_at/A2g33690</td>
<td>late embryogenesis abundant protein</td>
</tr>
<tr>
<td>265473_at/A1g25535</td>
<td>SLR1 binding protein core protein-relate...</td>
</tr>
<tr>
<td>264603_at/A1g04570</td>
<td>expressed protein</td>
</tr>
<tr>
<td>235151_at/A4g35570</td>
<td>glucoside hydrolase family 26 protein f/...</td>
</tr>
<tr>
<td>261943_at/A1g00660</td>
<td>ATPase 9</td>
</tr>
<tr>
<td>246227_at/A5g35020</td>
<td>expressed protein similar to ABA-induc...</td>
</tr>
<tr>
<td>241501_at/A5g43340</td>
<td>inorganic phosphate transporter identic...</td>
</tr>
<tr>
<td>265552_at/A2g07560</td>
<td>ATPase</td>
</tr>
<tr>
<td>246431_at/A5g17480</td>
<td>polcalon</td>
</tr>
<tr>
<td>259633_at/A3g60900</td>
<td>expressed protein</td>
</tr>
<tr>
<td>254033_at/A4g28590</td>
<td>vascular ATP synthase</td>
</tr>
<tr>
<td>265682_at/A3g28790</td>
<td>expressed protein</td>
</tr>
<tr>
<td>265682_at/A3g28840</td>
<td>expressed protein</td>
</tr>
<tr>
<td>265687_at/A5g23870</td>
<td>glycine-rich protein similar to H41 gen...</td>
</tr>
<tr>
<td>251988_at/A3g53300</td>
<td>cytochrome P450 family protein CYTOCHR...</td>
</tr>
</tbody>
</table>

Genes expressed preferentially (A) in stamina and pollen, (B) in seeds and siliques, (C) during seedling stage, and (D) during early flowering stage. For the description of growth stage groups (labeled A-J), see Table I. See also Supplemental Table II, which provides lists of genes expressed preferentially in roots, green tissues, photosynthetic active leaves, senescent leaves, stem and node, carpel, petal, sepal, and shoot apex.
output data computed by GENEVESTIGATOR. To facilitate data interpretation, the number of samples per category and the SEs of the means are indicated. Nevertheless, when working in a detailed level of granularity, a post-verification of individual genes is advised using the Digital Northern tool to confirm the origin of the effects observed.

CONCLUSION
Both the forward and reverse validation of GENEVESTIGATOR revealed that the combination of annotated data from various sources using the same technology platform is a valid approach to reveal contextual information about elements of the dataset. In our case, the expression profiles of more than 22,000 genes from Arabidopsis can be generated in the context of plant organ, plant development and environmental stress. Although not all annotated categories are currently well covered in terms of number of arrays, and therefore the output from these categories may be somewhat biased, the general quality of results obtained using GENEVESTIGATOR is high. The permanent submission of new datasets is expected to constantly improve the quality of the output. The resulting information can be used to confirm previous hypotheses or generate new hypotheses about gene expression network structures and genetic regulatory networks, resulting in the design of more precise and targeted experiments.

ACKNOWLEDGMENTS
We thank Eva Vranová and Franziska Humair for feedback on the use of the software in development. We are also grateful to the Functional Genomics Center Zurich for providing support and the Affymetrix platform for GeneChip experiments, as well as all public repositories for providing data.

Received May 14, 2004; returned for revision July 12, 2004; accepted July 16, 2004.

LITERATURE CITED

Zimmermann et al.