

*Chapter 9*

## ONLINE TOOLS FOR PRESENTATION AND ANALYSIS OF PLANT MICROARRAY DATA

***David Honys<sup>1,2,\*</sup>, Nikoleta Dupl'áková<sup>1,2</sup> and David Reňák<sup>1,2,3</sup>***

<sup>1</sup>Laboratory of Pollen Biology, Institute of Experimental Botany ASCR, v.v.i.,  
Rozvojová 263, 165 02 Prague 6, Czech Republic

<sup>2</sup>Department of Plant Physiology, Faculty of Science, Charles University,  
Viničná 5, 128 44, Prague 2, Czech Republic

<sup>3</sup>Department of Plant Physiology and Anatomy, Faculty of Science, University of South  
Bohemia, Branišovská 31, 370 05 České Budějovice, Czech Republic

### ABSTRACT

The complete sequencing and annotation of the *Arabidopsis thaliana* genome represented a major step in biological research. This knowledge enabled gene prediction, assignment of functional categories and gave an opportunity to characterize global gene expression patterns at the transcriptome level at different developmental stages and under various physiological and stress conditions. For the discovery of partial or whole regulatory or functional networks, the development of high-throughput technologies was inevitable with genome-wide transcriptomic studies providing an essential input. DNA microarray technologies are among the most frequently used methods for parallel global analysis of gene expression. Microarray technologies now belong to the standard functional genomics toolbox and have undergone massive development leading to increased genome coverage, accuracy and reliability. Whole genome microarrays developed by Affymetrix in collaboration with Syngenta represented the first standard in genome-wide transcriptomic studies in plants. Whole genome Affymetrix ATH1 GeneChips covers about 76% of the *Arabidopsis thaliana* genome. Moreover, the introduction of the Minimum Information About Microarray experiments (MIAME) has increased the value and reproducibility of microarray experiments and has become a standard in documentation of array experiments and in the creation of databases of comparable transcriptomic experiments. The number of experiments exploiting microarray technology has markedly increased in recent years. Not surprisingly, there are potential difficulties in the orientation in available data sets. Microarray expression data are deposited on servers, many of which are publicly accessible. Currently, these databases store several thousands of individual datasets and some of these offer online tools for data normalization, filtering,

---

\* Author for correspondence

statistical testing and pattern discovery. In parallel with the rapid accumulation of transcriptomic data, on-line analysis tools are being introduced to simplify their use. Global statistical data analysis methods contribute to the development of overall concepts about gene expression patterns and to query and compose working hypotheses. More recently, these applications are being supplemented with more specialized products offering visualization and specific data mining tools. In this chapter, an overview of available on-line databases and web-based applications using microarray data is presented together with the information about their structure and elementary principles.

## INTRODUCTION

Microarray technologies now belong to the standard functional genomics toolbox and have undergone massive development leading to increased genome coverage, accuracy and reliability. The number of experiments exploiting microarray technology has markedly increased in recent years. In parallel with the rapid accumulation of transcriptomic data, on-line analysis tools are being introduced to simplify their use. Global statistical data analysis methods contribute to the development of overall concepts about gene expression patterns and to query and compose working hypotheses. More recently, these applications are being supplemented with more specialized products offering visualization and specific data mining tools.

The knowledge of whole sequenced and annotated plant genomes that started with *Arabidopsis thaliana* [AGI 2000] enabled gene prediction, assignment of functional categories and gave an opportunity to study gene and chromosome organization including the distribution of transposable elements. Finally it has enabled the characterization of global gene expression patterns at the transcriptome level at different developmental stages and under various physiological and stress conditions. Efforts to reveal the biological functions of thousand of genes and their integration into proteome, metabolome and interactome networks has become the principal focus of many studies and represents a key objective of the 2010 Project (<http://www.nsf.gov/pubs/2006/nsf06612/nsf06612.htm>).

A number of efficient and accurate gene expression analysis technologies to determine the expression levels of individual genes have been widely exploited in recent decades (Northern blot analysis, quantitative reverse transcription-PCR, cDNA library screening). Most of these methods enable analysis of the expression of single or relatively few selected genes. For the discovery of partial or whole gene functional or regulatory networks, the development of high-throughput technologies is essential with genome-wide transcriptomic studies providing a major input [Donson et al. 2002]. Several such methods have been developed including, cDNA fingerprinting [Money et al. 1996], serial analysis of gene expression - SAGE [Velculescu et al. 1995], massively parallel signature sequencing - MPSS [Brenner et al. 2000], high-density DNA oligonucleotide probe microarrays [Lockhart et al. 1996, Lipshutz et al. 1999] or cDNA arrays [Schena et al. 1995]. DNA microarray technologies are among the most frequently used methods for parallel global analysis of gene expression. These methods are based on the principle of selective and differential hybridization between sample target molecules and immobilized DNA probes. Hybridisation to probes arrayed on a solid surface report the relative abundance of DNA or RNA target molecules by fluorescent signal detection [van Hall et al. 2000, Clarke and Zhu 2006]. Microarray technologies now belong to the standard functional genomics toolbox [Zhu 2003,

Aharoni and Vorst 2002] and have undergone massive development leading to increased genome coverage, accuracy and reliability. Whole Genome microarrays developed by Affymetrix (Santa Clara, CA, USA) in collaboration with Syngenta represented the first standard in genome wide transcriptomic studies in plants. Whole genome Affymetrix ATH1 GeneChips cover about 76% of the *Arabidopsis thaliana* genes [Hennig et al. 2003]. Moreover, the introduction of the Minimum Information About Microarray experiments (MIAME) as standard documentation for array experiments and in transcriptomic databases, increasing the value and comparability of microarray data [Brazma et al. 2001].

As microarray technology has become standard technique in large scale gene expression studies, the number of transcriptomic experiments has markedly increased in recent years. Not surprisingly, there are potential difficulties in navigating between different available data sets. Microarray expression data are deposited on servers, many of which are publicly accessible. Public plant microarray data are deposited in several databases including ArrayExpress, GEO, NASCArrays and the Stanford Microarray Database. Currently these databases store several thousands of individual datasets and some of these offer on-line tools for data normalization, filtering, statistical testing and pattern discovery [Parkinson et al. 2005, Barrett et al. 2005, Edgar et al. 2002, Gollub et al. 2003, Craigon et al. 2004]. In parallel with the rapid accumulation of transcriptomic data, on-line analysis tools are being introduced to simplify their use. Global statistical data analysis methods contribute to the identification of overall gene expression patterns and to query and compose new working hypotheses based on these findings [Clarke and Zhu 2006, Zhu 2003, Hughes et al. 2000, Mandaokar et al. 2006]. More recently, these applications are being supplemented with more specialized products offering visualization and specific data mining tools. Geneinvestigator, Botany Array Resource, Arabidopsis co-expression tool, and ArrayExpress Expression Profiler offer Web-based tools to analyse large microarray datasets.

In this chapter we describe and collate several available on-line portals devoted to storage and analyses of plant microarray data. Summary of individual portals is provided in Table 1. For each database, basic information is available like URL, annotation of Arabidopsis genome, microarray data resource, used platforms, data normalisation, number of microarray datasets covered and access options. It is also distinguished between data repositories and/or data analysis tools. Moreover, gene- or genome-centric database concepts are compared. Gene-centric approach means that the particular gene is in the centre of interest and relevant data can be acquired from individual experiments. On the contrary, genome-centric approach enables the identification of genes fulfilling given criteria. Most portals enable direct data download. However, they offer different formats that are listed in Table 1. In the following section we provide more detailed description of all portals. We always shortly introduce the overall database concept and general features with special attention paid to specific tools for data analysis.

**Table 1. Overview of plant microarray databases**

<b>Portal</b>	<b>ACT</b>	<b>aGFP</b>	<b>ArrayExpress</b>	<b>BAR</b>	<b>GEO</b>	<b>Genevestigator V3</b>	<b>NASCArrays</b>	<b>SMD</b>
<b>Database type</b>	Data repository and simple toolbox	Data visualisation toolbox	Data repository and toolbox	Data analysis and visualisation toolbox	Data repository	Data repository and toolbox	Data repository and simple toolbox	Data repository and toolbox
<b>Concept</b>	Genome-centric	Gene-centric	Gene-centric and genome-centric	Gene-centric and genome-centric	Gene-centric and genome-centric	Gene-centric and genome-centric	Gene-centric	Gene-centric and genome-centric
<b>Arabidopsis genome annotation</b>	TAIR7	TAIR7	Various	TAIR7	n/a	TAIR7	TIGR5	n/a
<b>Microarray data resource</b>	NASCArrays/GAR Net	AtGenExpress, NASCArrays	EBI Array express or user-uploaded data	Bio-Array Resource Database, AtGenExpress, NASCArrays	User-uploaded data compatible with GEO structure	NASCArrays, FGCZ, GEO, ArrayExpress, AtGenExpress, TAIR	NASCArrays	Various including Affymetrix, Agilent, and spotted arrays
<b>Platform(s)</b>	Affymetrix ATH1 Whole Genome Array	Affymetrix ATH1 Whole Genome Array	Various	Affymetrix ATH1 Whole Genome Array	Various high-throughput gene expression data	Affymetrix AG and ATH1 Arrays	Affymetrix AG and ATH1 Arrays	Various
<b>Data normalisation</b>	MAS5.0	MAS4.0 and MAS5.0	Various	MAS5.0	Various	MAS5.0	MAS5.0	Various
<b>No. of experiments</b>	73 experiments, 544 arrays (Arabidopsis)	120 experiments and 350 arrays	3,051 experiments and over 50,000 arrays from over 200 organisms	BAR 150 arrays, NASCArrays incl. AtGenExpress 330/2954 (Arabidopsis)	4,134 platforms, 181,922 samples and 7,202 series from over 100 organisms	201 experiments and 3132 arrays (Arabidopsis)	330 experiments and 2954 arrays	over 70,000 arrays and 15,451 experiments from 53 organisms

**Table 1. Continued**

<b>Portal</b>	<b>ACT</b>	<b>aGFP</b>	<b>ArrayExpress</b>	<b>BAR</b>	<b>GEO</b>	<b>Genevestigator V3</b>	<b>NASCArrays</b>	<b>SMD</b>
<b>Access</b>	Free	Free	Free with optional password-protected account	Free	Free	Open (free), Classic (free for academic users) and Advanced (paid)	Free	Free with restricted access to non-public data
<b>Data download</b>	Yes, via NASCArrays	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<b>Downloaded data format</b>	Tab delimited, XLS files, Gnumeric, CSV files	TAB-delimited text	TXT or XLS format, MEGA-ML format	Depends on selected tool	Web-based interactive forms, XLS files, MINiML, SOFT text files	Depends on selected tool	CSV or TAB-delimited files	XLS and TXT files
<b>URL</b>	<a href="http://www.arabidopsis.leeds.ac.uk/act/index.php">http://www.arabidopsis.leeds.ac.uk/act/index.php</a>	<a href="http://aGFP.ueb.cas.cz">http://aGFP.ueb.cas.cz</a>	<a href="http://www.ebi.ac.uk/arrayexpress">http://www.ebi.ac.uk/arrayexpress</a>	<a href="http://bbc.botany.utoronto.ca">http://bbc.botany.utoronto.ca</a>	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>	<a href="https://www.genevestigator.ethz.ch/">https://www.genevestigator.ethz.ch/</a>	<a href="http://affy.arabidopsis.info/narrays/experimentbrowse.pl">http://affy.arabidopsis.info/narrays/experimentbrowse.pl</a>	<a href="http://genome-www5.stanford.edu/">http://genome-www5.stanford.edu/</a>
<b>Key references</b>	<i>Jen et al. 2006, Manfield et al. 2006</i>	<i>Duplakova et al. 2007</i>	<i>Torrente et al. 2005, Parkinson et al. 2006</i>	<i>Toufighi et al. 2005, Winter et al. 2007</i>	<i>Edgar et al. 2002, Barrett et al. 2007</i>	<i>Zimmermann et al. 2004, Laule et al. 2006,</i>	<i>Craigon et al. 2004</i>	<i>Ball et al. 2005, Demeter et al. 2007</i>

## PORTALS

### Arabidopsis Co-expression Tool (ACT)

ACT is a web-based tool for gene expression analysis using large microarray data set from the Nottingham Arabidopsis Stock Centre. The database stores pre-calculated co-expression results for cca 21 800 genes based on data from over 400 arrays (Figure 1). ACT enables identification of gene co-expression patterns across user-selected single or multiple arrays. An additional Clique Finder tool provides quantitative method for the determination of correlation cut-offs leading to the generation of groups of genes that may share a common purpose [Manfield et al. 2006]. For further analysis of co-regulated genes, promoter element detection software is included for the identification of potential DNA upstream elements [Jen et al. 2006]. The bioinformatic analyses consist of three steps: (i) generation of a list of

**Arabidopsis Coexpression Data Mining Tools** UNIVERSITY OF LEEDS

Home ACT Utilities Experiments Downloads Miscellaneous

Tools to calculate how similar the expression patterns of probes on the Affymetrix array follow that of a "driver" gene of interest.

Analysis options:

1. [Coexpression analysis in the specified experiment](#)
2. [Coexpression analysis over available array experiments](#)
3. [Co-correlation scatter plot](#)
4. [Clique Finder](#)
5. [Word Counter](#)
6. [GO Term Counter](#)
7. [Keyword Search](#)

Need help? See [FAQ](#)

---

**1. Coexpression analysis in the specified experiment**

Note: Javascript must be enabled. The page may take some time to load.

Select the array type	AtGenome1 (8k) ▼
Gene ID / Probe ID	<input type="text"/> Choose
Choose array experiment(s)	Select experiment(s)
Reset	

**2. Coexpression analysis over available array experiments.**

Show Pearson Correlation Coefficients for a probe using 121 AtGenome1 (Exp\_ID: 1\_1 to 1\_21) or 322 ATH1 arrays (Exp\_ID: 2\_1 to 2\_41).

Probe ID	<input type="text"/> (e.g. 254831_at)	Look up the probe ID by using <a href="#">ID exchanger</a> .
limited to the first 50 genes in descending (positively correlated genes) order.		
Leave box blank to receive the full correlation list.		
Reset Submit		

Example analysis:

1. Ribosomal protein [At4g12600 \(254831\\_at\)](#);
2. Heat shock protein [At2g20560 \(263374\\_at\)](#);
3. Chlorophyll A/B binding protein [At3g61470 \(251325\\_s\\_at\)](#);

Figure 1. Homepage of the Arabidopsis Co-expression Tool available at <http://www.arabidopsis.leeds.ac.uk/act/index.php>.

negatively/positively co-expressed genes for a given gene of interest (driver); (ii) calculation of a co-correlation scatter plot of two selected genes; and (iii) identification of groups of genes that share statistical significant co-expression patterns. Co-correlation scatter plot analyses showing positively and negatively correlated genes may identify similar transcription regulatory mechanisms and can help to reveal the related biological functions of unannotated genes. Moreover, groups of genes with significant co-expression patterns often share the same biological theme and their members may overlap with other groups. It may then suggest their involvement in different biological processes. ACT provides a genome-centric approach allowing generation of a list of genes co-expressed within the driver gene throughout all experiments available. As an output, ACT offers a graphical view of gene expression patterns, list view of selected genes, co-correlation scatter plots; clusters of closely-associated probes and annotations of the most correlated genes to a given probe. Data can be downloaded as simple Tab-delimited TXT files, CSV, XLS files or in a Gnumeric format.

### Specific Tools

**Expression Pattern Displayer** allows an exploration of gene expression patterns in a specified experiment or over all arrays available in the database.

**Coexpression analysis** was designed for the generation of lists of genes co-expressed with the probe (driver gene) according to their r-, p- and E-values.

**Co-correlation scatter plot** results in 2-D visualization of co-correlated data within two probes.

**Clique Finder** identifies genesets that are consistently co-expressed with each other throughout all microarray data.

**Word count** highlights annotations of the most correlated genes to a probe.

**GO Terms** offers Gene Ontology terms of the most correlated probes.

**Sequence extractor** extracts the specific region of cis-upstream or downstream sequences of single or multiple genes.

**cis-Element Analyser** analyses of cis-elements in the given sequence while **cis-Element Locator** can be used for the localization of user-defined DNA motifs in promoters of single or multiple genes.

**ID/Function Linker** converts multiple Affymetrix probe IDs and AGI codes.

### Distinguishable Features

Even though ACT uses the simplest measuring algorithm using the Pearson correlation coefficient, it is suggested to be an effective tool for the calculation of statistical significance [Manfield et al. 2006]. Researchers looking for positively and negatively correlated genes may (not) consider similar transcription regulatory mechanisms involved in related biological processes when revealing possible functions of unannotated genes. Moreover, the result of overlapping genes in each group may lead to a conclusion of being involved in different biological processes. In addition, the 2-D co-correlation plots can help in identifying sets of genes acting in particular metabolic, signaling or developmental processes. Finally, ACT tools for promoter analysis can be used to identify motifs which are often seen to underpin the observed expression correlation pattern. Other tools for identifying over-represented (but not

necessarily characterized) sequence motifs such as the Inclusive Motif Sampler are also useful for analyzing promoters of correlated genes.

### Arabidopsis Gene Family Profiler (aGFP)

Currently, on-line analysis tools have been developed to simplify particular analyses of transcriptomic data from wide-ranging microarray analysis. Database Arabidopsis Gene Family Profiler (aGFP; <http://aGFP.ueb.cas.cz>) [Dupl'áková et al. 2007] is a web-based tool offering visualization of microarray expression data of single genes, pre-defined gene families or even custom gene sets (Figure 2). aGFP gives users the possibility to analyze current *Arabidopsis* developmental transcriptomic data starting with simple global queries that can be expanded and further refined to visualize comparative and highly selective gene expression profiles. One of new features of aGFP is the possibility to display gene expression of pre-defined gene families and to easily adjust their composition by addition or reduction their members.

**Arabidopsis**  
**Gene Family Profiler**

Laboratory of Pollen Biology  
Institute of Experimental Botany  
Rozvojová 135  
165 00 Prague 6 - Lysolaje  
Czech Republic

AGI number    AtGenExpress  
 NASCArrays  
 MAS5  
 MAS4

Family-oriented gene expression database

© 2005 Nikoleta Duplíšková, Patrik Hovanec  
David Reňák, Barbora Honyzová  
David Twell, David Honys

Figure 2. Homepage of the ArabidopsisGFP database available at <http://aGFP.ueb.cas.cz>.

aGFP database is based on a “gene-centric approach”. Authors adopted the general concept “from simple to complex”. In the first approximation, an arithmetical mean expression signal from multiple experiments is displayed. In subsequent steps the user can choose to display expression data for individual plant organs or tissues at particular growth stages. This is accompanied by the option of progressive replacement of arithmetical means



by individual expression values. So users have the option to choose different levels of visualization to suit their needs.

aGFP offers several data search options together with two other input parameters; the gene detection algorithm (MAS4.0 or MAS5.0) and the source of expression data (AtGenExpress or NASCArrays). At any stage of the query the user has the possibility to directly switch options between these pairs of parameters.

aGFP uses two large freely available transcriptomic databases for various tissues at different developmental stages of wild type plants *Arabidopsis thaliana* grown at physiological conditions. The first subset contains data obtained within a scope of the AtGenExpress project [Schmid et al. 2005], the second comprises all other datasets deposited at NASC and was labeled NASCArrays [Craigon et al. 2004]. Data in each subset are presented using several different graphical displays and, the user has an option to instantly switch between subsets in each environment. Only experiments using Affymetrix ATH1 whole genome arrays with at least two biological replicates were included.

The complete datasets used in the aGFP database are described in the Legend available from the homepage and it is possible to trace the origin of all datasets. Gene expression data can be downloaded for individual and selected gene sets as a TAB-delimited TXT files. This enables the direct import of downloaded data into spreadsheet editors such as Excel and database software like Microsoft Access and FileMaker.

Moreover, aGFP allows a direct comparison of the influence of the detection algorithm or data resource on expression profiles. The database presents data normalised using two different algorithms, empirical Affymetrix MAS 4.0 and statistical MAS 5.0 with the possibility to switch between them at any point.

Genes in arabidopsisGFP are organized into two hierarchical levels consisting of gene families and superfamilies. All data were assembled from various relevant resources, the majority from TAIR – Arabidopsis Gene Family Information (<http://www.arabidopsis.org/browse/genefamily/index.jsp>) and AGRIS (<http://arabidopsis.med.ohio-state.edu/>) [Palaniswamy et al. 2006].

An interactive "virtual plant" represents the main tool for the visualization of the spatial and developmental gene expression profiles. Virtual plant uses a white-yellow-green scale to show relative expression signals of individual genes, or gene families throughout the *Arabidopsis* life cycle. In first step, the „virtual plant“ is displayed an mean expression signal from multiple experiments. In subsequent steps, the user can choose to display expression data for individual plant organs or tissues at particular growth stages. This is accompanied by the option of progressive replacement of arithmetical means by individual expression values. Finally, the user can switch from "virtual plant" visualization to a simple bar chart (standard or log-scaled) or tabulated display. The user can also browse through individual experiments down to normalized or even raw data extracted from individual gene chips. Gene family expression data can also be visualized as a colorized spot chart (heat map) with colour scale identical to that of the virtual plant. The interactivity of modeling is based on the possibility of ad hoc addition and removal of genes to and from currently active set. Moreover, each spot contains mouse-over-activated information about expression signal value and experiment. Annotation of individual experiments meets MIAME standard [Brazma et al. 2001].

## Distinguishable Features

A novel feature of aGFP is that it enables the evaluation of the impact of normalization procedures on microarray gene expression data as well the possibility of rapid definition of user-defined gene families or other groups. Simultaneously, aGFP serves as a facile and synoptic developmental reference guide for expression profiles of individual genes or gene families in wild-type *Arabidopsis thaliana* plants. arabidopsisGFP contains lists of pre-defined gene families and superfamilies enabling the rapid comparative visualization of expression profiles of their members.

## ArrayExpress

ArrayExpress (Figure 3) was developed as a public repository for gene expression data at the centre of a wider microarray informatics system at European Bioinformatics Institute (EBI) [Brazma et al. 2003]. It comprises well-annotated raw and processed microarray data originated from several platforms supporting MIAME requirements [Parkinson et al. 2005]. Currently, the database stores data of 1,500,000 gene expression profiles from more than 50 000 hybridizations representing more than 200 organisms [Parkinson et al. 2007]. Microarray data are either part of EBI ArrayExpress or they can be user-uploaded from various resources including Affymetrix, Agilent, Illumina, Nimblegen, etc. [Brazma et al. 2003]. The majority of experimental data relates to gene expression profiling studies, the remainder are array-based chromatin immunoprecipitation or comparative genomics studies [Parkinson et al. 2005].

The screenshot shows the ArrayExpress homepage with the following sections:

- Navigation Bar:** EMBL-EBI logo, EB-eye Search, All Databases dropdown, Enter Text Here search box, Go button, Reset, Advanced Search, Give us feedback link.
- Menu:** Databases, Tools, EBI Groups, Training, Industry, About Us, Help, Site Index.
- ArrayExpress:** Public repository for transcriptomics and related data. Logo: AE.
- Experiments:** 3218 experiments available. Search for experiment, citation, sample and factor annotations. Includes a search box and a 'query' button. Links: Browse experiments, Advanced query interface, Submitter/reviewer login.
- Expression Profiles:** 267 experiments, 121891 genes available. Search for Gene(s) [e.g., NFKBIA], Species (dropdown: Arabidopsis thaliana), and Experiment or sample annotation [e.g., Leukemia]. Includes a search box and a 'query' button. Link: Query & browse interface.
- News:** 11/12/2007 - ArrayExpress database doubles in size to 100,000 hybridisations. ArrayExpress has doubled in size reaching the 100,000 hybridisation milestone...more.
- Links:**
  - ArrayExpress User Survey *new!*
  - How to link to ArrayExpress
  - How to submit data to ArrayExpress
  - ArrayExpress interfaces tutorial (pdf)
  - Documentation and online help
  - Downloads and statistics for ArrayExpress databases
  - FTP server for public data
  - Quality metrics for microarrays
  - ArrayExpress Scientific Advisory Board

Figure 3. Homepage of the ArrayExpress database available at <http://www.ebi.ac.uk/arrayexpress>.

ArrayExpress was created to achieve three major goals, first, to serve as an archive for microarray data, second, to provide access to genes expression profiles, and third, to facilitate the sharing of microarray design and experimental protocols [Sarkans et al. 2005; Brazma et al. 2006].

ArrayExpress infrastructure consists of database itself and individual modules enabling user data submission, database query and data analyses [Brazma et al. 2003]. For data submission, three types of data are accepted, arrays, experiments and protocols. Altogether, they provide full information about the respective dataset but they can be uploaded separately under individual accession numbers. It simplifies the process of microarray data submission using standard platforms like Affymetrix. Data are submitted directly using three major submission routes: (i) online via the MIAMExpress data submission tool [Brazma et al. 2003], (ii) via a spreadsheet suitable for all types and sizes based on technology type, species and experimental type, and (iii) via a MAGE-ML or MAGE-TAB (Microarray gene Expression Markup Language) [Spellman et al. 2002] pipelines set-up with external databases [Kapushensky et al. 2004]. As plant-based data represent about 25% of ArrayExpress datasets and vast majority of those based on Arabidopsis, specific open-source software application for the submission of Arabidopsis-based microarray data is available [Mukherjee et al. 2005]. MIAMExpress itself enables download of limited number of hybridisations with the requirement for filling out a web-form for each individual dataset. For this reason, BLoader software application was developed and implemented to generate the annotation information rapidly and submit the entire set in one batch [Schwager et al. 2005].

Since its introduction, ArrayExpress has undergone significant improvement as it has two major components. Original “ArrayExpress experiment repository” represents the main database comprising complete data while “ArrayExpress gene expression profile data warehouse” contains gene-indexed expression profiles from a curated subset of experiments from the repository [Brazma et al. 2006]. Both ArrayExpress and user-uploaded microarray data can be analysed and visualized exploiting integrated ExpressionProfiler toolbox that is described in detail below [Kapushesky et al. 2004].

## Specific Tools

### *General tools:*

**Browse Experiments** enables filters on species, array, experiment type, experimental factors, author, laboratory, publications, array design name and protocol type.

**Browse Expression Profiles** selects gene name or AGI, species, experiment or sample annotation.

**Similarity Search** is designed for arrangement of additional genes with similar expression profiles.

### *Expression Profiler — next generation:*

**Expression Profiler** (EP) [Kapushesky et al. 2004] represents microarray data analysis and visualization toolbox implemented in the ArrayExpress database [Brazma et al. 2003]. EP was originally designed as a web-based platform for microarray data and other functional genomics-related data analysis [Kapushesky et al. 2004]. After the implementation of new architecture enabling the modularization of the original design and de-centralization of database development, the toolbox was renamed to Expression profiler: next generation

(EP:NG). Individual tools are described below. EP:NG can be accessed from ArrayExpress page or directly at <http://www.ebi.ac.uk/expressionprofiler>.

*Individual EP:NG tools:*

User-defined **Data Upload** of microarray data following specific data-organization and data description matrices including optional various metadata.

**Data Selection** provides brief basic statistical overview of microarray data and user-defined data selection and sub-selection.

**Data Transformation** represents optional procedures preceding or following data subselection. They include K-nearest neighbour imputation [Troyanskaya et al. 2001] to fill in missing values, LOWESS normalization [Yang et al. 2002], data conversion in two-channel experiments.

**Similarity Search** extracts and visualizes of a group of co-expressed genes in relation to given one(s).

**Clustering** is application of hierarchical and partitioning-based K-groups clustering algorithms [Torrente et al. 2005].

**Clustering Comparison** is a tool for the comparison and matching of two independent K-groups clustering results.

**Signature Algorithm** identifies co-expressed subset of genes in user-submitted geneset.

**Ordination** offers several tools for multivariate analysis methods; principal component analysis and correspondence analysis.

**Between Group Analysis** allows multiple discriminant approach used with expression data matrices.

### **Distinguishable Features**


ArrayExpress database provides a free public repository of microarray gene expression data of various origins. In its modular architecture it links together gene expression data and number of data processing and analysis tools. Moreover, data submission in three separate steps resembling GEO enables future comparison and analyses of microarray data of different origin. Extra value was added by the subselection of specific gene expression data into growing curated ArrayExpress data warehouse. Data analysis and visualization tools are implemented in the integrated ExpressionProfiler package. This, together with unified data submission format, enables direct link to downloaded gene expression data, their efficient processing and analyses. Among most distinguishable features of EP:NG belong its gene clustering and statistical tools.

### **Bio-Array Resource (BAR)**








The Bio-Array Resource (BAR) is an extensive microarray data repository accompanied by number of very helpful tools designed for data visualization and detailed analyses (Figure 4) [Toufighi et al. 2005]. BAR is a completely web-based database, exploiting MySQL with interfaces implemented in Perl and C. The site was originally named Botany Array Resource as it was designed for Arabidopsis-based data. Recently, a toolbox for the visualisation of gene expression in mouse tissues have been recently added [Winter et al.

2007] so further development and extension of the coverage is likely to be expected. Currently, BAR database comprises microarray data collected from University of Toronto microarray facility as well as data loaded from NASCArrays and AtGenExpress Consortium. The database has been designed as MIAME-compliant [Brazma et al. 2001] and all provided data and tools are publicly available through a web interface. BAR offers large number of various data analysis tools. They include not only gene expression visualization and analysis tools but also set of assorted on-line tools especially useful for genetic mapping and data formatting, visualization and clean up. An overview and brief characterization of individual tools is provided in the specific section.

## The Bio-Array Resource for Arabidopsis Functional Genomics

Welcome to the Bio-Array Resource - the BAR! Below are user-friendly web-based tools for working with functional genomics and other data. Most are designed with the Arabidopsis researcher in mind, but a couple of them can be useful to the wider research community, e.g. [Mouse eFP Browser](#) or [BlastDigester](#). Click on the  icon to read the partner paper to a given tool. Mouse-over a given link for further information, and [click here](#) to restore the News panel.

### Expression Tools

- [Expression Angler](#) 
- [Sample Angler](#)
- [e-Northerns with Expression Browser](#) 
- [Arabidopsis eFP Browser, Cell eFP Browser](#) 
- [Mouse eFP Browser](#) 
- [Promomer](#) 
- [Arabidopsis Interactions Viewer](#) 
- [AGURF](#)
- [Classification SuperViewer](#) 
- [at to AGI converter](#)

### Molecular Markers

- [BlastDigester](#) 
- [Marker Tracker](#)
- [CapsID](#) 

### Other Genomic Tools

- [ClustaW with MView Output](#)
- [DataMetaFormatter](#)
- [Heatmapper](#)
- [Heatmapper Plus](#)
- [Duplicate Remover](#)
- [Venn Selector](#)
- [Venn SuperSelector](#)
- [Random ID list generator](#)

### Expression Browser

Perform electronic Northern blots using the Expression Browser, i.e. ask how up to 125 genes of interest are being expressed, with the gene expression data sets accumulated to date in the Botany Array Resource (BAR) DB or public data sets from the [AtGenExpress Consortium](#) ([Developmental](#) - see the Schmid et al. publication, [Abiotic Stress](#) - Kilian et al. 2007, [Pathogen](#), and [Hormone Series](#)), or others.

Automatic clustering can be performed. Our database contains more than 30 million expression measurements.

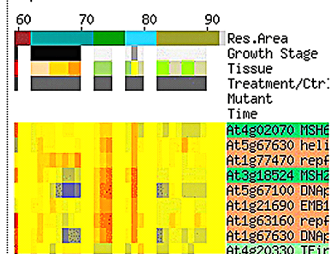


Figure 4. Homepage of the Bio-Array Resource database and toolbox available at <http://bbc.botany.utoronto.ca>.

## Specific Tools

### *Gene expression tools:*

**Expression Angler** identifies sets of co-regulated, anti-correlated, or condition/tissue-specific genes in various data sets.

**Sample Angler** identifies samples exhibiting similar expression profiles.

**e-Northerns with Expression Browser** is an electronic Northern tool providing a graphical display of expression data of particular set of genes across all selected datasets.

**Arabidopsis eFP Browser** provides intuitive visualization of particular gene expression patterns.

**Cell eFP Browser** enables intuitive visualization of particular protein subcellular localization patterns based on documented and predicted subcellular localizations according to the SUBA database [Heazlewood et al. 2005, 2007].

**Mouse eFP Browser** offers intuitive visualization of particular gene expression patterns based on the Zhang et al. [2004] Functional Landscape of Mouse Gene Expression.

**Promomer** was designed for the identification of over-represented motifs in the promoter of a gene of interest, or in promoters of co-regulated genes.

**Arabidopsis Interactions Viewer** returns predicted or documented interaction partners with a gene of interest.

**AGURR** (Arabidopsis Genetic Uniqueness and Redundancy Revealer) is a program for the identification of samples in which one gene is uniquely expressed or in which several genes are potentially redundantly expressed.

**Classification SuperViewer** returns functional classification a of AGI IDs list based on the MIPS database.

**at to AGI converter** converts 22k Affymetrix GeneChip IDs into AGI IDs and vice versa.

#### *Molecular Markers:*

**BlastDigester** analyses a nucleotide BLAST output for restriction enzymes differentially cutting the aligned sequences [Ilic et al. 2005].

**MarkerTracker** is a repository for genetic markers.

**CapsID** serves as automated CAPS marker-based genotyping system [Taylor and Provart 2006].

#### *Other general tools for basic genomic analyses:*

**ClustalW with MView Output** is a web-based version of ClustalW.

**DataMetaFormatter** can be used for Arabidopsis gene expression data reformatting and adding pieces of meta-information (protein-protein interactions, functional classification).

**Heatmapper and Heatmapper Plus** applies a third dimension of information to a 2-D table via colour-coding.

**Duplicate Remover** removes duplicates from data lists.

**Venn Selector** shows identifiers in common and unique to two sets of sequences and fetching annotation.

**Venn SuperSelector** organizes multiple lists of genes or terms with associated values with an option of data export to Excel.

**Random ID list generator** is useful for the generation of  $n$  sets of  $y$  genes containing  $z$  number of randomly generated Arabidopsis AGI IDs.

## **Distinguishable Features**

The BAR provides number of sophisticated and easy-to-use web-based tools for microarray data visualization and analyses. Moreover, these tools are accompanied by other “simple” but extremely helpful tools for data formatting, presentation and clean-up that certainly may be time-consuming. Microarray data are divided into several databases according to their origin and experiment type. In most tools, users can choose from Bio-Array

resource database, several AtGenExpress Consortium databases, NASC Arrays and sometimes even their own in-house databases.

## Gene Expression Omnibus (GEO)

The Gene Expression Omnibus (GEO) was originally designed as a public data repository and retrieval system for high throughput gene expression data with an ambition to complement number of existing in house gene expression databases and to act as the central data distribution hub (Figure 5) [Edgar et al. 2002]. It represents one of large number of computational resources for the analysis of data stored in GenBank and other biological data provided by NCBI and available through NCBI's website <http://www.ncbi.nlm.nih.gov> [Wheeler et al. 2005, 2007]. Currently, the database has a MIAME-compliant infrastructure and contains both raw and processed data derived from over 100 organisms submitted by over

NCBI

Gene Expression Omnibus

HOME SEARCH SITE MAP Handout NAR 2006 Paper NAR 2002 Paper FAQ MIAME Email GEO

NCBI > GEO Not logged in | Login

**Gene Expression Omnibus:** a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

**GEO navigation**

**QUERY**

- DataSets
- Gene profiles
- GEO accession
- GEO BLAST

**BROWSE**

- DataSets
- GEO accessions
  - Platforms
  - Samples
  - Series

**SUBMIT**

- Direct deposit / update
- Web deposit / update
- Create new account

**Public data**

GPL Platforms	4114
GSM Samples	180869
GSE Series	7166
<b>Total</b>	<b>192149</b>

**Site contents**

**Documentation**

- Overview
- FAQ
- Submission guide
- Linking & citing
- Journal citations
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

**Query & Browse**

- Repository browser
- Submitter contacts
- SAGEmap
- FTP site
- GEO Profiles
- GEO DataSets

**Deposit & Update**

- Direct deposit
- Web deposit
- New account

GEO help: Mouse over screen elements for information

Get GEO accession  Scope: **Self** Format: **HTML** Amount: **Quick**

**Depositors only** User:  Password:   Recover a password

NLM | NIH | GEO Help | Disclaimer | Section 508

Figure 5. Homepage of the GEO database available at <http://www.ncbi.nlm.nih.gov/geo/>.

1,500 laboratories [Barrett and Edgar 2006a, Barrett et al. 2007]. GEO is not devoted to one technology but it contains microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules, as well as non-array-based technologies such as SAGE and mass spectrometry peptide profiling [Barrett et al. 2005].

The unifying feature of GEO database is that all data are segregated and stored as three principle components, platforms, samples and series. Platform provides a summary description of the array and data table by defining the array template. Sample describes the biological material, the experimental conditions under which the sample was handled and a respective data table containing hybridization signal levels for each feature on the corresponding previously defined platform. Finally, series defines a set of related samples that are considered as a part of an individual experiment [Barrett et al. 2007].

### Specific Tools

There are two key options to query GEO using two NCBI Entrez databases [Barrett and Edgar 2006b]. **Entrez GEO DataSets** allows a genome-centric view on gene expression data in GEO. It returns a keyword-based identification of experiments of interest. **Entrez GEO Profiles** provides a gene-centric view on gene expression data in GEO. It returns expression profiles of genes of interest. Particular genes are identified using various search options. Both NCBI Entrez databases are supplemented with several additional tools. Alternatively, GEO database can be queried using GEO BLAST tool. It retrieves sequences and corresponding expression profiles similar to user-defined sequence.

*Entrez GEO DataSets supplementary tools:*

**DataSet clusters** is a hierarchical and K-means clustering tool for data visualization, selection and download.

**Query group A versus B** identifies genes with expression profiles meeting user-defined statistical criteria.

**Subset effect flags was** designed for the identification of potentially interesting genes by showing those with significantly different expression levels among experimental variables.

*Entrez GEO Profiles supplementary tools:*

**Profile neighbors link** allows connection and display of genes sharing similar expression profiles.

**Sequence neighbors link** enables connection and display of groups of genes related by sequence similarity.

**Homolog neighbors link** connects and displays groups of genes related by Homologene groups.

**Links link** links GEO data to related data in other NCBI resources (ie. PubMed, GenBank, Gene, UniGene, OMIM).

### Distinguishable Features

GEO is a fast-growing data repository representing large compendium of various gene expression data. Its main impact lies in the complexity of microarray data and in their integration with other resources like sequence information, mapping and bibliographic data.



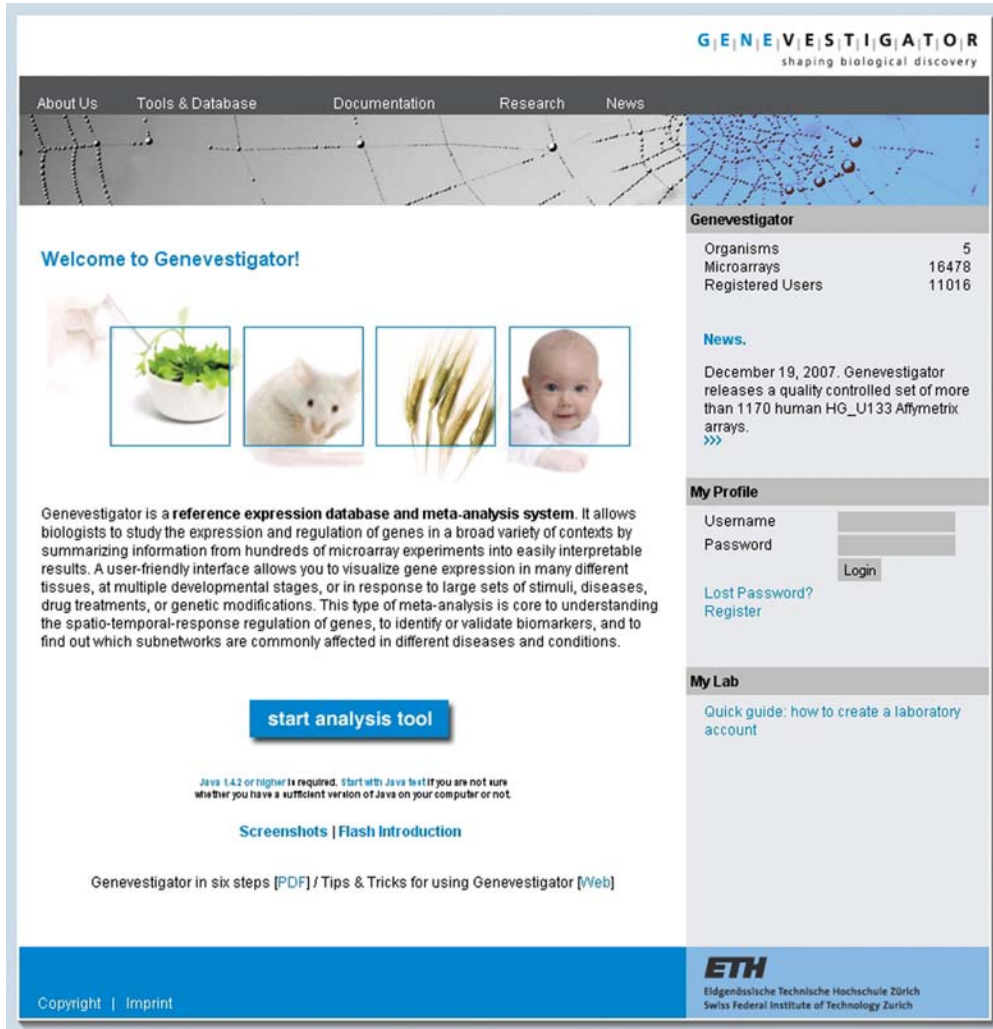
Moreover, GEO offers the possibility to download gene expression data to enable researchers the use of their own software. One example of such software package can be downloaded as a part of R/Bio/Conductor at <http://bioconductor.org/packages/release/bioc/html/GEOquery.html>. On the contrary, the diversity of area covered makes the development of robust and user-friendly statistical and data analysis tools difficult. Authors also plan further development of GEO and its extension in order to include non-gene-expression data types like chromatin-immunoprecipitation on arrays (ChIP-chip) studies, array comparative genomic hybridization (aCGH), SNP arrays and even proteomic data [Barrett et al. 2007].

### Genevestigator V3 (GV3)

In 2007, new version of Genevestigator (V3 beta) was released (Figure 6). It serves as a bioinformatic service for microarray gene expression data from five organisms (*Arabidopsis thaliana*, *Mus musculus*, *Rattus norvegicus*, *Hordeum vulgare*, *Homo sapiens*) with total number 16,319 microarrays [Zimmermann et al. 2004, 2005, Laule et al. 2006]. For *Arabidopsis*, Genevestigator V3 covers 3222 experiments. Initial Genevestigator database was launched in 2004 as a web-based MySQL/PHP software application to query *Arabidopsis* microarray database with several analysis tools – Digital Northern, Gene Correlator, The Gene Atlas, Gene Chronologer, Response Viewer and Meta-Analyser. Currently Genevestigator V3 operates as Java applet at a client's side. This client-server structure offers more sophisticated functions and allows more comfortable interface. Obviously, it is necessary to have Java runtime environment (JRE) version 1.4.2 or higher installed. The database is supplemented with several tools - Meta-Profile Analysis, Biomarker Search, Clustering Analysis, Pathway Projector. Genevestigator belongs among the most professional microarray analysis applications and provides many data mining tools allowing multilevel analyses. However, it does not serve as a data repository thus the user-controlled direct data upload is not possible. Currently, it hosts only expression data from Affymetrix GeneChips; for *Arabidopsis*, AG and ATH1 arrays are covered.

Transcriptomic data and their annotations were retrieved from many sources and repositories. Of these, NASC (<http://affymetrix.arabidopsis.info/>) [Craigon et al. 2004], FGCZ (<http://www.fgc.ethz.ch/>), GEO (<http://www.ncbi.nlm.nih.gov/geo/>), [Edgar et al. 2002, 2006, Barrett et al. 2005, 2007, Barrett and Edgar 2006a], ArrayExpress ([http://www.ebi.ac.uk/microarray-as/aer/?#ae-main\[0\]](http://www.ebi.ac.uk/microarray-as/aer/?#ae-main[0])) [Parkinson et al. 2005, 2007, Brazma et al. 2003, 2006, Sarkans et al. 2005, Rocca-Serra et al. 2003], AtGenExpress [Schmid et al. 2005] and TAIR (<http://www.arabidopsis.org/>) [Rhee et al. 2003] belong among the public repositories. Genevestigator enables two main query approaches – a gene-centric and genome-centric. It also offers an extensive selection of data-searching options that are well described in the accompanying documentation ([https://www.genevestigator.ethz.ch/index.php?option=com\\_content&task=view&id=34&Itemid=95](https://www.genevestigator.ethz.ch/index.php?option=com_content&task=view&id=34&Itemid=95)).

Genevestigator V3 is generally publicly available. However, it offers three levels of access – Open, Classic and Advanced. Open access does not require the registration of user account but it offers only two tools (Northern and Selection tool). Classic access enables also the Meta-Profile Analysis toolbox and is free for academic institutions. Advanced access offers all tools available in GV3 (also Biomarker search, Clustering analysis and Pathway projector) but this access option is paid.



**GENEVESTIGATOR**  
shaping biological discovery

About Us Tools & Database Documentation Research News

**Welcome to Genevestigator!**

Genevestigator is a **reference expression database and meta-analysis system**. It allows biologists to study the expression and regulation of genes in a broad variety of contexts by summarizing information from hundreds of microarray experiments into easily interpretable results. A user-friendly interface allows you to visualize gene expression in many different tissues, at multiple developmental stages, or in response to large sets of stimuli, diseases, drug treatments, or genetic modifications. This type of meta-analysis is core to understanding the spatio-temporal-response regulation of genes, to identify or validate biomarkers, and to find out which subnetworks are commonly affected in different diseases and conditions.

[start analysis tool](#)

Java 1.4.2 or higher is required. Start with Java test if you are not sure whether you have a sufficient version of Java on your computer or not.

[Screenshots | Flash Introduction](#)

[Genevestigator in six steps \[PDF\] / Tips & Tricks for using Genevestigator \[Web\]](#)

**Genevestigator**

Organisms	5
Microarrays	16478
Registered Users	11016

**News.**

December 19, 2007. Genevestigator releases a quality controlled set of more than 1170 human HG\_U133 Affymetrix arrays. >>>

**My Profile**

Username   
 Password

[Lost Password?](#)  
[Register](#)

**My Lab**

[Quick guide: how to create a laboratory account](#)

Copyright | Imprint

**ETH**  
Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

Figure 6. Homepage of the Genevestigator V3 database and toolbox available at <https://www.genevestigator.ethz.ch/>.

## Data Mining and Analysis Tools

**Meta-Profile Analysis** comprise six tools - Selection, Northern, Anatomy, Development, Stimulus and Mutation that allow the discovery of gene expression pattern across various arrays (Selection, Northern) or categories (Anatomy, Development, Stimulus and Station). All Meta-profile Analysis tools are based on „gene-centric approach“ so they return signal intensity values of pre-defined genes and genesets. Further, these tools can be divided into two categories according to conditions and gene-array approach used. The gene-condition approach groups all arrays with same conditions for instance identical developmental or anatomical category („meta-profiles“).

**Selection** tool displays how selected gene or genes are expressed across all arrays in the database. The user can choose arrays and genes for display. Moreover, each experiment can

be selected and highlighted in context of all experiments. Results are visualised as a scatter plot with magnifying rectangle.

**Northern** displays how selected gene or genes are expressed across chosen arrays in the database. Selection and Northern tools may depict expression profiles of many genes through many arrays simultaneously. Results are shown as an editable scatter plot with possibility to switch the axes (horizontal or vertical).

**Anatomy** depicts how gene(s) of interest are expressed in different plant tissues, organs and cell cultures organised in a tree of anatomical categories. In each category, the average signal intensity value (mean value) from all selected arrays belonging to this category is plotted. Expression values from children nodes are always involved in the mean expression value in a parent node. Results are visualised as a scatter plot or heat map. Users shall be aware that the scatter plot displays only up to ten genes.

The **Development** tool is very similar to the Anatomy tool (identical output formats and expression data types) but it shows how strongly the selected gene(s) are expressed in different developmental stages.

The **Stimulus** tool visualises gene expression responses to external stimuli and reveals up- or down-regulated genes. The same stimuli from different experiments are processed separately. The user then can see one stimulus in one plot repeatedly. Results are shown as scatter plot or heat map with the indication of number of arrays under treatment (T) and control (C) conditions used for the calculation of ratio and log ratio. Again, the scatter plot displays up to ten genes.

**Mutation** displays the gene expression response to the mutations or other genetic modifications in the genome as a ratio or log ratio value. The Mutation tool features are identical as in the Stimulus tool. Both tools also have identical layout, output and options.

**Biomarker Search** helps user to identify or validate biomarker genes in given category. It consists of the same toolbox as Meta-Profile Analysis (Anatomy, Development, Stimulus, Mutation). Specific biomarker identification is based on target-base search. The score defined to identify specifically expressed genes is derived from the sum of averages from the selected target and base categories. It is possible to display up to 400 genes per search. All available tools (**Anatomy, Development Stimulus and Mutation**) return the list of probe sets with the highest scoring genes in given target group over against a base. Whereas Anatomy and Development work with absolute signals, Stimulus and Mutation use relative expression values. The Anatomy tool is visualized as the expandable tree menu of individual categories. The Development tool is displayed as a pictorial list of development categories. The Mutation tool simply shows a list of mutation categories. Results are shown as a heat map with blue-white (absolute expression values) or green-red color coding (relative expression values)

**Clustering Analysis** is used to identify local patterns in gene expression data in given categories (Arrays, Anatomy, Development, Stimulus, Mutation) and to discover the possible relationship among gene expression patterns. Huge advantage of GV3 clustering tool is that it enables clustering of genes within their biological context using meta-profiles in the analysis (Anatomy, Development, Stimulus, Mutation). After selection of particular experiments and genes the user chooses the clustering method. There are two different algorithms available – Hierarchical Clustering and Biclustering. Whereas the first results in a two-dimensional matrix consisting of genes and conditions by selected gene cluster and profile cluster, the other uses exact BiMax algorithm that identifies all biclusters in a matrix.

**Pathway Projector** represents a reaction network analysis system. It combines expression data from transcriptomic studies and literature knowledge about signaling and metabolic pathways. Currently, the Pathway Projector analysis tool is available only for two organisms – *Arabidopsis thaliana* and *Mus musculus*. In few subsequent steps, the user defines comparison set (array set) that contains treatment and control sets and then selects appropriate subnetworks. The pre-defined or user-defined datasets for four ontologies (anatomy, development, stimulus, mutation) represent great advantage of this tool. In addition to display a network of interest, Pathway Projector also enables the user to build a new network from predefined pathways. Moreover, it is possible to import and export pathway maps as a GVP files.

### **Distinguishable Features**

The possibility to download graphical output in format and resolution ready for publication or web presentation represents very useful feature of newly released Genevestigator V3. The user can select between EPS, PNG or JPEG formats and set the figure parameters for export (width, height). Manual curation of microarray experiments represents great advantages of GV3. It results in Quality control (QC) that allows the detection of arrays with lower or poor quality and subsequently elimination of these arrays or experiment excluded from the microarray database. Identification of problematic arrays increases the output reliability of given data analysis. Genevestigator V3 categorizes experimental data that passed the QC check into separated subset called “High quality arrays only”.

### **NASCArrays**

NASCArrays [Craigon et al. 2004] is the Nottingham Arabidopsis Stock Centre's microarray database (<http://affy.arabidopsis.info/narrays/experimentbrowse.pl>) and it serves primarily as a repository for microarray data produced by NASC's transcriptomics service (Fig. 7). NASCArrays database harbours data from two types of Affymetrix GeneChips, AG Arrays and ATH1 Whole GenomeArrays. Microarray data are shared with other databases worldwide. In this database, individual arrays are referred to as “slides”. Index page of NASCArrays allows obtaining expression data by Search tools, Treeview and Data mining tools. The easiest way of data selection is the keyword search in Experiment Search, NASCArrays Reference Number and Slide search with using many selection criteria e.g. growth conditions, development stage, etc. Alternatively, individual experiments can be selected in a Treeview (list of experiments) containing all experiments categorized according to various criteria [Whetzel et al. 2006]. Moreover, the Slide selection tool allows researcher to perform desired data analysis by choosing individual slides. Immediately after selection of appropriate slides, set of data mining tools can be applied with the possibility of data download.

The experiment page shows details for each experiment (abstract, contact details of the author, information about the experiment and list of all slides belonging to the respective experiment with precise experimental information).

For download, one or several slides can be selected using “Slide selection“ function for the whole experiment or for up to 300 genes selected by bulk gene download. Data are supplied in one of two common data formats, CSV files or TAB-delimited text files. All data are well annotated with sample preparation details. Moreover, it is possible to download several data categories - Signal, StatPairsUsed, PresentCall, Detection P-value. For clustering purposes, only signal values may be downloaded.



**NASCArrays**

[start again](#), [selection](#), [spot history](#), [gene swinger](#), [two gene scatterplot](#), [digital northern](#), [bulk gene download](#), [tutorial](#), [help](#), [Affymetrix site](#), [NASC home](#)

## Welcome to NASCArrays.

NASCArrays is the [Nottingham Arabidopsis Stock Centre's](#) microarray database. Currently most of the data is for Arabidopsis thaliana experiments run by the NASC Affymetrix Facility. There are also experiments from other species, and experiments run by other centres too. *If you would like to see your data in this database, consult the [donation page](#).*

To navigate around this website, use the orange menus at the top of every page. "start again" will always bring you back to this page. To get the most out of this website, why not follow the tutorial (available from the "tutorial" link at the top of this page). Full documentation for all of the features on the website is available from the "help" section.

If you want to get large amounts of data from the database, you need AffyWatch! Affywatch is a CD subscription service, that for £50 will allow you quick access to all of the data on this website. [Click here](#) for more information.

For more information on NASC's Affymetrix service, you can visit our Affymetrix site by choosing Affymetrix site from the orange menu at the top, or by using Help.

There are three main ways into the data:

### 1) Search

**Experiment Search**

You can search for keywords to help you identify experiments you might be interested in.

Search for:

**Hints**

- Type in as many words as you like
- Type in words that you think are likely to be specific to what you are looking for rather than used a lot
- Consider words as "keywords" not phrases.

**Search by NASCArrays Reference Number**

Enter Reference Number (e.g. 343):  NASCARRAYS-

*(The number after the "NASCARRAYS-" prefix is synonymous with the Experiment ID.)*

**Slide Search**

This search allows you to search for slides that match your criteria. Type in terms you wish to search for, and the fields you wish to search for them in, below.

	<input type="text"/>	Name	▼	contains	<input type="text"/>
AND	<input type="text"/>	Name	▼	contains	<input type="text"/>
AND	<input type="text"/>	Name	▼	contains	<input type="text"/>
AND	<input type="text"/>	Name	▼	contains	<input type="text"/>
AND	<input type="text"/>	Name	▼	contains	<input type="text"/>

### 2) Use a data-mining tool

There are currently the following data mining tools available. All of these tools allow you to type in a gene(s) of interest, and identify experiments or slides that you might be interested in. Click on the links to use.

- [Spot History](#): This tool allows you to see the pattern of gene expression over all slides in the database. Easily identify slides (and therefore experimental treatments) where genes are highly, lowly, or unusually expressed.
- [Two gene scatter plot](#): This tool allows you to see the pattern of gene expression over all slides for two genes as a scatter plot. If you are interested in two genes, you can find out if they act in tandem, and highlight slides (and therefore experimental conditions) where these two genes behave in an unusual manner.
- [Gene Swinger](#): If you have a gene of interest, this tool will show you which experiment the gene expression varied most.
- [Bulk Gene Download](#): This tool allows you to download the expression of a list of genes over all experiments. You can get all genes over all experiments (*the entire database*) from the [Super Bulk Gene Download](#).

### 3) Treeview

The treeview allows you to browse the list of experiments. We have categorised all the experiments using various criteria.

To work the treeview, click on the categories. The branches below will expand, showing you all of the experiments in that category. To visit an experiment you are interested in, click on the link and you will be taken to a page about that experiment. One experiment can appear under more than one branch of the tree.

**Date** view experiments sorted by date

Figure 7. Homepage of the NASCArrays database available at <http://affy.arabidopsis.info/narrays/experimentbrowse.pl>.

Arabidopsis genome annotation supplied from Arabidopsis Ensembl (<http://atensembl.arabidopsis.info/index.html>) is based on TIGR3 version [Childs et al. 2007], (<http://www.tigr.org/>). For this reason, data must be interpreted carefully. However, user-supplied annotation for each sample is MIAME-compliant [Brazma et al. 2001] including information on how the original RNA sample preparation and treatment. The information about the each experiment and all used microarrays is very comprehensive. It can be accessed on the experiment page where each slide is represented by a slide box. Moreover, specially formatted data for clustering analysis using EBI-EPCLUST software [Kapushesky et al. 2004; <http://ep.ebi.ac.uk/EP/EPCLUST/>) can be acquired and all microarray data produced by NASC's Affymetrix service can be also mailed to users on CDs (AffyWatch, paid service).

### Specific Tools

**Spot history** displays the distribution of expression of a gene of interest over all experiments in the database and gives the user the chance to find experiments or individual slides with distant expression values for given gene and therefore to identify “unusual experiments”.

**Two-gene scatter plot** visualises expression profiles of two genes and reveals the relationships between these genes (co-expression) over all slides in a database.

**Gene swinger** is a web-based tool for experiment ranking according to the Coefficient of Variance value. It identifies experiments where the gene of interest shows the highest variability.

**Digital Northern** is an elementary tool for the visualization of relative expression signal and detection call (present, absent) of up to ten genes over all slides or over selected experiment or slides.

**Simple Pairwise Analysis** is a simple mining tool for identification of upregulated genes. After selection of just two slides it shows top 200 upregulated genes at each slide.

**Bulk gene download** enables download of up to 300 genes. After submitting list of AGI codes, related data from all the experiments in the database are downloaded.

### Distinguishable Features

Extensive user-supplied annotation for each sample meets MIAME requirements [Brazma et al. 2001]. Data describing each experiment are very comprehensive including information on how the original RNA sample was prepared and how the RNA sample was subsequently hybridised. This information can be accessed on the experiment page. Among advantageous features belongs the availability of expression data specially formatted for clustering analysis by EBI-EPCLUST software [Kapushesky et al. 2004; <http://ep.ebi.ac.uk/EP/EPCLUST/>) and the availability of paid AffyWatch mailing service.

### The Stanford Microarray Databases (SMD)

Originally, the database was established in 1999 for Stanford investigators but it is no longer restricted to local Stanford users. Now SMD serves as a data repository and microarray research database that is publicly available for all researches worldwide (Figure 8) [Sherlock



et al. 2001, Gollub et al. 2003]. Currently, the database contains the data for 53 organisms that represents almost 70,000 experiments (<http://smd.stanford.edu/statistics.html>). SMD supports wide collection of microarray platforms and software packages such as Affymetrix, Agilent, Combimatrix, GenePix, ScanAlyze, etc. Most data are public; however access to non-public data is limited to registered Stanford researchers and their collaborators. In principle, users can upload the data to the repository, subsequently retrieve and analyze their data using variety of data mining, analysing and displaying tools, and then, after publishing, made them available for the scientific community [Ball et al. 2005].

**Stanford MicroArray Database**

SMD Links Help

**SMD Login**

User Name:   
 Password:

**Public Data**

**SMD Announcements**

The XBabelPhish XML/MAGE-ML translator is available as XBabelPhish\_dist.zip at [SMD FTP Transfer Site](#)

**SMD Release 2.01**

- Primarily a bug-fix deployment
- SMD Release 2.00
- New biosequence data is implemented. This change

**Recent Publications**

 [Transcriptional modulation of genes encoding structural characteristics of differentiating enterocytes during development of a polarized epithelium in vitro.](#) Halbleib JM, et al. (2007) *Mol Biol Cell* 18(11):4261-78

 [Parallels between Global Transcriptional Programs of Polarizing Caco-2 Intestinal Epithelial Cells In Vitro and Gene Expression Programs in Normal Colon and Colon Cancer.](#) Saaf AM, et al. (2007) *Mol Biol Cell* 18(11):4245-60

 [Intrinsic androgen-dependent gene expression patterns revealed by comparison of genital fibroblasts from normal males and individuals with complete and partial androgen insensitivity syndrome.](#) Holterhus PM, et al. (2007) *BMC Genomics* 8(1):376

 [Expression of a pathogen-response program in peripheral blood cells defines a subgroup of Rheumatoid Arthritis patients.](#) van der Pouw Kraan TC, et al. (2007) *Genes Immun*

 [MicroRNA expression signature of human sarcomas.](#) Subramanian S, et al. (2007) *Oncogene*

 [Transcriptional program induced by wnt protein in human fibroblasts suggests mechanisms for cell cooperativity in defining tissue microenvironments.](#) Klapholz-Brown Z, et al. (2007) *PLoS ONE* 2(9):e945

 [Identification of a peripheral blood transcriptional biomarker panel associated with operational renal allograft tolerance.](#) Brouard S, et al. (2007) *Proc Natl Acad Sci U S A* 104(39):15448-53

 [Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors.](#) Kosinski C, et al. (2007) *Proc Natl Acad Sci U S A* 104(39):15418-23

**SMD Access:** Access to non-public data is limited to registered Stanford researchers and their collaborators. Please see [SMD Registration](#) for more specific information. If you have further questions regarding access, please e-mail the *Stanford Microarray Database* curators at [array@genome.stanford.edu](mailto:array@genome.stanford.edu).

**Proprietary Data:** Please note that some data in the database are proprietary and subject to legal restriction on their use, re-use and distribution. This includes but is not limited to Affymetrix and Agilent oligonucleotide sequences and patented sequences. It is the responsibility of the person viewing or downloading such data to ensure that such use does not infringe on the intellectual property rights of others.

**Project Funding:** The [National Cancer Institute](#) at the [US National Institutes of Health](#), the [Howard Hughes Medical Institute](#), and the [School of Medicine, Stanford University](#) fund the Microarray Database. The database is a joint project in the Departments of [Biochemistry](#) and [Genetics](#) at the [School of Medicine, Stanford University](#).

Database Copyright © 2001-2007 The Board of Trustees of Leland Stanford Junior University. Permission to use the information contained in this database was given by the researchers/institutes who contributed or published the information. Users of the database are solely responsible for compliance with any copyright restrictions, including those applying to the author abstracts. Documents from this server are provided "AS-IS" without any warranty, expressed or implied.

Please send comments or questions to: [array@genome.stanford.edu](mailto:array@genome.stanford.edu)

Figure 8. Homepage of the Stanford Microarray Database available at <http://genome-www5.stanford.edu/>.

After data retrieval, SMD provides various data processing and analysing tools. First, all data undergoes normalization and centering enabling cross-array comparison. Second, clustering algorithms reveal specific patterns within data using hierarchical clustering, self organising maps, singular value decomposition to identify missing data, alternatively, KNNImpute to estimate missing values in data matrices. Third, quality assessment tools offer other options including ANOVA analysis to detect spatial and print-tip bias on the array, HEEBO/MEEBO plots to view diagnostic and doping control, and graphing tools to produce histograms or scatter plots of user-selected fields [Demeter et al. 2007].

When searching for particular data, users can access variety of data mining tools in two different search forms – “Basic Search” and “Advanced Search”. Options available in these forms give users great flexibility in their searches to specify the organism of interest or authors, and to further refine categories and subcategories by application of cellular, biochemical or physiological limits. Moreover, researchers can select from particular experiments, specify genes and follows other filtering criteria and are allowed to edit or delete the annotations associated with the data. Finally, obtained results can be displayed in different formats and links are provided to explore more details about genes, arrays and experiments of interest.

### Specific Tools

**Quality Assessment** is an exploratory graphic tool producing histograms or scatter plots of user-selected fields to look at data points distribution, an ANOVA analysis to detect spatial and print tip bias on the array and a tool from BioConductor’s ArrayQuality package to view diagnostic and doping control plots for HEEBO/MEEBO arrays.

**Hierarchical Clustering** and other Data Analysis Tools are intended for data clustering and visualizing of co-expressed genes.

**Singular Value Decomposition** provides a mathematical algorithm for the determination of unique orthogonal gene and corresponding array expression patterns.

**KNNImpute** enables computation of missing values in data matrices before applying singular value decomposition (SVD) displayed as raster image or bar graph; SVD is a mathematical approach allowing determination of unique orthogonal genes and corresponding array expression patterns.

**GO TermFinder** determines whether a list of genes produced by any number of microarray analysis software tools has a significant enrichment of GO terms using Gene Ontology annotation.

**Array Color Tool** provides a simplified view of the data ratio for a given microarray allowing the user to quickly examine the microarray for evidence of global effects such as printing biases using ANOVA calculation to measure the dependence of per-spot ratio on printing plate and on sector.

**Experiment Selection Tool** selects from variety of experiments by slide name, organism, category and subcategory of experiments.

**Expresion History Tool** allows users to find microarray expression data for a gene throughout all experiments available.

**Data Filtering Options** offers variety of statistical measurements to filter and select data.

**Clustering and Image Generation** provides choice of self organizing maps or clustering tools using Pearson Correlation or Euclidean Distance.



**Gene Selection and Annotation Tool** enables selection of one or all genes and their annotation by description, gene model, GenBank accession, clone ID, or BAC locus.

### **Distinguishable Features**

SMD is a fast expanding database using extensive data mining tools for microarray data processing, analyzing, visualizing and sharing. Since employing wide spectra of microarray platforms, software, organisms and experimental conditions, the database is able to meet broad users' expectancy. SMD is currently the largest public repository for Arabidopsis Affymetrix microarray data. However, using complete data repository and all associated tools is possible only after paid registration for long-term data storage. SMD users can store dataset before and after filtering and centering in preclustering files. One can find useful to be able to precisely set up criteria for data mining and use the Expression History tool to explore the behavior of a gene of interest across all microarrays. SMD has also implemented pipe-line that converts sets of microarray data into MEGA-ML files and deposits them directly to ArrayExpress and GEO database.

## **CONCLUSION**

Enormous increase of gene expression profiling studies in recent years have supported rapid formation of extensive data repositories and development of number of on-line software applications allowing researchers to exploit available gene expression data. This chapter provided brief description and comparison of microarray-based gene expression databases and web-based tools. Special attention was paid to Affymetrix GeneChip platform that is the most widely used and serves as a standard for the model species *Arabidopsis thaliana*. However, several included databases comprise data acquired from various platforms of microarray-based experiments (ArrayExpress, SMD) or even non-array based technologies (GEO). Moreover, these databases contain expression data from different organisms such as SMD storing data from 53 organisms, GEO from over 100 individual organisms and ArrayExpress carrying data from even more than 200 organisms. From Affymetrix-focused portals, number of databases contains data from both array types – the first generation GeneChip® 8K AG Genome Arrays covering approximately one third of Arabidopsis genome and later and more complete ATH1 Whole Genome Array (22K). Fewer databases are focused solely on the later type of Affymetrix ATH1 GeneChip (BAR, aGFP).

The largest data repositories and databases with associated analysis tools often stand as integrated components of extensive bioinformatic projects run by large institutions like European Bioinformatics Institute (EMBL-EBI) or National Center for Biotechnology Information (NCBI). Other databases and web-based tools represent extra value added to microarray service (NASCArrays) or were created by established laboratories working in transcriptomic or bioinformatic field (BAR, aGFP). A subset of microarray databases were originally designed as public data repositories and retrieval systems for high throughput expression data. They serve mainly as data storage and offer fewer analysis tools (GEO, ArrayExpress, SMD). On the other hand, other software applications were designed primarily for data analysis. They exploit microarray data from many gene expression databases and offer large number of various analysis tools (BAR, Genevestigator V3).

Microarray data analysis tools can be divided into two main categories. The first category comprises gene expression visualization tools, whereas tools forming the second one offer more thorough analyses of gene expression data, for example gene grouping and display of co-expressed genes. The most advanced software applications (Genevestigator V3 or BAR) offer both types of analyses. Other tools were designed specifically for intuitive visualization of gene expression at various morphological and developmental stages (aGFP).

Almost all databases and web-based tools presented in this chapter are freely publicly available. Moreover, some databases offer the possibility of password-protected account to increase the work comfort (ArrayExpress). As an exception, Genevestigator V3 offers three levels of access – Open, Classic and Advanced. Classic and Advanced access options supplement free Open access, they are paid (with the exception of Classic access for academic users) and offer more data analysis tools as described previously. In addition, paying registration is required in the Stanford MicroArray Database for long-term storage.

Individual databases exploit various microarray data resources. Most databases process data obtained from publicly available data resources and repositories (NASC, FGCZ, GEO, TAIR, etc). In specialized cases like NASCArrays, they represent a functional overlay of existing microarray service. As a supplement, several projects allow user-controlled data submission (GEO, ArrayExpress),

And how to recognize the quality of experimental data in the database? First, the use of curated experiments is important prerequisite for correct interpretation of expression data. Currently, the large number of databases is strictly MIAME-compliant. The minimum information about microarray experiment (MIAME) contributes to increased comparability and of microarray data and so increases the value of interpretation of obtained results [Brazma et al. 2001]. In some cases, the use of alternative normalization algorithms may prove useful (aGFP). So, the user can identify the quality database by means of well-annotated raw and processed microarray data too. Further criteria comprise the assessment of expression data quality by measurement and evaluation of several quality control metrics. Quality control (QC) allows the detection of arrays with lower or poor quality and subsequent elimination of these arrays or whole experiments from the microarray database. Identification of problematic arrays increase the output reliability of given data analysis. Genevestigator V3 categorizes experimental data that passed the QC check into separated subset called “High quality arrays only”. Moreover, several databases allow users not only to extract gene expression data, but also facilitate the sharing of microarray experiment design and experimental protocols. This possibility is very helpful as it markedly improves the design of new experiments. In addition to gene expression data characterized by expression signal (sometimes along with detection call, p-value) and data obtained as a result of database query and data analysis, selected databases also include other data labels (gene annotation, gene ontology). On the other hand, some databases do not clearly mention the version of Arabidopsis genome annotation and this imperfection can contribute to data misinterpretation.

In addition to the database query, data analyses and data submission, most databases offer also data download. Database commonly support various data format options, from well-known XLS, CSV or Tab-delimited TXT files to specialised formats like SOFT text file (GEO). When downloading defined datasets, the user can often acquire additional information e.g. StatPairsUsed, PresentCall, Detection P-value (Genevestigator V3).

To date, the vast amount of transcriptomic data is publicly available in number of data repositories together with various data analysis tools providing instruments for their exploitation as well as assigning putative functional information to analysed genes. The large number of software applications has a user-friendly web-interface and broad scale of analysis tool, which may help researchers to open up new perspective of gene expression data and to exploit them for the planning of specifically targeted experiments and verification of their hypotheses.

## ACKNOWLEDGMENTS

Authors gratefully acknowledge the financial support from GACR (522/06/0896), GAAVCR (KJB600380701) and MSMT CR (LC06004, OC08011).

## REFERENCES

- Aharoni, A; Vorst, O. DNA microarrays for functional plant genomics. *Plant Mol Biol*, 2002, 48, 99-118.
- Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408, 796-815.
- Ball, CA; Awad, IAB; Demeter, J; Gollub, J; Hebert, JM; Hernandez-Boussard, T; Jin, H; Matese, JC; Nitzberg, M; Wymore, F; Zachariah, ZK; Brown, PO; Sherlock, G. *Nucleic Acids Research*, 2005, 33 D580-D582
- Barrett, T; Suzek, TO; Troup, DB; Wilhite, SE; Ngau, WC; Ledoux, P; Rudnev, D; Lash, AE; Fujibuchi, W; Edgar, R. NCBI GEO: mining millions of expression profiles - database and tools. *Nucleic Acids Res*, 2005, 33 Database Issue, D562-566.
- Barrett, T; Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, 2006a, 411, 352-369.
- Barrett, T; Edgar, R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO). *Methods Mol Biol*, 2006b, 338, 175-190.
- Barrett, T; Troup, DB; Wilhite, SE; Ledoux, P; Rudnev, D; Evangelista, C; Kim, IF; Soboleva, A; Tomashevsky, M; Edgar, R. NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, 2007, 35 Database issue, D760-765.
- Brazma, A; Hingamp, P; Quackenbush, J; Sherlock, G; Spellman, P; Stoeckert, C; Aach, J; Ansorge, W; Ball, CA; Causton, HC; Gaasterland, T; Glenisson, P; Holstege, FC; Kim, IF; Markowitz, V; Matese, JC; Parkinson, H; Robinson, A; Sarkans, U; Schulze-Kremer, S; Stewart, J; Taylor, R; Vilo, J; Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 2001, 29(4), 365-371.
- Brazma, A; Parkinson, H; Sarkans, U; Shojatalab, M; Vilo, J; Abeygunawardena, N; Holloway, E; Kapushesky, M; Kemmeren, P; Lara, GG; Oezcimen, A; Rocca-Serra, P; Sansone, SA. ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 2003, 31(1), 68-71.

- Brazma, A; Kapushesky, M; Parkinson, H; Sarkans, U; Shojatalab, M. Data storage and analysis in ArrayExpress. *Methods Enzymol*, 2006, 411, 370-386.
- Brenner, S; Johnson, M; Bridgham, J; Golda, G; Lloyd, DH; Johnson, D; Luo, S; McCurdy, S; Foy, M; Ewan, M; Roth, R; George, D; Eletr, S; Albrecht, G; Vermaas, E; Williams, SR; Moon, K; Burcham, T; Pallas, M; DuBridge, RB; Kirchner, J; Fearon, K; Mao, J; Corcoran, K. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 2000, 18, 630-634.
- Childs, KL; Hamilton, JP; Zhu, W; Ly, E; Cheung, F; Wu, H; Rabinowicz, PD; Town, CD; Buell, CR; Chan, AP. The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res*, 2007, 35(Database issue):D846-851.
- Clarke, JD; Zhu, T. Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems. practical considerations and perspectives. *Plant J*, 2006, 45, 630-650.
- Craigon, DJ; James, N; Okyere, J; Higgins, J; Jotham, J; May, S. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*, 2004, 32(Database issue), D575-577.
- Demeter, J; Beauheim, C; Gollub, J; Henandez- Bouccard, T; Jin, H; Maier, D; Matese, JC; Nitzberg, M; Wymore, F; Zachariah, ZK; Brown, PO; Sherlock, G; Ball, CA. *Nucleic Acids Research*, 2007, 35, D766-D770
- Donson, J; Fang, Y; Espiritu-Santo, G; Xing, W; Salazar, A; Miyamoto, S; Armendarez, V; Volkmuth, W. Comprehensive gene expression analysis by transcript profiling. *Plant Mol Biol*, 2002, 48, 75-97.
- Dupřáková, N; Reňák, D; Hovanec, P; Honysová, B; Twell, D; Honys D. Arabidopsis Gene Family Profiler (aGFP)--user-oriented transcriptomic database with easy-to-use graphic interface. *BMC Plant Biol*, 2007, 23, 7:39.
- Edgar, R; Domrachev, M; Lash, AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 2002, 30(1), 207-210.
- Edgar, R; Barrett, T. NCBI GEO standards and services for microarray data. *Nat Biotechnol*, 2006, 24(12), 1471-1472.
- Gollub, J; Ball, CA; Binklay, G; Demeter, J; Pinknelstein, DB; Hebert, JM; Hernandez-Boussard, T; Heng, J; Kaloper, M; Matese, JC; Schroeder, M; Brown, PO; Botstein, D; Sherlock, D. *Nucleic Acids Research*, 2003, 31 94-96
- Heazlewood, JL; Tonti-Filippini, J; Verboom, RE; Millar, AH. Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. *Plant Physiol*, 2005, 139(2), 598-609.
- Heazlewood, JL; Verboom, RE; Tonti-Filippini, J; Small, I; Millar, AH. SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res*, 2007, 35(Database issue), D213-218.
- Hennig, L; Menges, M; Murray, JA; Gruissem, W. Arabidopsis transcript profiling on Affymetrix GeneChip arrays. *Plant Mol Biol*, 2003, 53, 457-465.
- Hughes, TR; Marton, MJ; Jones, AR; Roberts, CJ; Stoughton, R; Armour, CD; Bennett, HA; Coffey, E; Dai, H; He, YD; Kidd, MJ; King, AM; Meyer, MR; Slade, D; Lum, PY; Stepaniants, SB; Shoemaker, DD; Gachotte, D; Chakraburttty, K; Simon, J; Bard, M; Friend, SH. Functional discovery via a compendium of expression profiles. *Cell*, 2000, 102, 109-126.

- Ilic, K; Berleth, T; Provar, NJ. BlastDigester - a web-based program for efficient CAPS marker design. *Trends Genet*, 2005, 21(1), 36.
- Jen, CH; Manfield, IW; Michalopoulos, I; Piney, JW; Willats, WGT; Gilmartin, PM; Westhead, DR. *Plant Journal*, 2006, 46, 336-348.
- Kapushesky, M; Kemmeren, P; Culhane, AC; Durinck, S; Ihmels, J; Körner, C; Kull, M; Torrente, A; Sarkans, U; Vilo, J; Brazma, A. Expression Profiler: next generation - an online platform for analysis of microarray data. *Nucleic Acids Res*, 2004, 32, W465-470.
- Laule, O; Hirsch-Hoffmann, M; Hruz, T; Gruissem, W; Zimmermann, P. Web-based analysis of the mouse transcriptome using Genevestigator. *BMC Bioinformatics*, 2006, 7:311.
- Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet*, 1999, 21, 20-24.
- Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 1996, 14, 1675-1680.
- Mandaokar A, Thines B, Shin B, Markus Lange B, Choi G, Koo YJ, Yoo YJ, Choi YD, Browse J. Transcriptional regulators of stamen development in Arabidopsis identified by transcriptional profiling. *Plant J*, 2006, 46, 984-1008.
- Manfield, WI; Jen, CH; Pinney, JW; Michalopoulos, I; Bradford, JR; Gilmartin, PM; Westhead, DR. *Nucleic Acids Research*, 2006, 34, 504-509.
- Money, T; Reader, S; Qu, LJ; Dunford, RP; Moore, G. AFLP-based mRNA fingerprinting. *Nucleic Acids Res*, 1996, 24, 2616-2617.
- Mukherjee, G; Abeygunawardena, N; Parkinson, H; Contrino, S; Durinck, S; Farne, A; Holloway, E; Lilja, P; Moreau, Y; Oezcimen, A; Rayner, T; Sharma, A; Brazma, A; Sarkans, U; Shojatalab, M. Plant-based microarray data at the European Bioinformatics Institute. Introducing AtMIAMExpress, a submission tool for Arabidopsis gene expression data to ArrayExpress. *Plant Physiol*, 2005, 139(2), 632-636.
- Palaniswamy, SK; James, S; Sun, H; Lamb, RS; Davuluri, RV; Grotewold, E. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol*, 2006, 140(3), 818-829.
- Parkinson, H; Sarkans, U; Shojatalab, M; Abeygunawardena, N; Contrino, S; Coulson, R; Farne, A; Lara, GG; Holloway, E; Kapushesky, M; Lilja, P; Mukherjee, G; Oezcimen, A; Rayner, T; Rocca-Serra, P; Sharma, A; Sansone, S; Brazma, A. ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 2005, 33 Database Issue, D553-555.
- Parkinson, H; Kapushesky, M; Shojatalab, M; Abeygunawardena, N; Coulson, R; Farne, A; Holloway, E; Kolesnykov, N; Lilja, P; Lukk, M; Mani, R; Rayner, T; Sharma, A; William, E; Sarkans, U; Brazma, A. ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 2007, 35 Database issue, D747-750.
- Rhee, SY; Beavis, W; Berardini, TZ; Chen, G; Dixon, D; Doyle, A; Garcia-Hernandez, M; Huala, E; Lander, G; Montoya, M; Miller, N; Mueller, LA; Mundodi, S; Reiser, L; Tacklind, J; Weems, DC; Wu, Y; Xu, I; Yoo, D; Yoon, J; Zhang, P. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, 2003, 31(1), 224-228.

- Rocca-Serra, P; Brazma, A; Parkinson, H; Sarkans, U; Shojatalab, M; Contrino, S; Vilo, J; Abeygunawardena, N; Mukherjee, G; Holloway, E; Kapushesky, M; Kemmeren, P; Lara, GG; Oezcimen, A; Sansone, SA. ArrayExpress: a public database of gene expression data at EBI. *C R Biol*, 2003, 326(10-11), 1075-1078.
- Sarkans, U; Parkinson, H; Lara, GG; Oezcimen, A; Sharma, A; Abeygunawardena, N; Contrino, S; Holloway, E; Rocca-Serra, P; Mukherjee, G; Shojatalab, M; Kapushesky, M; Sansone, SA; Farne, A; Rayner, T; Brazma, A. The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics*, 2005, 21(8), 1495-1501.
- Schena, M; Shalon, D; Davis, RW; Brown, PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 1995, 270, 467-470.
- Schmid, M; Davison, TS; Henz, SR; Pape, UJ; Demar, M; Vingron, M; Scholkopf, B; Weigel, D; Lohmann, JU. A gene expression map of Arabidopsis thaliana development. *Nat Genet*, 2005, 37(5), 501-506.
- Schwager, C; Blake, J. Bload - a batch loader application for MIAMExpress. *Bioinformatics*. 2005, 21(8):1727-1729.
- Sherlock, G; Hernandez-Boussard, T; Kasarskis, A; Binkley, G; Matese, JC; Dwight, SS; Kaloper, M; Weng, S; Jin, H; Ball, CA; Eisen, MB; Spellman, PT; Brown, PO; Botstein, D; Cherry, JM. Nucleic Acids Research, 2001, 152-155.
- Spellman, PT; Miller, M; Stewart, J; Troup, C; Sarkans, U; Chervitz, S; Bernhart, D; Sherlock, G; Ball, C; Lepage, M; Swiatek, M; Marks, WL; Goncalves, J; Markel, S; Iordan, D; Shojatalab, M; Pizarro, A; White, J; Hubley, R; Deutsch, E; Senger, M; Aronow, BJ; Robinson, A; Bassett, D; Stoeckert, CJ Jr; Brazma, A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 2002, 3, RESEARCH0046.
- Taylor, J; Provart, NJ. CapsID: a web-based tool for developing parsimonious sets of CAPS molecular markers for genotyping. *BMC Genet*, 2006, 10, 7:27.
- Torrente, A; Kapushesky, M; Brazma, A. A new algorithm for comparing and visualizing relationships between hierarchical and flat gene expression data clusterings. *Bioinformatics*, 2005, 21(21), 3993-3999.
- Toufighi, K; Brady, SM; Austin, R; Ly, E; Provart, NJ. The Botany Array Resource: e-Northerns; Expression Angling; and promoter analyses. *Plant J*, 2005, 43(1), 153-63.
- Troyanskaya, O; Cantor, M; Sherlock, G; Brown, P; Hastie, T; Tibshirani, R; Botstein, D; Altman, RB. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 2001, 17(6), 520-525.
- van Hal, NL; Vorst, O; van Houwelingen, AM; Kok, EJ; Peijnenburg, A; Aharoni, A; van Tunen, AJ; Keijer, J. The application of DNA microarrays in gene expression analysis. *J Biotechnol*, 2000, 78, 271-280.
- Velculescu, VE; Zhang, L; Vogelstein, B; Kinzler, KW. Serial analysis of gene expression. *Science*, 1995, 270, 484-487.
- Wheeler, DL; Barrett, T; Benson, DA; Bryant, SH; Canese, K; Church, DM; DiCuccio, M; Edgar, R; Federhen, S; Helmberg, W; Kenton, DL; Khovayko, O; Lipman, DJ; Madden, TL; Maglott, DR; Ostell, J; Pontius, JU; Pruitt, KD; Schuler, GD; Schriml, LM; Sequeira, E; Sherry, ST; Sirotkin, K; Starchenko, G; Suzek, TO; Tatusov, R; Tatusova, TA; Wagner, L; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2005, 33(Database issue), D39-45.

- Wheeler, DL; Barrett, T; Benson, DA; Bryant, SH; Canese, K; Chetvernin, V; Church, DM; DiCuccio, M; Edgar, R; Federhen, S; Geer, LY; Kapustin, Y; Khovayko, O; Landsman, D; Lipman, DJ; Madden, TL; Maglott, DR; Ostell, J; Miller, V; Pruitt, KD; Schuler, GD; Sequeira, E; Sherry, ST; Sirotkin, K; Souvorov, A; Starchenko, G; Tatusov, RL; Tatusova, TA; Wagner, L; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2007, 35(Database issue), D5-12.
- Whetzel, PL; Parkinson, H; Causton, HC; Fan, L; Fostel, J; Fragoso, G; Game, L; Heiskanen, M; Morrison, N; Rocca-Serra, P; Sansone, SA; Taylor, C; White, J; Stoeckert, CJ Jr. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 2006, 22(7), 866-73.
- Winter, D; Vinegar, B; Nahal, H; Ammar, R; Wilson, GV; Provart, NJ. An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, 2007, 2(1), e718.
- Yang, YH; Dudoit, S; Luu, P; Lin, DM; Peng, V; Ngai, J; Speed, TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 2002, 30(4), e15.
- Zhang, W; Morris, QD; Chang, R; Shai, O; Bakowski, MA; Mitsakakis, N; Mohammad, N; Robinson, MD; Zirngibl, R; Somogyi, E; Laurin, N; Eftekharpour, E; Sat, E; Grigull, J; Pan, Q; Peng, WT; Krogan, N; Greenblatt, J; Fehlings, M; van der Kooy, D; Aubin, J; Bruneau, BG; Rossant, J; Blencowe, BJ; Frey, BJ; Hughes, TR. The functional landscape of mouse gene expression. *J Biol*. 2004, 3, 21.
- Zhu, T. Global analysis of gene expression using GeneChip microarrays. *Curr Opin Plant Biol*, 2003, 6, 418-425.
- Zimmermann, P; Hirsch-Hoffmann, M; Hennig, L; Gruissem, W. GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. *Plant Physiol*, 2004, 136(1), 2621-2632.
- Zimmermann, P; Hennig, L; Gruissem, W. Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci*, 2005, 10(9), 407-409.