Genevestigator

arabidopsis microarray database and analysis toolbox

GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox[1][w]

Philip Zimmermann[2], Matthias Hirsch-Hoffmann[2], Lars Hennig, and Wilhelm Gruissem*

Institute of Plant Sciences, Swiss Federal Institute of Technology and Zurich-Basel Plant Science Center, ETH Center, CH–8092 Zurich, Switzerland (P.Z., M.H.-H., L.H., W.G.); and Functional and Genomics Center Zurich, UNI Irchel, Y32 H52, CH–8057 Zurich, Switzerland (W.G.)

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 (Boyes 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

[1] This work was supported by ETH, Strategic Excellence Project 2–74213–02/TH–8/02–2, and by the Functional Genomics Center Zurich.

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail wilhelm.gruissem@ipw.biol.ethz.ch; fax 41–1–632–10–79.

[w] The online version of this article contains Web-only data. www.plantphysiol.org/cgi/doi/10.1104/pp.104.046367.

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.

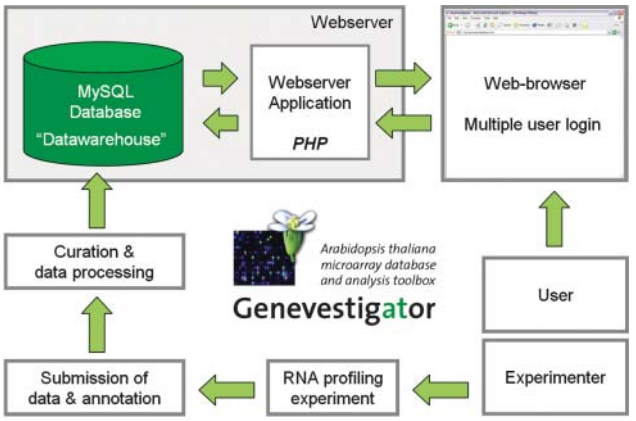


Figure 1. Concept and design of GENEVESTIGATOR. The experimenter submits RNA profiling data to the database curator, who processes the data and uploads it to the database. The datawarehouse contains raw signal intensity and *P* values, as well as preprocessed tables. A Webserver application acts as an interface between users and the GENEVESTIGATOR database.

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 datasets 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

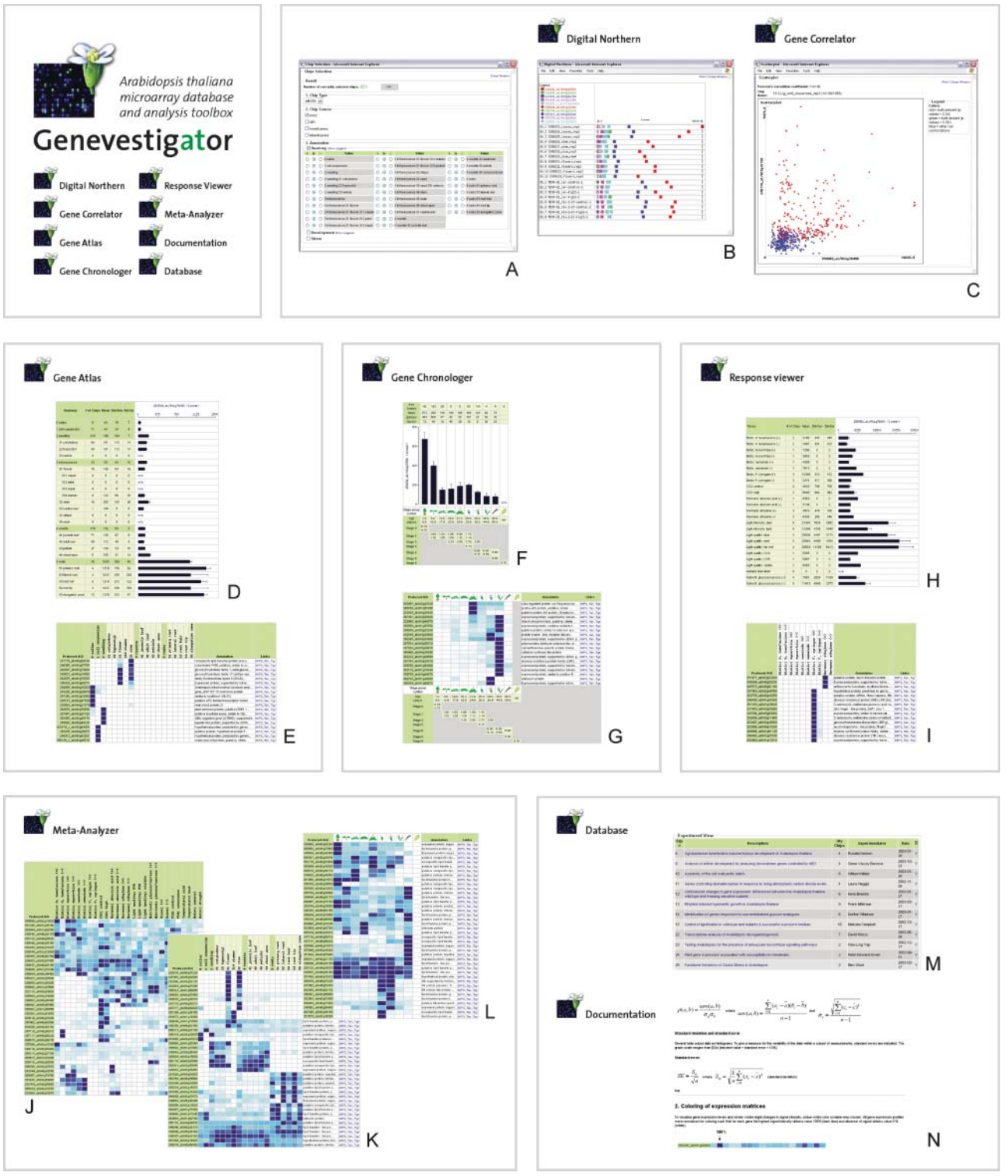


Figure 2. Screenshots of some of the features of GENEVESTIGATOR. Top left, Logo and available tools. A, Chip Selection tool; B, Digital Northern; C, Gene Correlator; D and E, Gene Atlas (relates to plant anatomy); F and G, Gene Chronologer (relates to the plant growth stages); H and I, Response Viewer (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,

Table I. *Annotation categories incorporated in GENEVESTIGATOR as of July 2004*

Plant Tissues/Organs	Developmental Stages	Environmental Factors (Continued)
0 Callus	10 Categories based on the	*Hormones*
1 Cell suspension	Boyes key ontology:	Ethylene
2 Seedling	A) 0.10 … 0.70	Auxin
21 Cotyledons	B) 1.00 … 1.02	Abscisic acid
22 Hypocotyl	C) 1.03 … 1.05	Gibberellin
23 Radicle	D) 1.06 … 1.08 / 3.20	*Atmosphere*
3 Inflorescence	E) 1.09 … 1.12 / 3.50	Ozone
31 Flower	F) 1.13 / 1.14 / 3.70 / 5.10	Carbon dioxide
311 Carpel	G) 3.90 / 6.00 / 6.10	
312 Petal	H) 6.30 / 6.50	*Illumination*
313 Sepal	I) 6.90 / 8.00	Light intensity
314 Stamen	J) 9.70	Light
315 Pedicel		Dark
32 Silique		Light quality
33 Seed		Far-red
34 Stem		Blue
35 Node		UVA
36 Shoot apex		UVB
37 Cauline leaf		Visible
4 Rosette	Environmental Factors	
41 Juvenile leaf		*Biotic interactions*
42 Adult leaf	*Nutrients/heavy metals*	*Pseudomonas syringae*
43 Petiole	Phosphate	*Gigaspora rosea*
44 Senescent leaf	Nitrate	*Agrobacterium tumefaciens*
5 Roots	Sulfate	*Heterodera schachtii*
51 Primary root	Potassium	*Erisyphe cichoracearum*
52 Lateral root	Water	*Programmed cell death*
53 Root hair	Suc/Glc	Senescence
54 Root tip	Lead	Heat
55 Elongation zone	Zinc	Cold

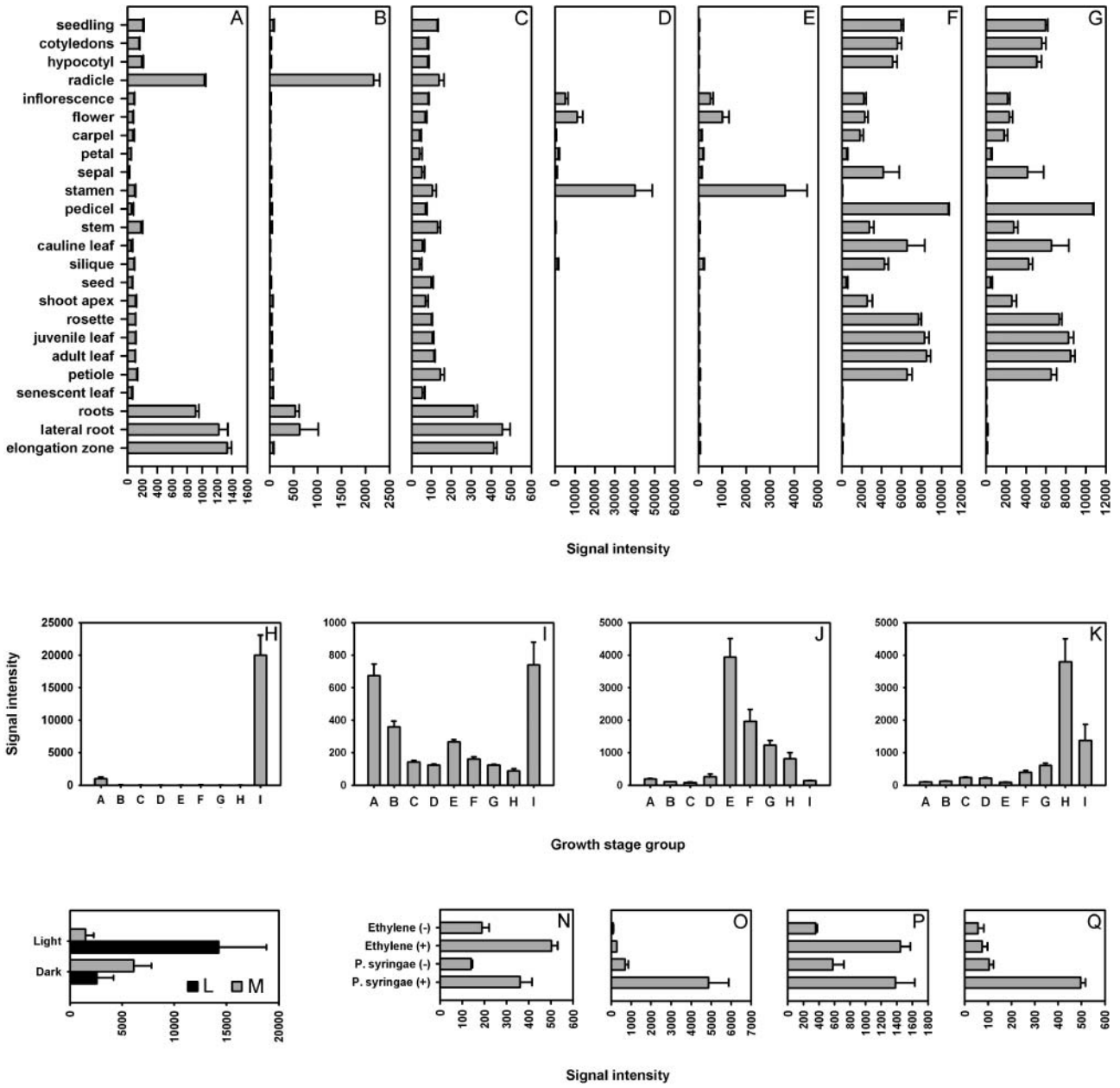


Figure 3. Validation of the quality of data generated by GENEVESTIGATOR. A to G, Expression of organ or tissue-specific marker genes used for testing the Gene Atlas tool (A, *AGL12*, At1g71692; B, *AGL14*, At4g11880; C, *AGL17*, At2g22630; D, At1g55570; E, At2g25600; F, At1g19150; G, At3g08940). H to K, Expression of growth stage specific marker genes used to validate the Gene Chronologer tool (H, *ATEM1*, At3g51810; I, At4g37580; J, *APETALA1*, At1g69120; K, *FLOWERING LOCUS T*, At1g65480). L to Q, Expression of environmental factor specific marker genes to validate the Response Viewer tool (L, At4g14690; M, At5g54190; N, *ERF1*, At3g23240; O, *AtERF1*, At4g17599; P, *AtERF2*, At5g47220; Q, *AtERF13*, At2g44840).

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).

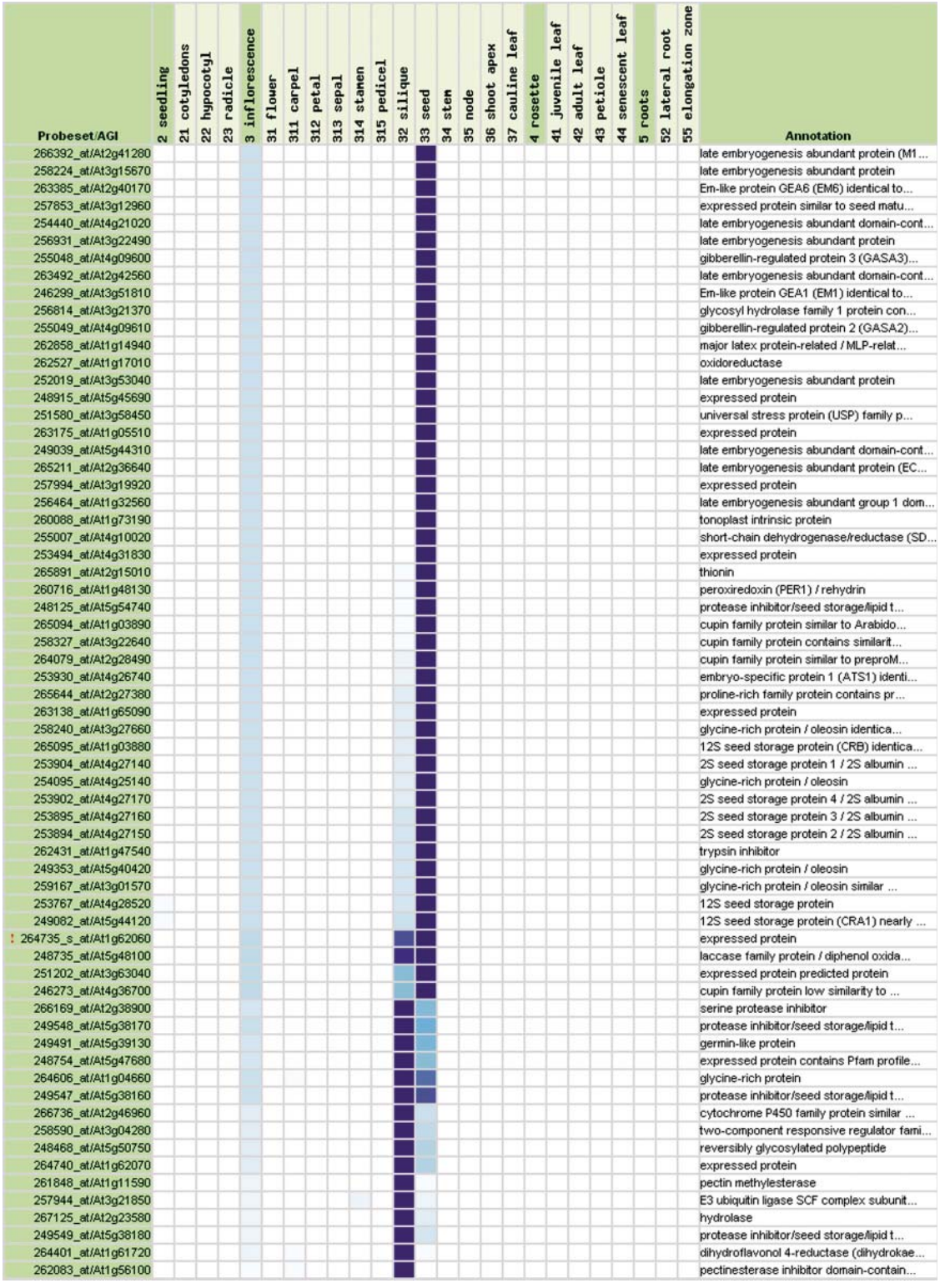
Third, the Response Viewer tool was used for several genes known to be responsive to particular stresses (Fig. 3, L–Q). GENEVESTIGATOR correctly showed the expression pattern of a light-induced gene encoding a light-harvesting chlorophyll *a*/*b* binding protein (AT4G14690, Jansson et al., 2000) and 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*

Probeset/AGI	2 seedling	21 cotyledons	22 hypocotyl	23 radicle	3 inflorescence	31 flower	311 carpel	312 petal	313 sepal	314 stamen	315 pedicel	32 silique	33 seed	34 sten	35 node	36 shoot apex	37 cauline leaf	4 rosette	41 juvenile leaf	42 adult leaf	43 petiole	44 senescent leaf	5 roots	52 lateral root	55 elongation zone	Annotation
261943_at/At1g80660										■																ATPase 9
248227_at/At5g53820										■																expressed protein similar to ABA-induci...
256587_at/At3g28780										■																glycine-rich protein similar to H41 gen...
249048_at/At5g44300										■																dormancy/auxin associated family protei...
265080_at/At1g55570										■																multi-copper oxidase type I family prot...
256588_at/At3g28790										■																expressed protein
254716_at/At4g13560										■																late embryogenesis abundant domain-cont...
252161_at/At3g50580										■																proline-rich family protein contains pr...
247897_at/At5g57810										■																senescence-associated protein-related s...
263144_at/At1g54070										■																dormancy/auxin associated protein-relat...
248194_at/At5g54095										■																expressed protein
! 252085_s_at/At3g52000										■																serine carboxypeptidase S10 family prot...
261045_at/At1g01310										■																allergen V5/Tpx-1-related family protei...
247402_at/At5g62750										■																expressed protein predicted proteins
267476_at/At2g02720										■																pectate lyase family protein similar to...
251180_at/At3g62640										■																expressed protein
246761_at/At5g27980										■																seed maturation family protein similar ...
267443_at/At2g19000										■																expressed protein
250608_at/At5g07420										■																pectinesterase family protein contains ...
262022_at/At1g35490										■																bZIP family transcription factor
258600_at/At3g02810										■																protein kinase family protein contains ...
247843_at/At5g58050										■																glycerophosphoryl diester phosphodieste...
261532_at/At1g71680										■																lysine and histidine specific transport...
254762_at/At4g13230										■																late embryogenesis abundant domain-cont...
265552_at/At2g07560										■																ATPase
255101_at/At4g08670										■																protease inhibitor/seed storage/lipid t...
266697_at/At2g19770										■																profilin 4 (PRO4) (PFN4) identical to p...
267447_at/At2g33870										■																Ras-related GTP-binding protein
253153_at/At4g35700										■																zinc finger (C2H2 type) family protein ...
264269_at/At1g60240										■																apical meristem formation protein-relat...
259044_at/At3g03430										■																polcalcin
252061_at/At3g52620										■																hypothetical protein phosphate actyltra...
250121_at/At5g16500										■																protein kinase family protein contains ...
246878_at/At5g26060										■																S1 self-incompatibility protein-related...
263659_at/At1g04470										■																expressed protein EST gb|ATTS5672 comes...
255815_at/At1g19890										■																histone H3
259265_at/At3g01250										■																expressed protein
249536_at/At5g38760										■																expressed protein similar to ABA-induci...
252771_at/At3g42880										■																leucine-rich repeat transmembrane prote...
247759_at/At5g59040										■																copper transporter family protein simil...
250902_at/At5g03590										■																GDSL-motif lipase/hydrolase protein-rel...
258671_at/At3g08560										■																vacuolar ATP synthase subunit E
262385_at/At1g72960										■																root hair defective 3 GTP-binding (RHD3...
265587_at/At2g19980										■																allergen V5/Tpx-1-related family protei...
258843_at/At3g04690										■																protein kinase family protein contains ...
249381_at/At5g40040										■																60S acidic ribosomal protein P2 (RPP2E)...
260702_at/At1g32250										■																calmodulin
256200_at/At1g58210										■																kinase interacting family protein simil...
265280_at/At2g28355										■																expressed protein contains similarity t...
255530_at/At4g02140										■																expressed protein
249150_at/At5g43340										■																inorganic phosphate transporter identic...
253961_at/At4g26440										■																WRKY family transcription factor identi...
250561_at/At5g08030										■																glycerophosphoryl diester phosphodieste...
255386_at/At4g03620										■																myosin heavy chain-related contains wea...
259451_at/At1g13890										■																SNAP25 homologous protein
249428_at/At5g39870										■																hypothetical protein
256261_at/At3g12160										■																Ras-related GTP-binding family protein ...
257065_at/At3g18220										■																phosphatidic acid phosphatase family pr...
245232_at/At4g25590										■																actin-depolymerizing factor
261015_at/At1g26480										■																14-3-3 protein GF14 iota (GRF12) identi...
248367_at/At5g52360										■																actin-depolymerizing factor
246646_at/At5g35090										■																expressed protein
265473_at/At2g15535										■																SLR1 binding pollen coat protein-relate...
264603_at/At1g04670										■																expressed protein
267449_at/At2g33690										■																late embryogenesis abundant protein
262156_at/At1g52680										■																late embryogenesis abundant protein-rel...
253151_at/At4g35670										■																glycoside hydrolase family 28 protein /...
246431_at/At5g17480										■																polcalcin
259693_at/At1g63060										■																expressed protein
254033_at/At4g25950										■																vacuolar ATP synthase
262742_at/At1g28550										■																Ras-related GTP-binding protein
260306_at/At1g70540										■																invertase/pectin methylesterase inhibit...

(*Table continues on following page.*)

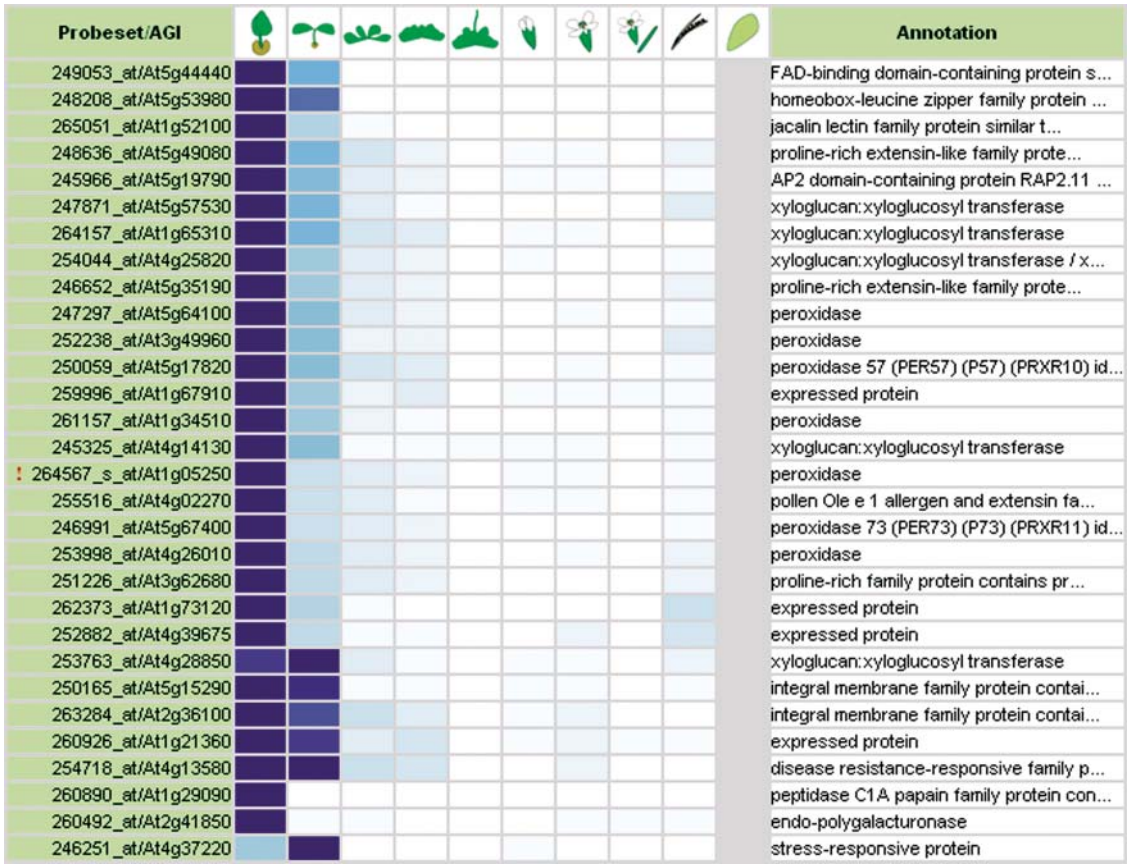
Table IIB.



Probeset/AGI	2 seedling	21 cotyledons	22 hypocotyl	23 radicle	3 inflorescence	31 flower	311 carpel	312 petal	313 sepal	314 stamen	315 pedicel	32 silique	33 seed	34 stem	35 node	36 shoot apex	37 cauline leaf	4 rosette	41 juvenile leaf	42 adult leaf	43 petiole	44 senescent leaf	5 roots	52 lateral root	55 elongation zone	Annotation
266392_at/At2g41280																										late embryogenesis abundant protein (M1...
258224_at/At3g15670																										late embryogenesis abundant protein
263385_at/At2g40170																										Em-like protein GEA6 (EM6) identical to...
257853_at/At3g12960																										expressed protein similar to seed matu...
254440_at/At4g21020																										late embryogenesis abundant domain-cont...
256931_at/At3g22490																										late embryogenesis abundant protein
255048_at/At4g09600																										gibberellin-regulated protein 3 (GASA3)...
263492_at/At2g42560																										late embryogenesis abundant domain-cont...
246299_at/At3g51810																										Em-like protein GEA1 (EM1) identical to...
256814_at/At3g21370																										glycosyl hydrolase family 1 protein con...
255049_at/At4g09610																										gibberellin-regulated protein 2 (GASA2)...
262858_at/At1g14940																										major latex protein-related / MLP-relat...
262527_at/At1g17010																										oxidoreductase
252019_at/At3g53040																										late embryogenesis abundant protein
248915_at/At5g45690																										expressed protein
251580_at/At3g58450																										universal stress protein (USP) family p...
263175_at/At1g05510																										expressed protein
249039_at/At5g44310																										late embryogenesis abundant domain-cont...
265211_at/At2g36640																										late embryogenesis abundant protein (EC...
257994_at/At3g19920																										expressed protein
256464_at/At1g32560																										late embryogenesis abundant group 1 dom...
260088_at/At1g73190																										tonoplast intrinsic protein
255007_at/At4g10020																										short-chain dehydrogenase/reductase (SD...
253494_at/At4g31830																										expressed protein
265891_at/At2g15010																										thionin
260716_at/At1g48130																										peroxiredoxin (PER1) / rehydrin
248125_at/At5g54740																										protease inhibitor/seed storage/lipid t...
265094_at/At1g03890																										cupin family protein similar to Arabido...
258327_at/At3g22640																										cupin family protein contains similarit...
264079_at/At2g28490																										cupin family protein similar to preproM...
253930_at/At4g26740																										embryo-specific protein 1 (ATS1) identi...
265644_at/At2g27380																										proline-rich family protein contains pr...
263138_at/At1g65090																										expressed protein
258240_at/At3g27660																										glycine-rich protein / oleosin identica...
265095_at/At1g03880																										12S seed storage protein (CRB) identica...
253904_at/At4g27140																										2S seed storage protein 1 / 2S albumin ...
254095_at/At4g25140																										glycine-rich protein / oleosin
253902_at/At4g27170																										2S seed storage protein 4 / 2S albumin ...
253895_at/At4g27160																										2S seed storage protein 3 / 2S albumin ...
253894_at/At4g27150																										2S seed storage protein 2 / 2S albumin ...
262431_at/At1g47540																										trypsin inhibitor
249353_at/At5g40420																										glycine-rich protein / oleosin
259167_at/At3g01570																										glycine-rich protein / oleosin similar ...
253767_at/At4g28520																										12S seed storage protein
249082_at/At5g44120																										12S seed storage protein (CRA1) nearly ...
! 264735_s_at/At1g62060																										expressed protein
248735_at/At5g48100																										laccase family protein / diphenol oxida...
251202_at/At3g63040																										expressed protein predicted protein
246273_at/At4g36700																										cupin family protein low similarity to ...
266169_at/At2g38900																										serine protease inhibitor
249548_at/At5g38170																										protease inhibitor/seed storage/lipid t...
249491_at/At5g39130																										germin-like protein
248754_at/At5g47680																										expressed protein contains Pfam profile...
264606_at/At1g04660																										glycine-rich protein
249547_at/At5g38160																										protease inhibitor/seed storage/lipid t...
266736_at/At2g46960																										cytochrome P450 family protein similar ...
258590_at/At3g04280																										two-component responsive regulator fami...
248468_at/At5g50750																										reversibly glycosylated polypeptide
264740_at/At1g62070																										expressed protein
261848_at/At1g11590																										pectin methylesterase
257944_at/At3g21850																										E3 ubiquitin ligase SCF complex subunit...
267125_at/At2g23580																										hydrolase
249549_at/At5g38180																										protease inhibitor/seed storage/lipid t...
264401_at/At1g61720																										dihydroflavonol 4-reductase (dihydrokae...
262083_at/At1g56100																										pectinesterase inhibitor domain-contain...

(*Table continues on following page.*)

Table IIC.



Probeset/AGI	Annotation
249053_at/At5g44440	FAD-binding domain-containing protein s...
248208_at/At5g53980	homeobox-leucine zipper family protein ...
265051_at/At1g52100	jacalin lectin family protein similar t...
248636_at/At5g49080	proline-rich extensin-like family prote...
245966_at/At5g19790	AP2 domain-containing protein RAP2.11 ...
247871_at/At5g57530	xyloglucan:xyloglucosyl transferase
264157_at/At1g65310	xyloglucan:xyloglucosyl transferase
254044_at/At4g25820	xyloglucan:xyloglucosyl transferase / x...
246652_at/At5g35190	proline-rich extensin-like family prote...
247297_at/At5g64100	peroxidase
252238_at/At3g49960	peroxidase
250059_at/At5g17820	peroxidase 57 (PER57) (P57) (PRXR10) id...
259996_at/At1g67910	expressed protein
261157_at/At1g34510	peroxidase
245325_at/At4g14130	xyloglucan:xyloglucosyl transferase
! 264567_s_at/At1g05250	peroxidase
255516_at/At4g02270	pollen Ole e 1 allergen and extensin fa...
246991_at/At5g67400	peroxidase 73 (PER73) (P73) (PRXR11) id...
253998_at/At4g26010	peroxidase
251226_at/At3g62680	proline-rich family protein contains pr...
262373_at/At1g73120	expressed protein
252882_at/At4g39675	expressed protein
253763_at/At4g28850	xyloglucan:xyloglucosyl transferase
250165_at/At5g15290	integral membrane family protein contai...
263284_at/At2g36100	integral membrane family protein contai...
260926_at/At1g21360	expressed protein
254718_at/At4g13580	disease resistance-responsive family p...
260890_at/At1g29090	peptidase C1A papain family protein con...
260492_at/At2g41850	endo-polygalacturonase
246251_at/At4g37220	stress-responsive protein

(*Table continues on following page.*)

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

Table IID.

Probeset/AGI	Annotation
262697_at/At1g75940	glycosyl hydrolase family 1 protein / a...
257220_at/At3g27810	myb family transcription factor (MYB3) ...
256381_at/At1g66850	protease inhibitor/seed storage/lipid t...
260038_at/At1g68875	expressed protein
262675_at/At1g75930	family II extracellular lipase 6 (EXL6)...
245622_at/At4g14080	glycosyl hydrolase family 17 protein / ...
264430_at/At1g61680	terpene synthase/cyclase family protein...
255101_at/At4g08670	protease inhibitor/seed storage/lipid t...
249048_at/At5g44300	dormancy/auxin associated family protei...
261532_at/At1g71680	lysine and histidine specific transport...
259265_at/At3g01250	expressed protein
249536_at/At5g38760	expressed protein similar to ABA-induci...
260306_at/At1g70540	invertase/pectin methylesterase inhibit...
245232_at/At4g25590	actin-depolymerizing factor
248367_at/At5g52360	actin-depolymerizing factor
262156_at/At1g52680	late embryogenesis abundant protein-rel...
261015_at/At1g26480	14-3-3 protein GF14 iota (GRF12) identi...
262742_at/At1g28550	Ras-related GTP-binding protein
267449_at/At2g33690	late embryogenesis abundant protein
265473_at/At2g15535	SLR1 binding pollen coat protein-relate...
264603_at/At1g04670	expressed protein
253151_at/At4g35670	glycoside hydrolase family 28 protein /...
261943_at/At1g80660	ATPase 9
248227_at/At5g53820	expressed protein similar to ABA-induci...
249150_at/At5g43340	inorganic phosphate transporter identic...
265552_at/At2g07560	ATPase
246431_at/At5g17480	polcalcin
259693_at/At1g63060	expressed protein
254033_at/At4g25950	vacuolar ATP synthase
256588_at/At3g28790	expressed protein
256582_at/At3g28840	expressed protein
256587_at/At3g28780	glycine-rich protein similar to H41 gen...
251988_at/At3g53300	cytochrome P450 family protein CYTOCHR...

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.

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

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

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

Albani D, Sardana R, Robert LS, Altosaar I, Arnison PG, Fabijanski SF (1992) A Brassica napus gene family which shows sequence similarity to ascorbate oxidase is expressed in developing pollen. Molecular characterization and analysis of promoter activity in transgenic tobacco plants. Plant J **2:** 331–342

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25–29

Becker JD, Boavida LC, Carneiro J, Haury M, Feijo JA (2003) Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol **133:** 713–725

Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol **2:** E9

Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN (2003) A gene expression map of the Arabidopsis root. Science **302:** 1956–1960

Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Gorlach J (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell **13:** 1499–1510

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton H C, et al (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet **29:** 365–371

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res **31:** 68–71

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

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

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science **303:** 799–805

Hennig L, Gruissem W, Grossniklaus U, Köhler C (2004) Transcriptional programs of early stages of plant reproduction. Plant Physiol **135:** 1765–1775

Hennig L, Menges M, Murray JA, Gruissem W (2003) Arabidopsis transcript profiling on Affymetrix GeneChip arrays. Plant Mol Biol **53:** 457–465

Honys D, Twell D (2003) Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiol **132:** 640–652

Ihmels J, Levy R, Barkai N (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat Biotechnol **22:** 86–92

Jansson S, Andersson J, Kim SJ, Jackowski G (2000) An Arabidopsis thaliana protein homologous to cyanobacterial high-light-inducible proteins. Plant Mol Biol **42:** 345–351

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjolander K, Gruissem W, Baginsky S (2004) The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol **14:** 354–362

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298:** 799–804

Lehman A, Black R, Ecker JR (1996) HOOKLESS1, an ethylene response gene, is required for differential cell elongation in the Arabidopsis hypocotyl. Cell **85:** 183–194

Lelandais G, Le Crom S, Devaux F, Vialette S, Church GM, Jacq C, Marc P (2004) yMGV: a cross-species expression data mining tool. Nucleic Acids Res (Database issue) **32:** D323–D325

Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP (2002) Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics **18:** 1593–1599

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol **14:** 1675–1680

Menges M, Hennig L, Gruissem W, Murray JA (2003) Genome-wide gene expression in an Arabidopsis cell suspension. Plant Mol Biol **53:** 423–442

Mouline K, Very AA, Gaymard F, Boucherez J, Pilot G, Devic M, Bouchez D, Thibaud JB, Sentenac H (2002) Pollen tube development and competitive ability are impaired by disruption of a Shaker K(+) channel in Arabidopsis. Genes Dev **16:** 339–350

Onate-Sanchez L, Singh KB (2002) Identification of Arabidopsis ethylene-responsive element binding factors with distinct induction kinetics after pathogen infection. Plant Physiol **128:** 1313–1322

Parenicova L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, et al (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell **15:** 1538–1551

Pelaz S, Gustafson-Brown C, Kohalmi SE, Crosby WL, Yanofsky MF (2001) APETALA1 and SEPALLATA3 interact to promote flower development. Plant J **26:** 385–394

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science **297:** 1551–1555

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

Ruiz-Garcia L, Madueno F, Wilkinson M, Haughn G, Salinas J, Martinez-Zapater JM (1997) Different roles of flowering-time genes in the activation of floral initiation genes in Arabidopsis. Plant Cell **9:** 1921–1934

Runge S, Sperling U, Frick G, Apel K, Armstrong GA (1996) Distinct roles for light-dependent NADPH:protochlorophyllide oxidoreductases (POR) A and B during greening in higher plants. Plant J **9:** 513–523

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet **34:** 166–176

Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature **420:** 190–193

Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA **100:** 8348–8353

Vicient CM, Hull G, Guilleminot J, Devic M, Delseny M (2000) Differential expression of the Arabidopsis genes coding for Em-like proteins. J Exp Bot **51:** 1211–1220

Wille A, Zimmermann P, Vranová E, Bleuler S, Fürholz A, Hennig L, Laule O, Prelíc A, von Rohr P, Thiele L, et al (2004) Sparse graphical gaussian modeling for genetic regulatory network inference. Genome Biol (in press)