TAIR User guide

Getting Started

Browser compatibility and configuration.

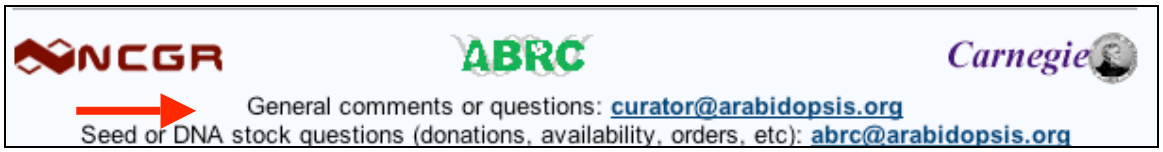
The majority of the website has been tested for compatibility with different browsers on Mac and Windows operating systems. We recommend the following browsers:
PC: Internet Explorer 6 and above, Netscape 6 and above, Firefox
Mac: Firefox, Internet Explorer 5 and above, Safari

For registration and processing stock orders you need to have cookies enabled in your browser. The site makes extensive use of Javascript, therefore you should also enable this function in your browser.

Some features on the website may not work as expected if you have pop-ups blocked on your browser.

Help is always available from the navigation bar

or contact us directly…



Additional Resources

For help in setting up your browser see the section “Configuring your browser to use TAIR” on the following web page.
http://arabidopsis.org/help/

Finding help documents for TAIR tools

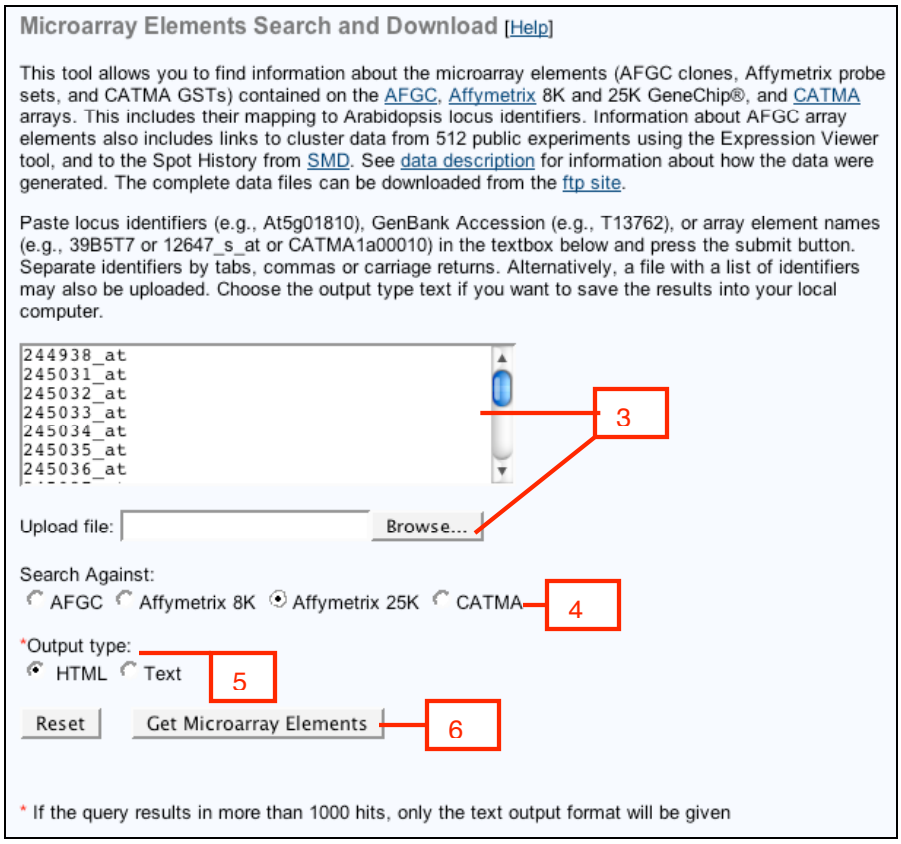
Most TAIR searches and analysis tools have links to on-line help documents that will guide you in how to perform searches and use the results. A list of these documents can be found at http://www.arabidopsis.org/help/helpcontents.jsp. All of the help documents, tutorials, glossary of terms used in TAIR, Quickstart guide and a FAQ can be found in the on-line help section (http://www.arabidopsis/org/help/index.jsp).

Requesting Help.

For general problems and questions about TAIR contact the TAIR curators at curator@arabidopsis.org.
For problems with stock orders or questions about stocks: abrc@arabidopsis.org

Finding Genes and Annotations for Microarray Elements

The Microarray Element Search can be used to find genes that correspond to an array element using array element names, or GenBank accessions (for spotted cDNA arrays). Alternatively, you can use locus identifiers to find the corresponding array element on a given array.



Using the Microarray Element Search

1. From the TAIR home page, find the Advanced Search section and click on the link to Microarray element. Or type in the URL: http://www.arabidopsis.org/tools/bulk/microarray/index.jsp
2. Go to the TAIR ftp site tmp directory (ftp://ftp.arabidopsis.org/home/tair/tmp/) and locate the sample file (probeset_sample).
3. Paste the list of probe names into the text input box. Alternatively, if you have a file saved on your computer you can upload the file from your computer.
4. Choose the array design to search against. You can only search one type of array design at a time.
5. Choose the HTML output option. Choose text if you want to save the file to your personal computer as a text file.
6. Submit the search by clicking the 'Get Microarray Elements' button.

Microarray Elements Search Results [Help]

Array Element	Locus Identifier	Annotation	Organism	Probe Type	Is Control
244938_at	ATCG01120	rps15 ribosomal protein S15	Arabidopsis thaliana	oligonucleotide	no
245031_at	AT2G26360	mitochondrial substrate carrier family protein contains Pfam profile: PF00153 mitochondrial carrier protein	Arabidopsis thaliana	oligonucleotide	no
245032_at	AT4G04635	hypothetical protein	Arabidopsis thaliana	oligonucleotide	no
245033_at	AT2G26380	disease resistance protein-related / LRR protein-related contains leucine rich-repeat domains Pfam:PF00560, INTERPRO:IPR001611; similar to Hcr2-2A [Lycopersicon pimpinellifolium] gi|3894389|gb|AAC78594	Arabidopsis thaliana	oligonucleotide	no
245034_at	AT2G26390	serpin, putative / serine protease inhibitor, putative similar to phloem serpin-1 [Cucurbita maxima] GI:9937311; contains Pfam profile PF00079: Serpin (serine protease inhibitor)	Arabidopsis thaliana	oligonucleotide	no
245035_at	AT2G26400	acireductone dioxygenase (ARD/ARD') family protein similar to iron-deficiency induced gene [Hordeum vulgare] GI:14522834, SIPL [Homo sapiens] GI:16551383; contains Pfam profile PF03079: ARD/ARD' family	Arabidopsis thaliana	oligonucleotide	no
245036_at	AT2G26410	calmodulin-binding family protein similar to SF16 protein [Helianthus annuus] GI:560150; contains Pfam profile PF00612: IQ calmodulin-binding motif	Arabidopsis thaliana	oligonucleotide	no
245037_at	AT2G26420	1-phosphatidylinositol-4-phosphate 5-kinase, putative / PIP kinase, putative / PtdIns(4)P-5-kinase, putative / diphosphoinositide kinase, putative similar to phosphatidylinositol-4-phosphate 5-kinase AtPIP5K1 [Arabidopsis thaliana] GI:3702691; contains Pfam profiles PF01504: Phosphatidylinositol-4-phosphate 5-Kinase, PF02493: MORN repeat	Arabidopsis thaliana	oligonucleotide	no

Microarray Element Search HTML results page

The file lists the corresponding locus name which is hyperlinked to the TAIR locus details. The file also includes the gene description field (shown here as 'Annotation)**.**

245030_at	AT2G26620 AT2G15450 AT2G15470 AT2G15460	[AT2G26620, glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein similar to SP|P35339 Exopolygalacturonase precursor (EC 3.2.1.67) (Pectinase) (Galacturan 1,4-alpha-galacturonidase) {Zea mays}; contains Pfam profile PF00295: Glycosyl hydrolases family 28 (polygalacturonases)];[AT2G15450, glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein similar to SP|P35339 Exopolygalacturonase precursor (EC 3.2.1.67) (Pectinase) (Galacturan 1,4-alpha-galacturonidase) {Zea mays}; contains Pfam profile PF00295: Glycosyl hydrolases family 28 (polygalacturonases)];[AT2G15470, glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein similar to SP|P35339 Exopolygalacturonase precursor (EC 3.2.1.67) (Pectinase) (Galacturan 1,4-alpha-galacturonidase) {Zea mays}; contains Pfam profile PF00295: Glycosyl hydrolases family 28 (polygalacturonases)];[AT2G15460, glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein similar to SP|P35339 Exopolygalacturonase precursor (EC 3.2.1.67) (Pectinase) (Galacturan 1,4-alpha-galacturonidase) {Zea mays}; contains Pfam profile PF00295: Glycosyl hydrolases family 28 (polygalacturonases)]	Arabidopsis thaliana	oligonucleotide	no

Example of an array element that maps to more than one locus.

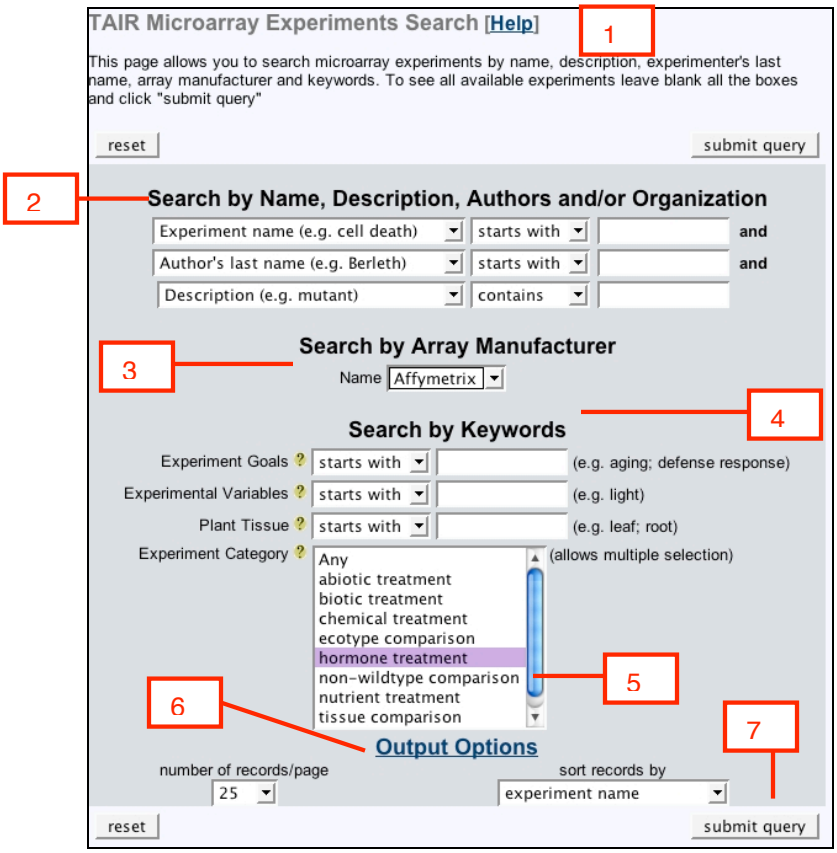
Some array elements were designed to detect paralogs and have more than one associated locus.

Additional Resources

The entire set of TAIR's mappings between array elements and loci can be downloaded in tab-delimited format from our FTP site (ftp://ftp.arabidopsis.org/home/tair/Microarrays/). Please see the README files for descriptions of the files. These files are updated whenever the genome annotation changes. The most recent version is based on the TIGR5.0 genome release (Jan 2004).

Finding microarray experiments and datasets

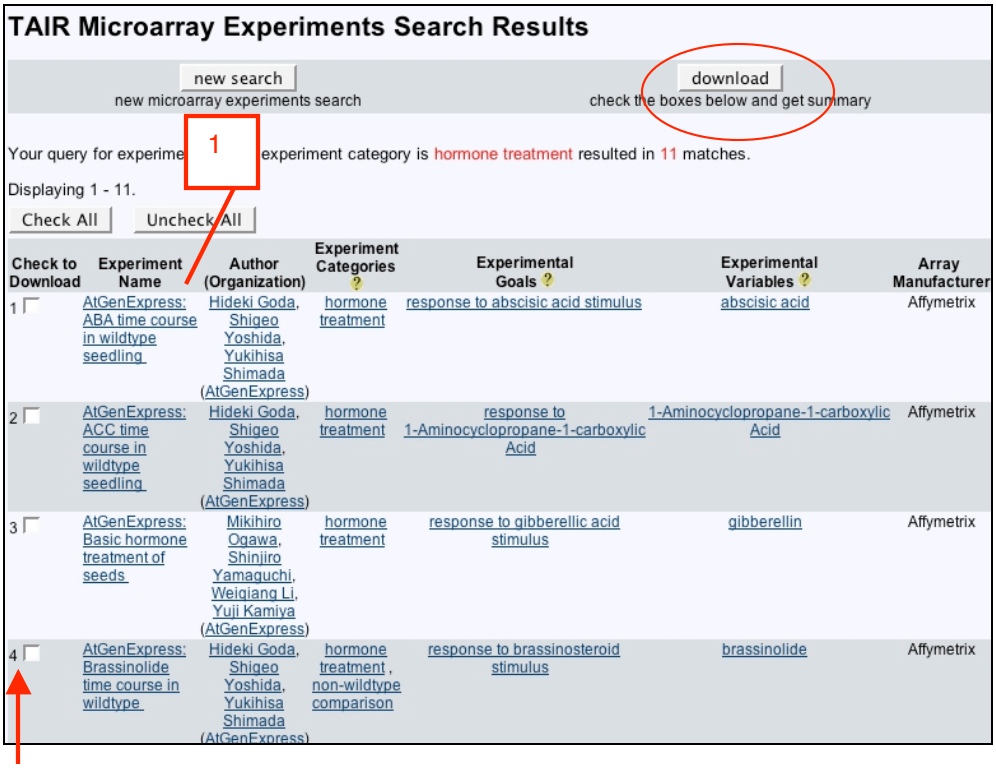
The next section describes how to use the Microarray Experiment Search to find data and information about microarray elements stored in TAIR's database.



Using the Microarray Experiment Search

1. Start at the TAIR home page and in the Advanced Search Section, click on Microarray Experiment Search page.
2. The first option allows you to search for experiments by experiment name, submission number ,description, author's name or organization. If more than one option is specified, the second parameter is included as an implicit AND. A search with organization = "ATGenExpress" and description contains "Atlas" would find the AtGenExpress Developmental Atlas experiment.
3. Next, choose the array manufacturer. The default option is set to 'Any' which will retrieve experiments from all types of manufacturers.
4. Searching with Keywords. You can use keywords to limit your results based specific experiment parameters such as goals, variables, tissue used for RNA extraction or category. If terms are entered in multiple options, the search is treated as an implicit AND.
5. Selecting by experiment category. To find all hormone treatments, use this option. To choose more than one category press the CTRL key (PCs) or the Apple key (Mac) when making your selections with the mouse. The default option (ANY) will include all types of experiments in the results set.
6. Select the output format. The output options can be set to display up to 200 records per page of results. The format of the results page can be ordered by experiment category, name, experimenter's name, goals or variables.
7. Click the submit query button.

IMPORTANT NOTE: If you are not sure of exactly what you are looking for, use less rather than more parameters. If you get too many results you can always go back and apply more filters.



The results of a query for hormone treatments using Affymetrix chips.

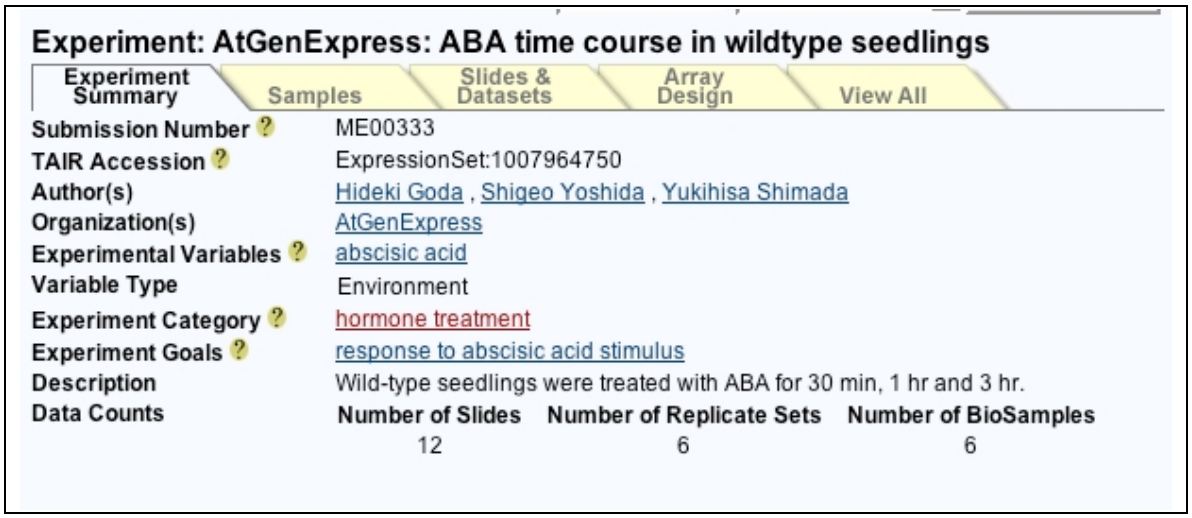
1. Click on the experiment named 'AtGenExpress ABA Time Course' to view the experiment detail page.

Clicking on the authors/organizations name will display their community detail page with contact information. Clicking on any of the keywords such as the experiment category keyword, experimental goal, experimental variables links to the keyword detail page where you can find microarray experiments, genes and papers associated to the same term. For example, click on the experimental goal 'response to abscisic acid stimulus to find other microarray experiments, genes involved in responding to ABA and papers about ABA responsiveness.

The check boxes (arrow) can be used to select search results to download (circled button). NOTE the download only downloads what you see on the results page, it does NOT download the experimental data itself. If you want the entire data sets- use the ExpressionSet identifier on the Experiment details to locate the file in the FTP site (ftp://ftp.arabidopsis.org/home/tair/Microarrays/Datasets/).

Understanding and using the Experiment Details

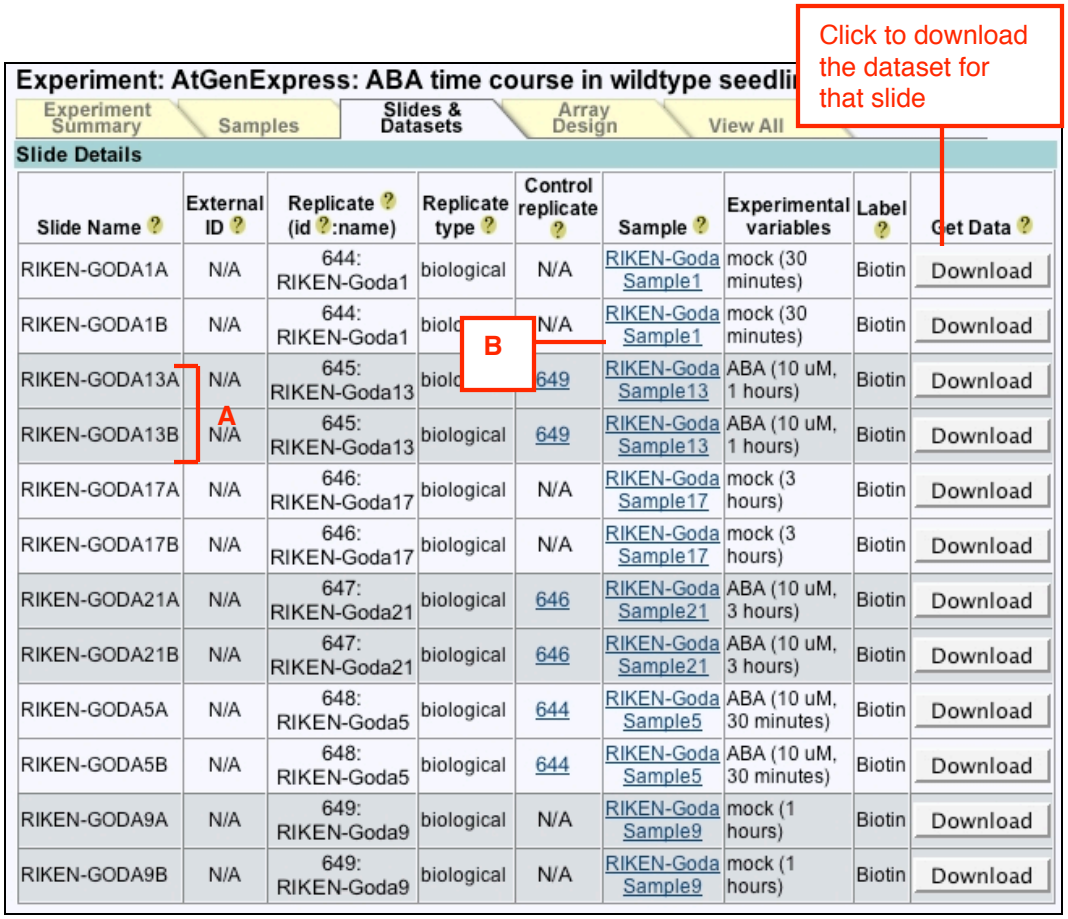
The Experiment detail page can be accessed by clicking on the name of an experiment in the list of results that matched your query. A set of tabs at the top of the page allows you to navigate quickly to different sub sections of the data. This section describes the contents of the Experiment Details and their uses. More information is available in the Experiment Search/Results and Detail page help document. To navigate between sections of the experiment details, click on the tab.



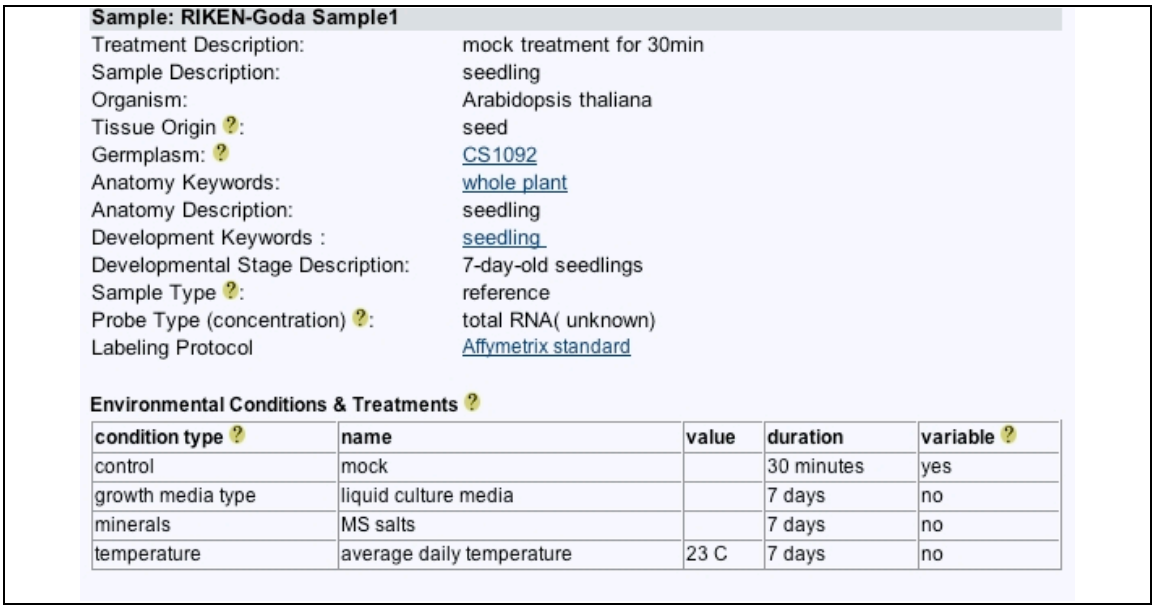
Experiment Summary page

The first section displayed shows information about the experimenter, experimental variables, number of slides in the experiment and an abstract summarizing the experiment.. Each experiment in TAIR is considered an "ExpressionSet" that includes multiple slides. The total number of slides in the experiment is shown on the bottom of this page along with the number of those slides which are either biological or technological replicates. The abstracts submitted by the experimenters, should provide an overview of the goals of the experiment. If there are papers associated to the experiment, these will also be displayed on the summary page. The tabs are used to navigate to different sections of the data.

The hyperlinks on this page function like the ones on the results. They link to detail pages in TAIR such as people/labs or to the keyword details.



Slides and Datasets:
This is the section where the main information about the slides that comprise the experiment is stored. Replicates are grouped together in alternating color bands (A). You can scan through the list of slides in the experiment and download the data for the slides you are most interested in. Each slide has a link to the sample data section where you can find information about the RNA sample used for hybridization (B). For each data set you want to download, click on the 'Download data' button. The data files are in a tab delimited text file which can be opened in a spreadsheet program such as Microsoft Excel.



Sample: RIKEN-Goda Sample1

Treatment Description:	mock treatment for 30min
Sample Description:	seedling
Organism:	Arabidopsis thaliana
Tissue Origin:	seed
Germplasm:	CS1092
Anatomy Keywords:	whole plant
Anatomy Description:	seedling
Development Keywords :	seedling
Developmental Stage Description:	7-day-old seedlings
Sample Type:	reference
Probe Type (concentration):	total RNA(unknown)
Labeling Protocol	Affymetrix standard

Environmental Conditions & Treatments

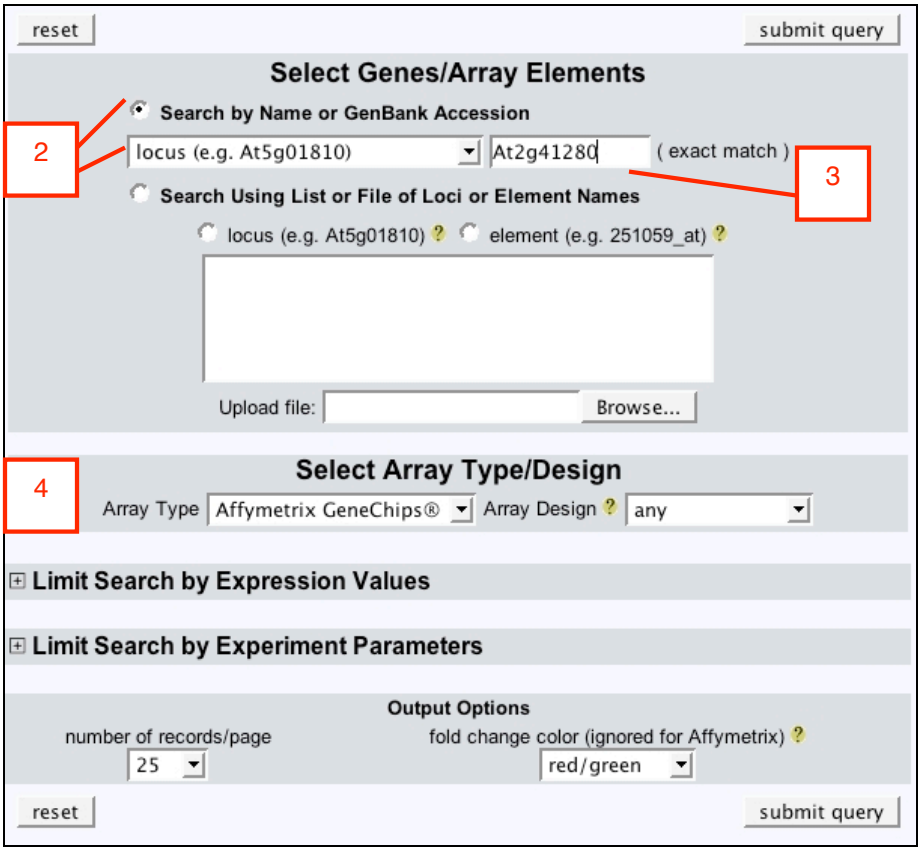
condition type	name	value	duration	variable
control	mock		30 minutes	yes
growth media type	liquid culture media		7 days	no
minerals	MS salts		7 days	no
temperature	average daily temperature	23 C	7 days	no

Sample Details

Each tissue sample used to prepare RNA for the experiment is described in this section. Each sample data has a table which lists all of the environmental conditions applied to that sample. In addition to the sample descriptions provided by the data donor, TAIR annotates the sample data using controlled vocabularies (Plant Ontologies) to describe anatomy and development. These keywords are in turn linked to keyword details where you can find other types of data (or other microarray experiments) which used similar tissue types. For each entry, the experimental variables are listed which allows you to find specific datasets that examine a variable of interest. You can scan the variables to find the tissue samples of interest. For example, if you want to compare expression values between mock and treated tissues, you can select and download these hybridization data. If you were only interested in the differences between genes expressed in different ABA concentrations then you might want to only download and analyze that subset of data.

Finding information about the expression of a gene or set of genes

The Microarray Expression Search tool can be used to perform a simple search by name for expression data from a single gene or set of genes. The Advanced Options allow you to restrict your search to expression data that meets specific criteria.

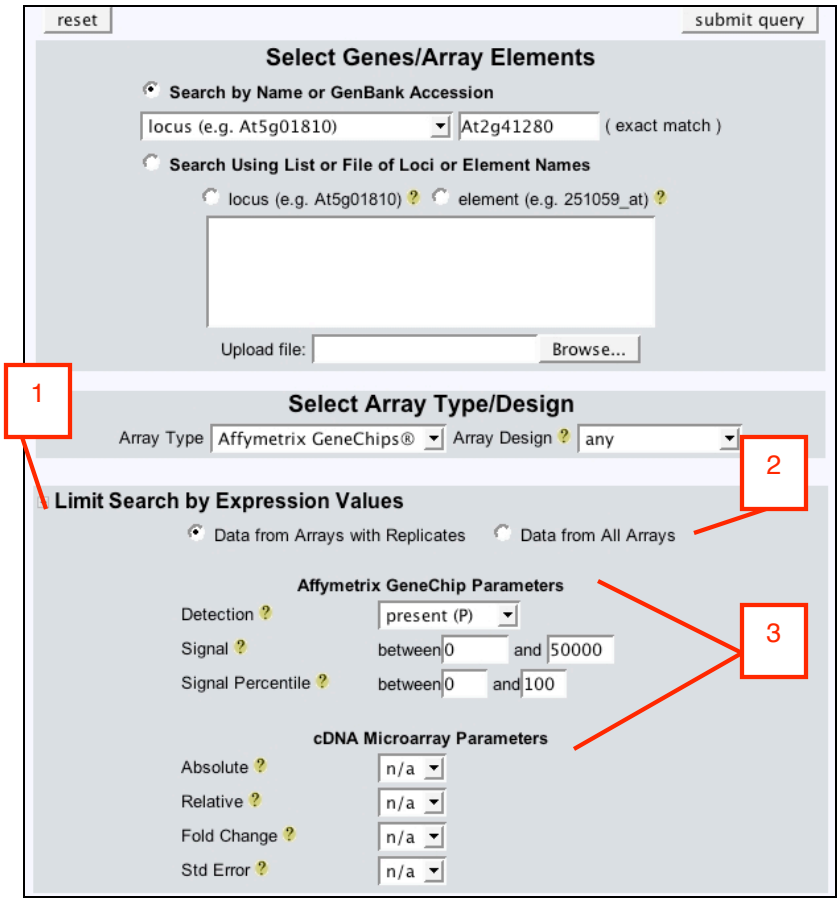


Using the Microarray Expression Basic search functions

1. From the TAIR home page, click on the link to the Microarray Expression in the Advanced Search section.
2. Choose the locus name from the name type drop down list.
3. Enter the name AT2G41280.
4. Select Affymetrix for the type of array/array design

This option allows you to limit the results by array platform and design. The default option only includes results from single channel arrays (e.g. Affymetrix). To search only within cDNA arrays, choose this option. As of January 2005, all cDNA array data in TAIR is from the AFGC project.

IMPORTANT NOTE: If you are searching with array element names or GenBank accessions you MUST choose the appropriate array type, otherwise you may get false negative results. We recommend using the broadest possible options -for either platform, choose any array design.



Using the Microarray Expression Advanced Search options

The advanced options can be accessed by clicking on the plus sign next to each of the optional fields.

Limiting the search by expression values

The default search will return results only for replicate hybridizations from single channel arrays. Depending on the type of array selected in the previous step, different parameters are available for restricting search results based upon expression values. These are optional parameters.

1. Expand this selection by clicking on the plus [+] sign.
2. If you prefer to return results from all hybridizations, select the Data from All Arrays option. This will include data from hybridizations without replicates which may be of lower significance.
3. Choose expression value options depending on the platform you selected before.

Affymetrix Array Options

- Detection: This option allows you to limit results based on whether or not expression of a gene was detectable above background. The default option is set to "Present" meaning only hybridizations where the gene is 'expressed' will be included. Choosing the "Absent" options will return results for which the level of expression was not significantly increased over background.
- Signal: This option allows you to specify a range of expression values for the gene(s) of interest. The signal strength between arrays are comparable as all Affymetrix data is normalized to a target value of 200. An approximation of signal intensity to transcript abundance is shown below.

> 20: not expressed or very low abundance; 20-50: low; >50-200: moderate'>200, high

- Signal Percentile: This option allows you to restrict results to only those hybridizations in which the relative expression of the target gene is above a certain threshold. This option is useful for selecting only those hybridizations in which your gene(s) of interest are most highly induced relative to other genes represented on the array.

cDNA array options.

- Absolute Expression: The default option is 'Expressed' which includes only those experiments in which absolute level of expression of a gene was above a defined threshold once the background is subtracted. Choosing the not expressed option allows you to find experiments/conditions under which the target gene does not appear to be expressed above background.

- Relative Expression: The default option (Any) includes all hybridizations regardless of the degree of increased or decreased expression. You can use this option to limit the results to only those conditions under which the target gene is increased, decreased or unchanged.

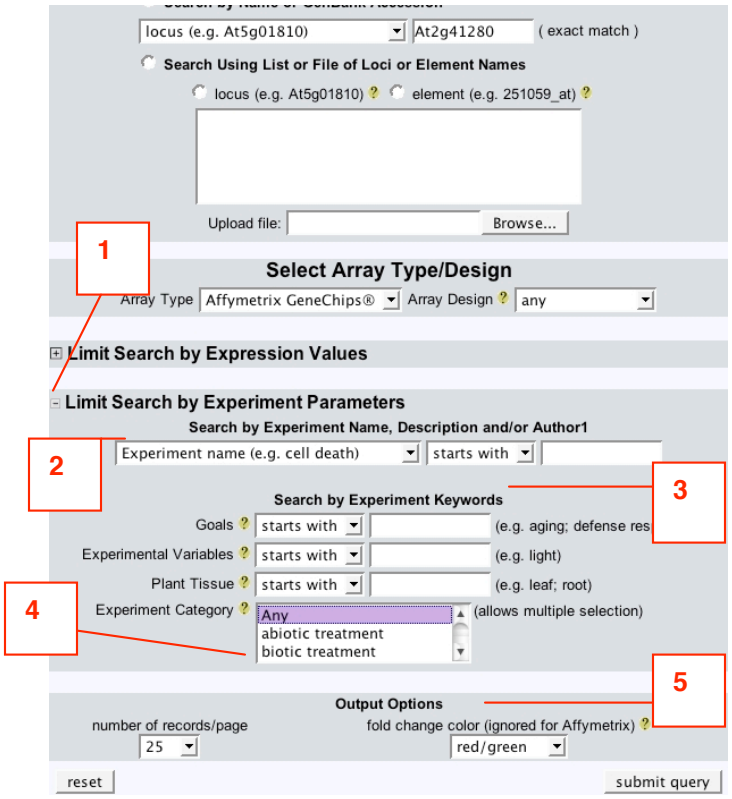
- Fold Change: This option can be used in combination with the Relative Expression option, to indicate the degree of increased or decreased expression.

- Standard Error: This refers to the standard error for the overall fold change. You can use this option to set a 'quality' threshold for results (e.g. a smaller value means there is less variation among replicates). For best results leave the default value, Any. If necessary you can go back and re-do the query with more restrictive parameters.

Limiting the search by experiment parameters

The optional parameters in this section can be used to define a subset of expression values to display based upon characteristics of the experiment. For example, if you are interested in finding out how the expression of your gene is affected by environmental or developmental conditions. This option is particularly useful for narrowing down conditions under which your gene(s) of interest have the most varied expression. Also, it can be useful for obtaining smaller and more manageable data sets.

Remember, it is NOT necessary to select any of these options. The default parameters are the least restrictive and will return results regardless of the experimental parameters. First try the search without changing these parameters. If you get too many results you can always go back and refine your search.

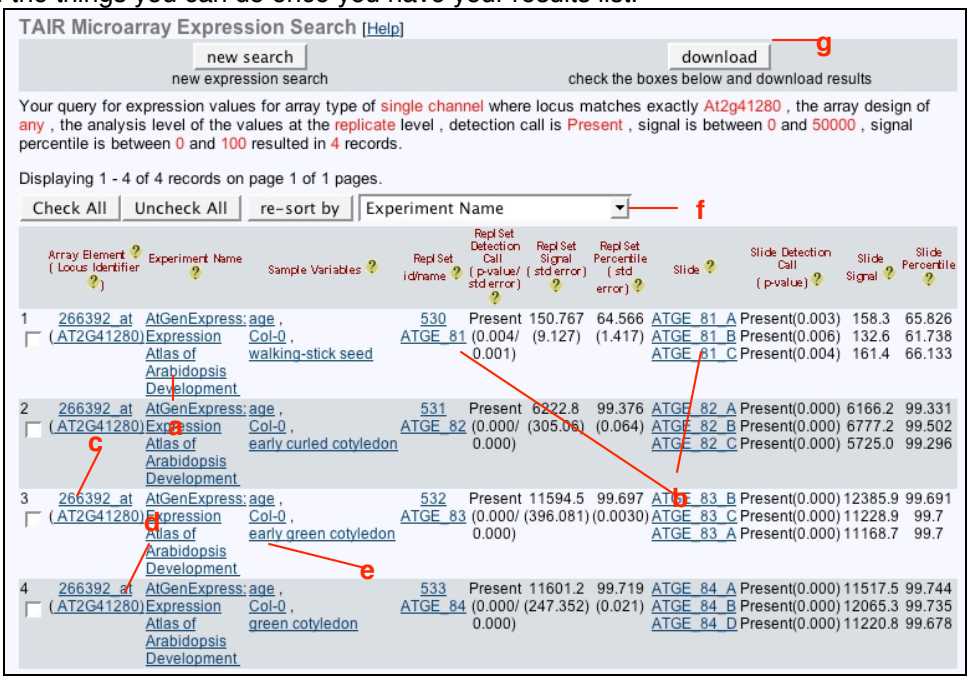


1. Expand the section by clicking on the plus [+] sign.
2. Limit the search by Experiment name. These options can be used to limit the expression results set to include only the defined named experiments, or experimenters.
3. Limit the search by keywords . Within this section are several options which allow you to input keywords and find expression values for all experiments annotated with those keywords.
4. Limit the search by experiment category. Select one or more categories of experiments to include in the search. The default option includes all experiments regardless of type. To select more than one category, hold down the CTRL key (PC) or Apple key (Mac) when making your selections with a mouse click.
5. Define the output format. Select the number of results per page to display and the color scheme for showing the fold change. You can choose to display up to 200 individual results per page. Choosing the most records per page is a good idea, especially if you plan on downloading the results. You can always go back and redo the search with more filters.

Understanding and interpreting the Expression Search Results

A successful query will return a list of results that match your search criteria. The format of the results will differ depending upon the array type option you selected in step 5. If you have not done the sample query, you can view the sample Single Channel Results or Dual Channel Results.

The results page lists all of the replicate hybridizations that match your query (and may include non-replicated hybridizations if you chose that option).The upper portion of the results shows what search criteria were used and lists the number of matching records. The following items list some of the things you can do once you have your results list.

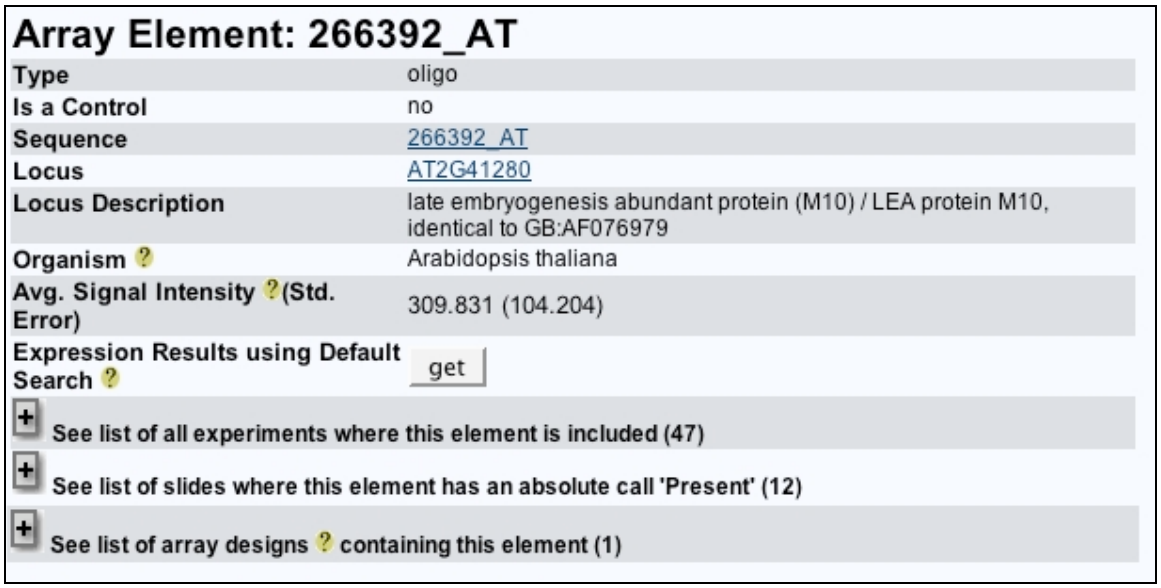


a. Find information about the experimental methods and sample treatments, click on the experiment name. For more information about the contents of the experiment details and navigating expression set data, see the Microarray Experiment Search tutorial (http://www.arabidopsis.org/help/tutorials./micro_intro.jsp).
b. Find and download the datasets. Click on the name of the replicate set, or the individual slide name if you just want information about that specific hybridization. From the slide/dataset details you can choose to download the dataset or find out more about the RNA sample used for the hybridization.
c. Find other experiments that include this array element. Click on the array element name to view the detailed information about this element including a list of all experiments in which the element is included on the array.
d. Find other information about the locus by clicking on the (AGI) locus name. This will open a new view showing the TAIR locus detail page. From this page you can find other information such as functional annotations, alleles/polymorphisms, gene and protein features and publications.
e. View a description of the sample treatment for each slide variables. Click on the sample variable terms to view the sample details for that hybridization.
f. This option allows you to sort the results by different parameters, such as locus or array element name (useful if you have uploaded a file of more than one element or locus), experiment, expression values/fold change. The different options allow you to find

experiments in which the expression of your gene of interest varies with different conditions, or to find experiments in which the expression values were highest or lowest.

1. Select the appropriate field from the drop down menu (e.g. Experiment Name). Click on the 're-sort by' button. If you chose the example above, the results would be displayed according to the name of the experiment. All replicate sets belonging to one experiment will be grouped together.

g. One or more rows of results can be downloaded as a tab-delimited text file. These files can then be opened using a simple text editor or spreadsheet program such as Microsoft Excel.
   1. Select the records to download by checking the box at the far left side of each row.
   2. Alternatively, if you want to download ALL of the records on a single page, use the 'Check All' option next to the re-sort button.
   3. Download the file by clicking on the 'download' button below the TAIR toolbar. You will need to do this for each of the pages of results. Currently the download button only functions for a page of results at a time.
   4. Save the file to the hard disk of your computer.



Array Element Detail page.
From this page you can a) find all experiments where expression has been assayed using the default expression search parameter, b) find slides where expression was detected -signal call was ‘Present’. c)You can also find all the experiments in TAIR which included the element. For example, for array elements that exist on more than one array design.

Additional resources

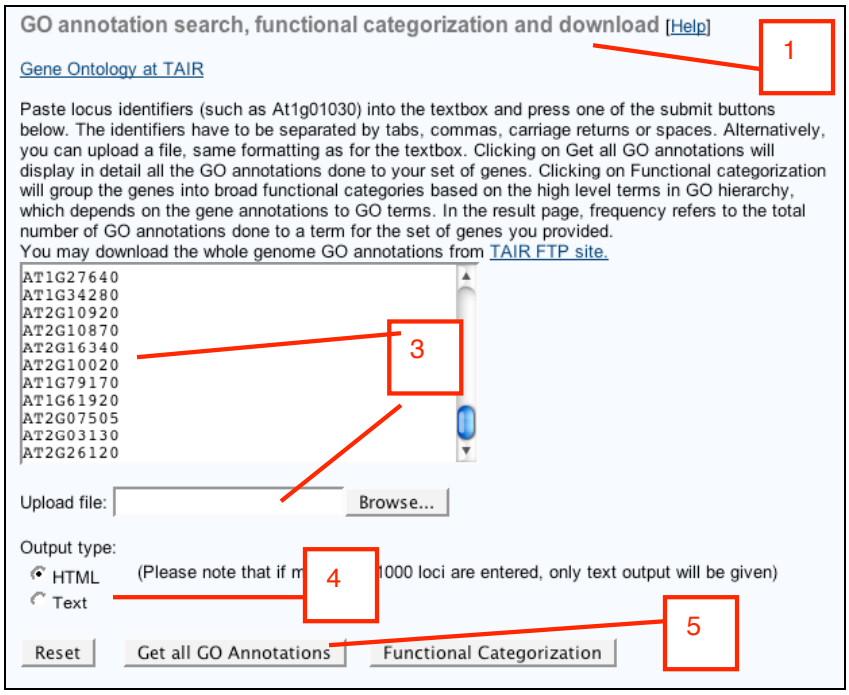
An introduction to microarray resources and tutorial can be found at:
HTTP://WWW.ARABIDOPSIS.ORG/HELP/TUTORIALS/MICRO_INTRO.JSP

Using TAIR's Gene Ontology resources to classify sets of clustered genes

The Gene Ontologies are controlled vocabularies that are used by many databases (including TAIR) for annotating the molecular function, biological roles and sub-cellular location of gene products. Annotations are made to specific (granular) terms which are in turn associated to more general terms (GOSlim).

Annotations for specific subsets of genes can be accessed through the GO annotation bulk download and analysis tool (http://www.arabidopsis.org/tools/bulk/go). The data can be downloaded as tab-delimited text files or as an HTML page with links to entries in TAIR and the Gene Ontology databases. The 'Functional Categorization' option can be used to classify sets of genes according to broad (GOSlim) categories which can in turn be displayed as a graphical pie chart.

Some of the uses of GO annotations for analyzing cluster data are to: 1) infer the functions of unknown genes in a cluster by evaluating the functions of known genes in the same cluster, 2) identify members of a cluster that may function in a similar pathway and may be co-regulated.

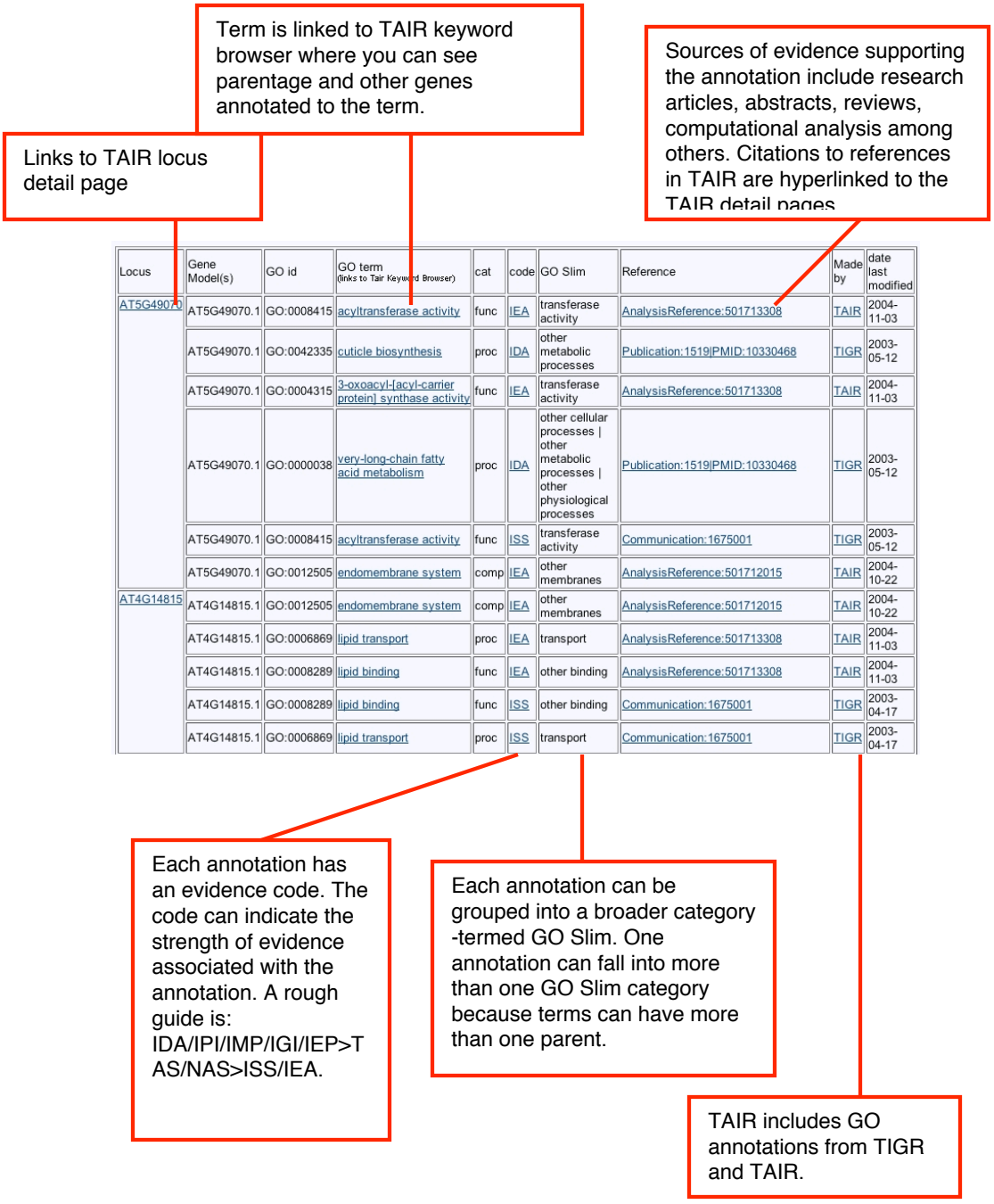


Obtaining GO annotations for a set of genes

1. Go to the GO Annotation Bulk Download tool. Type in the URL http://www.arabidopsis.org/tools/bulk/go/index.jsp or from the TAIR home page (http://www.arabidopsis.org) click on the link to "GO Annotation" in the Advanced Search section.
2. In a new window, open the sample data file (ftp://ftp.arabidopsis.org/home/tair/tmp/cluster_sample.txt).This file contains a list of 7 locus identifiers representing a cluster of genes identified from a microarray experiment. Alternatively you can try one of the larger cluster datasets (unk-cluster.txt) to see if it is possible to predict the functions of the unknown genes in this list.
3. Paste the locus names from the text file into the text input box in the GO Annotation download page.
4. Check the HTML output option.
5. Click on the button to Get All GO Annotations.

Using the GO Annotation results set

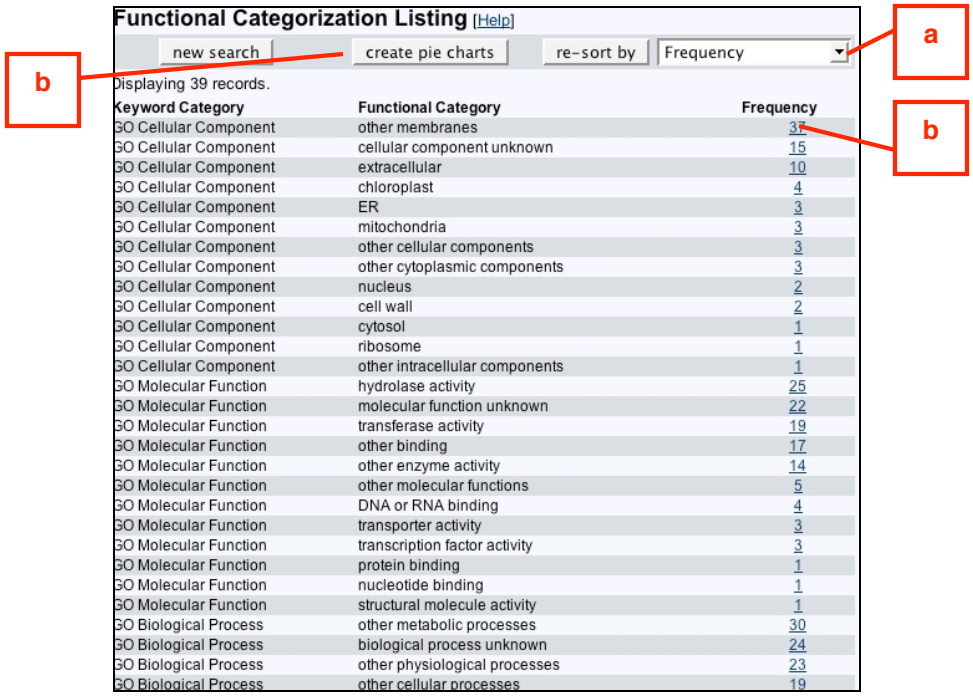
Choosing either the text or HTML option will return a list containing the following fields. The HTML (web page) includes hyperlinks to additional web pages that may be useful in analyzing and interpreting the results.



Locus	Gene Model(s)	GO id	GO term (links to Tair Keyword Browser)	cat	code	GO Slim	Reference	Made by	date last modified
AT5G49070	AT5G49070.1	GO:0008415	acyltransferase activity	func	IEA	transferase activity	AnalysisReference:501713308	TAIR	2004-11-03
	AT5G49070.1	GO:0042335	cuticle biosynthesis	proc	IDA	other metabolic processes	Publication:1519|PMID:10330468	TIGR	2003-05-12
	AT5G49070.1	GO:0004315	3-oxoacyl-[acyl-carrier protein] synthase activity	func	IEA	transferase activity	AnalysisReference:501713308	TAIR	2004-11-03
	AT5G49070.1	GO:0000038	very-long-chain fatty acid metabolism	proc	IDA	other cellular processes | other metabolic processes | other physiological processes	Publication:1519|PMID:10330468	TIGR	2003-05-12
	AT5G49070.1	GO:0008415	acyltransferase activity	func	ISS	transferase activity	Communication:1675001	TIGR	2003-05-12
	AT5G49070.1	GO:0012505	endomembrane system	comp	IEA	other membranes	AnalysisReference:501712015	TAIR	2004-10-22
AT4G14815	AT4G14815.1	GO:0012505	endomembrane system	comp	IEA	other membranes	AnalysisReference:501712015	TAIR	2004-10-22
	AT4G14815.1	GO:0006869	lipid transport	proc	IEA	transport	AnalysisReference:501713308	TAIR	2004-11-03
	AT4G14815.1	GO:0008289	lipid binding	func	IEA	other binding	AnalysisReference:501713308	TAIR	2004-11-03
	AT4G14815.1	GO:0008289	lipid binding	func	ISS	other binding	Communication:1675001	TIGR	2003-04-17
	AT4G14815.1	GO:0006869	lipid transport	proc	ISS	transport	Communication:1675001	TIGR	2003-04-17

Classifying the functions for a of a set of genes

1. Follow steps 1-4 of the previous protocol. Or use the browsers back button to go back to the filled out sample query page.
2. Click on the button labeled Functional Categorization.



Using the results

The Functional Categorization results are displayed on a table, which is first grouped by keyword category (type) and within each type, by functional category (GO slim term). Within each category the frequency for each bin is shown. The frequency corresponds to the number of times a given combination of GO term+gene appears in each category. To see a complete list of annotations to genes within a category, click on the number in the frequency column (a).

You can choose to re-sort the results to display similar GOSlim terms in adjacent rows by choosing ‘functional category ‘ from the drop down menu and then clicking the ‘re-sort by’ button (b). You can also choose to display the data in a graphical format as a pie chart (c).

Creating a pie chart showing the distribution of functional categories for a set of genes.

1. From the functional categorization results, click on the button to 'create pie charts'.

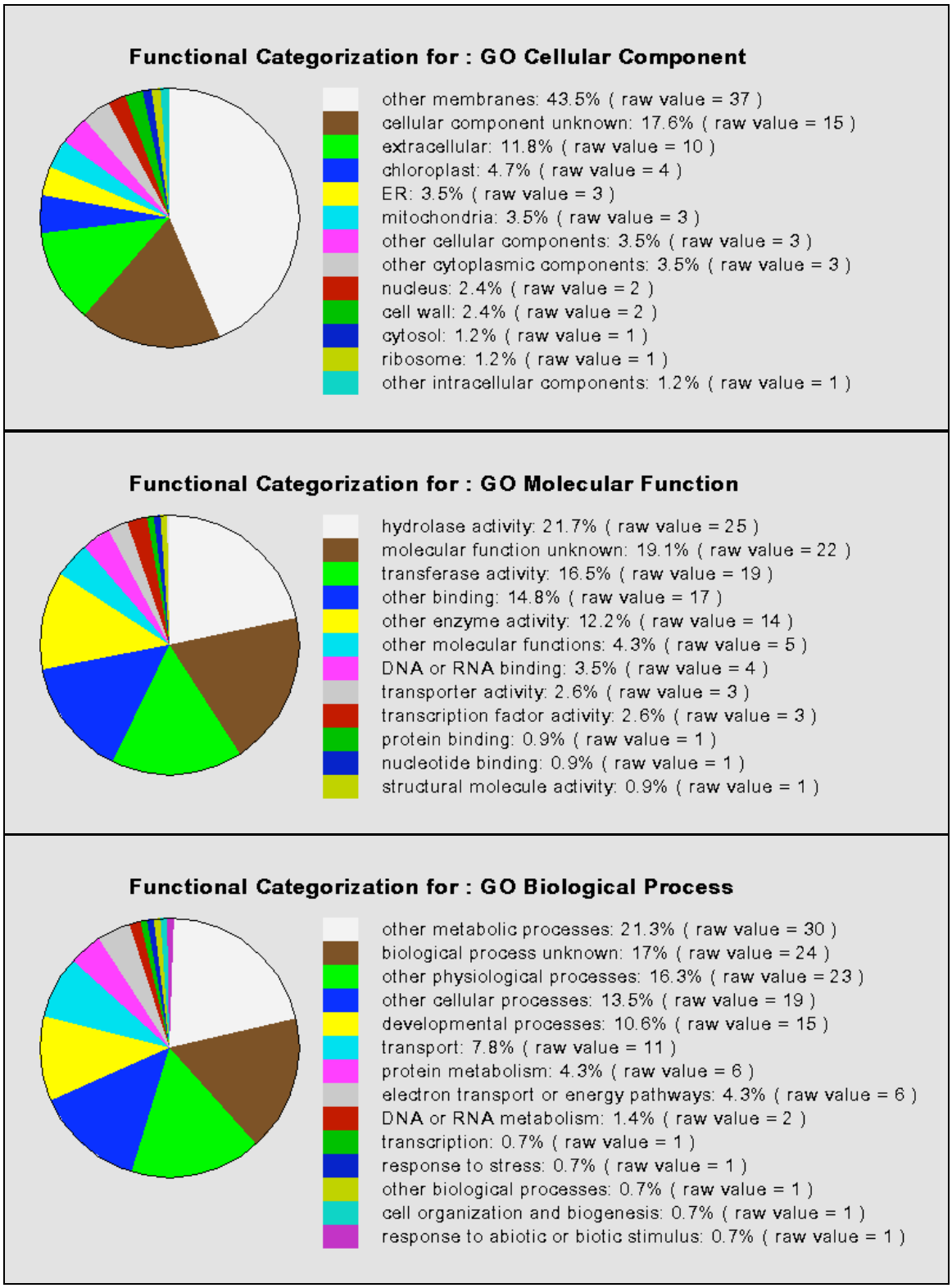
This will generate 3 pie charts, one for each aspect of the GO ontologies. Each segment of the graph is labeled with the category name, percentage of the total and the raw values for the number of annotations represented in the graph. Depending on how you have chosen to sort the results set, the pie chart display will order the segments of each pie either by frequency or category. If you want to show similar categories close together in your pie chart, sort the functional categorization results by category before making the pie chart.

2. To save each graph as a GIF, right click with your mouse (PC) or hold down the CTRL key (Mac) and save the image as a file onto your personal computer.

Once you have obtained a list of categories for your genes of interest, you may wish to compare the distribution of genes into functional categories in your cluster data relative to the distribution in the whole genome. You can obtain the entire set of GO annotations for Arabidopsis from our FTP site (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/).

Additional Resources

http://www.arabidopsis.org/help/tutorials/go_intro.jsp
http://www.geneontology.org



Pie chart for the dataset unk-cluster.txt. Each ontology aspect is represented in a single graph.

GO slim categories and their definitions.
Each table lists the GO slim categories for one of the three aspects of the GO. A GOSlim term MAY correspond to a single GO term or may not be a GO term at all. The multiple parentage of GO terms means that some genes may be included in more than one GO slim category. The complete table with hyperlinks to the corresponding terms in TAIR database is available at http://www.arabidopsis.org/help/helppages/go_slim_help.jsp

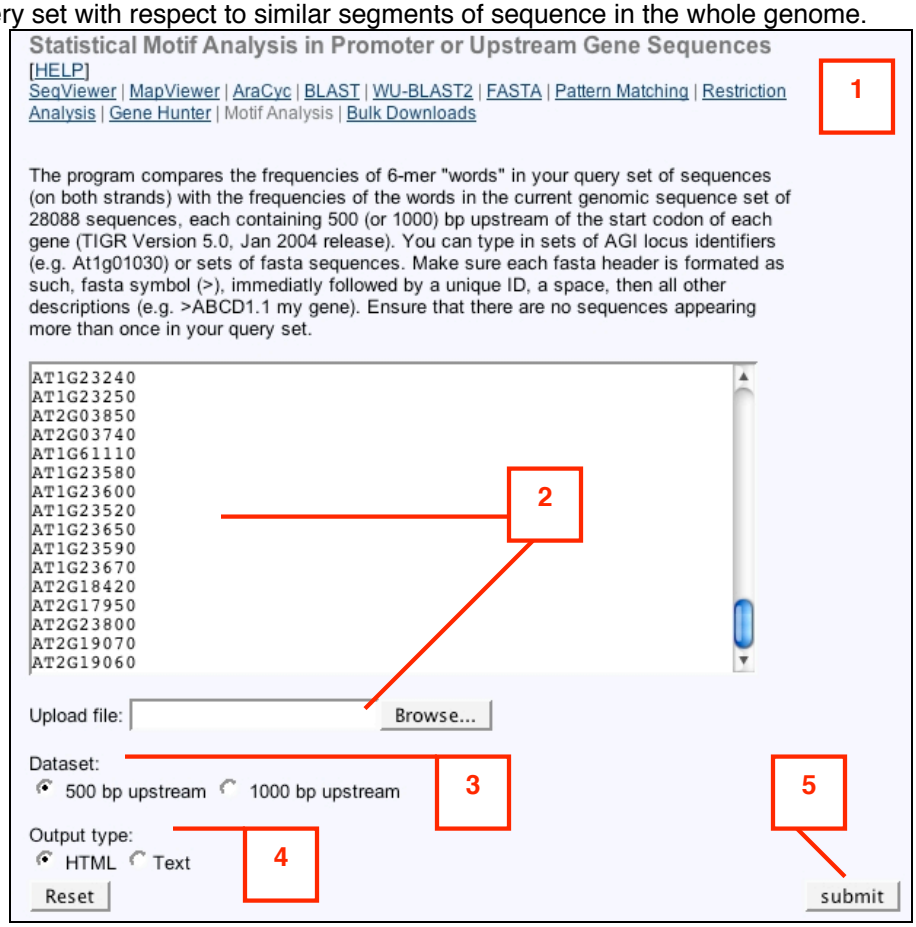
GO Molecular Function	GO Slim term	Definition
	hydrolase activity (GO:0016787)	Includes this term and all of its children
	kinase activity (GO:0016301)	Includes this term and all of its children
	transferase activity (GO:0016740)	Includes this term and all of its children
	other enzyme activity (GO:0003824)	Excludes hydrolase, kinase and transferase activities
	transcription factor activity (GO:0003700)	Includes this term and all of its children
	DNA or RNA binding	Includes DNA binding GO:0003677 or RNA binding GO:0003723 and excludes transcription factor activity GO:0003700
	other nucleic acid binding (GO:0003676)	Excludes DNA binding GO:0003677, RNA binding GO:0003723 and transcription factor activity GO:0003700
	nucleotide binding (GO:0000166)	Includes this term and all of its children
	protein binding (GO:0005515)	Includes this term and all of its children
	receptor binding and activity	Includes receptor binding GO:0005102 or receptor activity GO:0004872 and all of their children
	other binding (GO:0005488)	Excludes nucleic acid binding (GO:0003676),nucleotide binding (GO:0000166),DNA binding GO:0003677, RNA binding GO:0003723,transcription factor activity GO:0003700, protein binding (GO:0005515), receptor binding GO:0005102, receptor activity GO:0004872
	structural molecule activity (GO:0005198)	includes this term and all of its children terms
	transporter activity (GO:0005215)	Includes this term and all of its children
	molecular function unknown (GO:0005554)	Genes for which the function is not known or cannot be inferred
	other molecular functions (GO:0003674)	Excludes all of the other Molecular function GO slim categories

GO Biological Process	GO Slim Term	Includes/excludes
	biological_process unknown (GO:0000004)	Genes for which the process is not known or cannot be inferred
	developmental processes(GO:0007275)	Includes this term and all of its children
	transport (GO:0006810)	Includes this term and all of its children
	signal transduction (GO:0007165)	Includes this term and all of its children
	cell organization and biogenesis (GO:0016043)	Includes this term and all of its children
	other cellular processes (GO:0009987)	Includes DNA metabolism GO:0006259 or RNA metabolism GO:0006403
	protein metabolism GO:0019538	Includes this term and all of its children
	electron transport and energy pathways	Includes electron transport GO:0006118 or energy pathways GO:0006091
	transcription GO:0006350	Includes this term and all of its children
	other metabolic processes GO:0008152	Excludes protein metabolism GO:0019538,DNA metabolism GO:0006259, RNA metabolism GO:0006403, electron transport GO:0006118, energy pathways GO:0006091, transcription GO:0006350.
	response to abiotic and biotic stimulus	Includes response to abiotic stimulus (GO:0009628) and response to biotic stimulus (GO:0009607)
	response to other stresses (GO:0006950)	Excludes everything that is a child of response to abiotic stimulus or response to biotic stimulus.
	other physiological processes GO:0007582	Excludes response to abiotic stimulus (GO:0009628), response to biotic stimulus (GO:0009607), response to stress (GO:0006950), transport (GO:0006810), cell organization and biogenesis (GO:0016043), and other metabolic processes

GO Cellular Component	GO Slim term	Definition
	mitochondrion (GO:0005739)	Includes this term and all of its children
	chloroplast (GO:0009507)	Includes this term and all of its children
	plastid (GO:0009536)	Includes this term and all of its children
	ribosome (GO:0005840)	Includes this term and all of its children
	cytosol (GO:0005829)	Includes this term and all of its children
	endoplasmic reticulum (GO:0005829)	Includes this term and all of its children
	Golgi apparatus (GO:0005794)	Includes this term and all of its children
	other cytoplasmic components (GO:0005737)	Excludes, mitochondrion (GO:0005739), plastid (GO:0009536), ribosome (GO:0005840), cytosol (GO:0005829),endoplasmic reticulum (GO:0005829) and Golgi apparatus (GO:0005794).
	nucleus (GO:0005634)	Includes this term and all of its children
	other intracellular components (GO:0005622)	Includes this term and all of its children
	plasma membrane (GO:0005886)	Includes this term and all of its children
	other membranes (GO:0016020)	Excludes plasma membrane (GO:0005886)
	unknown cellular component (GO:0008372)	Used when the sub-cellular localization is not known or cannot be inferred
	extracellular (GO:0005576)	Includes this term and all of its children
	cell wall (GO:0005618)	Includes this term and all of its children
	other cellular components (GO:0005575)	Excludes all of the other cellular component GO slim terms.

Using the motif finder for identifying putative cis-regulatory elements

The Motif Finder was developed for the Arabidopsis Functional Genomics Consortium (AFGC). It searches for sixmer oligos in a set of query sequences and finds those that are over represented in the query set with respect to similar segments of sequence in the whole genome.

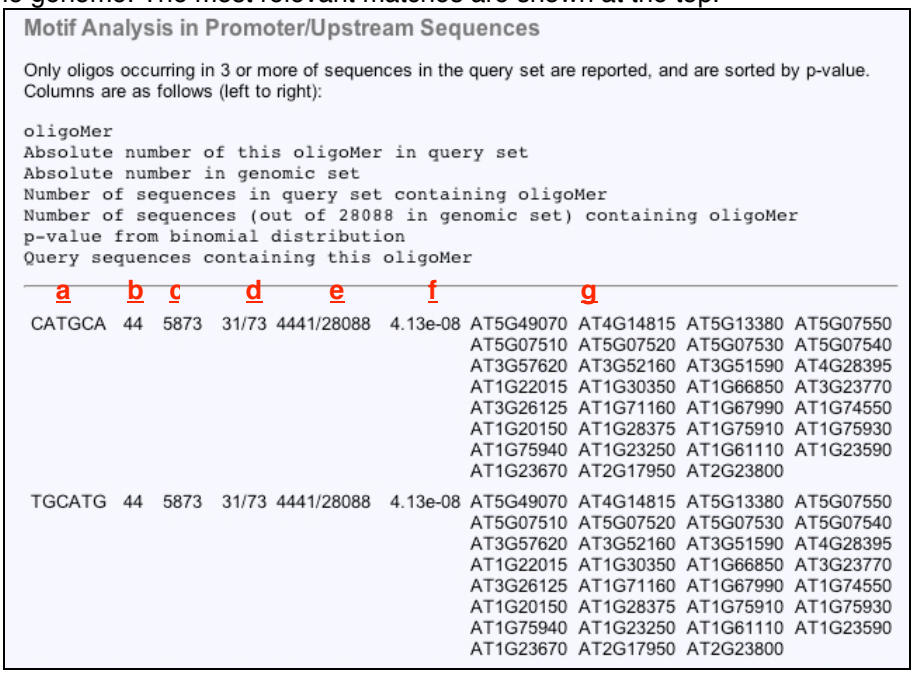


Performing a search for motifs

1. Open the Motif Finder page in your browser. From the TAIR home page, locate the section titled 'Tools' and click on the link to 'Motif Analysis' Or, enter the following URL: (http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp).
2. In a new window or tab, locate and open either the sample files used for the GO annotation exercise (http://www.arabidopsis.org/help/tutorials/unk-cluster.txt OR cluster_sample.txt) or use a set of FASTA formatted sequences (http://www.arabidopsis.org/help/tutorials/cis_fasta_sample.txt).
3. Copy the file contents into the text input box in the Motif Finder page.
4. Select the 500 bp upstream sequence dataset.
5. Choose the HTML option. Alternatively you can choose the text option if you want to save the results as a tab-delimited text file onto your personal computer.
6. Click on the 'Submit' button.

Evaluating the results

The results page displays a list the sixmer sequences which were found in the query dataset that are over-represented in the set with respect to the same subset of sequences (500 bp upstream) in the whole genome. The most relevant matches are shown at the top.



Motif Analysis in Promoter/Upstream Sequences

Only oligos occurring in 3 or more of sequences in the query set are reported, and are sorted by p-value.
Columns are as follows (left to right):

```
oligoMer
Absolute number of this oligoMer in query set
Absolute number in genomic set
Number of sequences in query set containing oligoMer
Number of sequences (out of 28088 in genomic set) containing oligoMer
p-value from binomial distribution
Query sequences containing this oligoMer
```

a	b	c	d	e	f	g
CATGCA	44	5873	31/73	4441/28088	4.13e-08	AT5G49070 AT4G14815 AT5G13380 AT5G07550 AT5G07510 AT5G07520 AT5G07530 AT5G07540 AT3G57620 AT3G52160 AT3G51590 AT4G28395 AT1G22015 AT1G30350 AT1G66850 AT3G23770 AT3G26125 AT1G71160 AT1G67990 AT1G74550 AT1G20150 AT1G28375 AT1G75910 AT1G75930 AT1G75940 AT1G23250 AT1G61110 AT1G23590 AT1G23670 AT2G17950 AT2G23800
TGCATG	44	5873	31/73	4441/28088	4.13e-08	AT5G49070 AT4G14815 AT5G13380 AT5G07550 AT5G07510 AT5G07520 AT5G07530 AT5G07540 AT3G57620 AT3G52160 AT3G51590 AT4G28395 AT1G22015 AT1G30350 AT1G66850 AT3G23770 AT3G26125 AT1G71160 AT1G67990 AT1G74550 AT1G20150 AT1G28375 AT1G75910 AT1G75930 AT1G75940 AT1G23250 AT1G61110 AT1G23590 AT1G23670 AT2G17950 AT2G23800

Examine the first match displayed: CATGCA.

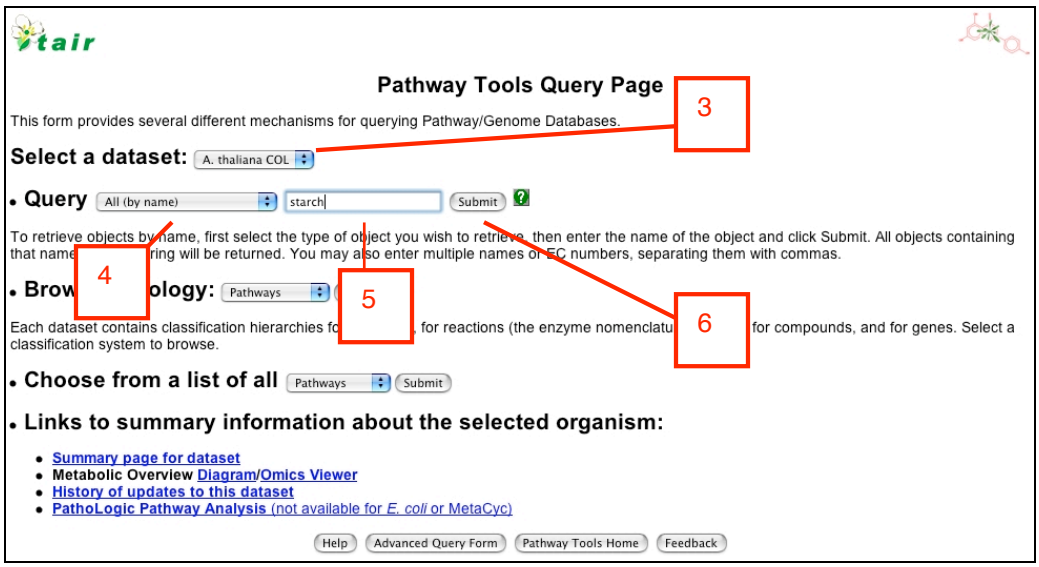
a. The first column of the results gives the sixmer sequence (CATGCA)
b. In the second column, the total number of times the sequence was found in your query set was 44. Since the total number of sequences was 73, this means that the sixmer occurs at least twice in some of the sequences.
c. The next column shows the total number of times the sequence appears in the 500 bp upstream sequence dataset (5873).
d. The fourth column is a ratio of the total number of sequences in the genome data set that contain the sixmer (4441) to the total number of sequences in the genome data set (in this case 28088).
e. The p_value score for this sixmer sequence is shown in the next column. Numbers closer to zero are better scores indicating that distribution of sixmer sequences is less likely to be random.
f. The last column is a list of the sequences in your query set that contains the sixmer sequence.

Additional Resources

For any potential cis-element experimentation is the obvious next step. You may want to see if the motifs you identified correspond to a previously described element by searching one of the many cis-element databases (http://www.arabidopsis.org/links/cis_element.jsp).The simplest way to view the location of the putative cis-element in its genomic context is to access the Nucleotide Sequence View for the locus in the SeqViewer (http://www.arabidopsis.org/servlets/sv). Once you have the nucleotide sequence view for a locus in front of you, use the 'Find' option in your browser to locate the sixmer sequence. If you want to find all of the upstream regions of the genome that contain the sixmer.
the PatMatch tool (http://www.arabidopsis.org) can be used to find short sequences in the genome and their relative coordinates. Both of these tools have on-line help documents.

Finding information about pathways, reactions,enzymes and compounds in AraCyc.

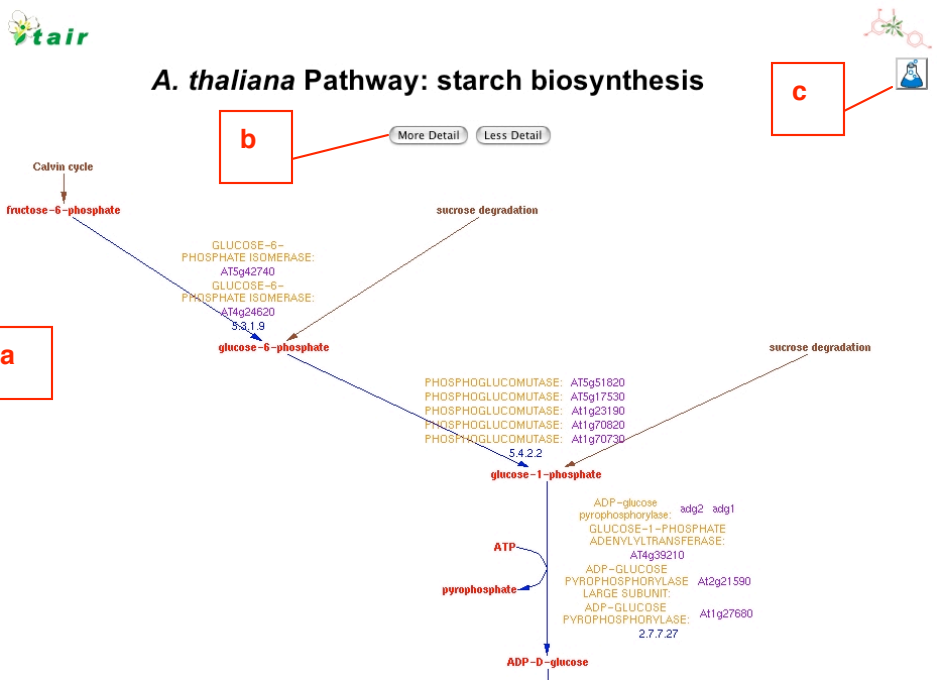
AraCyc is a database of metabolic pathways, enzymes, reactions, compounds and proteins. Pathways were initially computationally predicted and are also manually curated. See the AraCyc home page (http://www.arabidopsis.org/tools/aracyc) for a list of newly curated pathways, newly added pathways and predicted pathways that have been manually updated.



Searching for pathways by name

1. From your browser go to the AraCyc main page, type in the URL http://www.arabidopsis.org/tools/aracyc/ or from the TAIR home page (http://www.arabidopsis.org) find the link to AraCyc Pathways in the Tools section.
2. Click on the link to the Main Query Page (http://www.arabidopsis.org:1555/ARA/server.html).
3. Ensure that the selected dataset is set to A.thaliana-COL
4. In the query section, select ALL.
5. Enter the term ‘sucrose’ into the text input box for the query selection.
6. Click submit.

The results page will show a list of all of the objects in the database that include the term ‘starch in the name. The results are grouped by type and include proteins, compounds, pathways and reactions. Using the compound starch as a starting point, you can navigate through the different types of data in the database related to starch. Click on the starch biosynthesis pathway.

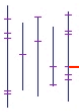


a

b

c

Locations of Mapped Genes:



d

Superclasses: Pathways -> Biosynthesis -> Sugars and Polysaccharides

e

Comment:

Starch is an α-glucan. Its counter part in animals is glycogen. There are two types of starch, amylose and amylopectin. Amylose contains up to several thousand α-glucosyl units linked almost exclusively in α(1->4) linkage with very few branches of α(1->6) linkage. Amylose accounts for 30% of starch. Amylopectin, on the other hand is a much more branched molecule and contains up to several million glucosyl residues. Amylopectin accounts for 70% of starch.

f

Starch is synthesized in plastids, including chloroplasts in photosynthetic tissues and amyloplasts in non-photosynthetic tissues such as seeds, roots, and tubers. Starch synthesized in chloroplasts of photosynthetic tissues is degraded to hexoses during the dark period. The derived hexoses are exported to the cytosol and used in sucrose synthesis. But the majority of hexoses derived from starch degradation are converted to triose phosphate which is then transported to the cytosol for sucrose synthesis. Sucrose can be readily transported to non-photosynthetic tissues to support plant growth or for starch synthesis in amyloplasts. The starch biosynthesis pathway depicted here includes both chloroplast and amyloplast pathways. The starting point for chloroplast pathway is fructose-6-phosphate, a product of photosynthetic carbon fixation. The starting point for amyloplast pathway is glucose-1-phosphate, a product of sucrose degradation. Studies from potato, pea, and maize indicate that glucose-6-phosphate, in addition to glucose-1-phosphate, can be imported into the amyloplast and can serve as the starting point for starch biosynthesis [Tauberger00] .

Citations: [Pa , Tauberger00 , Cathie95]

Unification Links: METACYC:PWY-622

g

Pathway Evidence Glyph:

h

Key to pathway glyph edge colors:

- green: reactions in which the enzyme is present in this organism
- black: reactions for which the enzyme is not identified in this organism
- orange: reactions unique to this pathway, and which, if an enzyme is identified, the enzyme is unique to this pathway
- magenta: reactions that are spontaneous, or edges that do not represent reactions at all (e.g. in polymerization pathways)

Using the AraCyc pathway detail page

a. The pathway detail page shows all of the reactions in blue, enzymes in yellow, genes in purple,compounds in red and related pathways are brown. Clicking on any of these will display the corresponding detail page from the database.
b. You can show more or less details of the pathway by using the zoom controls. At the highest zoom level all of the reactions and chemical structures for the compounds are shown.
c. The evidence icon indicates the type of evidence supporting the pathway. experimentally verified pathways have a flask icon. click on this icon to show the definition.

d. The genomic location of the genes encoding the enzymes is shown. mouse over the tick marks and click to show the name of the gene.

e. You can find related pathways by following the pathway hierarchy. click on the pathway super class to show a list of pathways in this class.

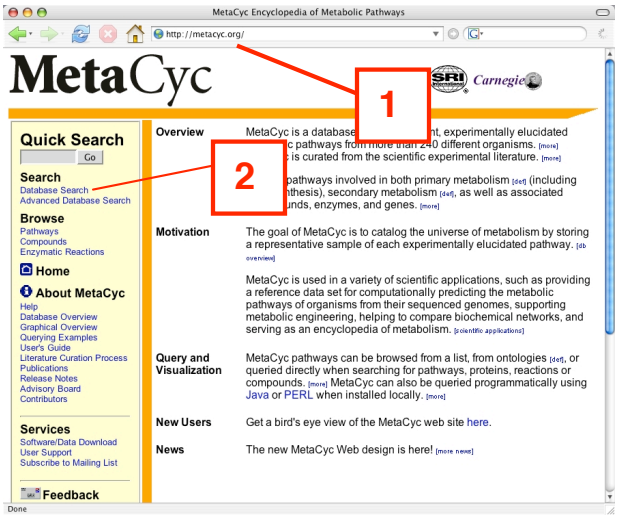
f. This section includes a summary of the pathway, and includes links to the papers used to curate the pathway information.

g. If the pathway is included in MetaCyc (e.g. if it is experimentally determined), you can use this link to see the pathway entry in MetaCyc (and find related pathways).

h. The pathway evidence glyph indicates the evidence supporting each of the reactions shown in the pathway.

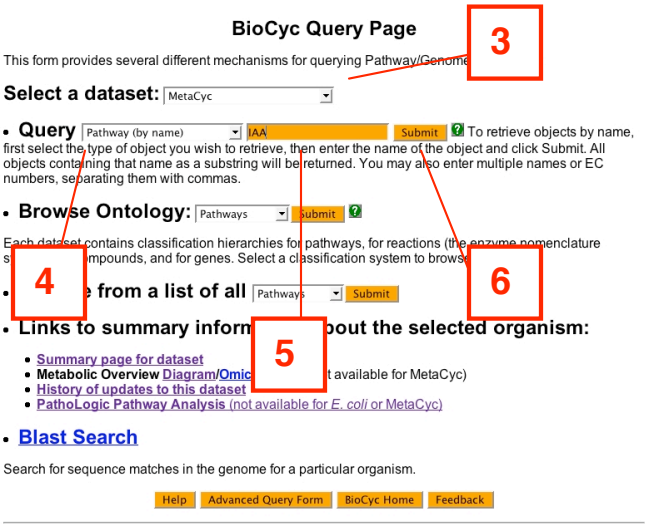
Finding pathways, reactions, enzymes and compounds in MetaCyc

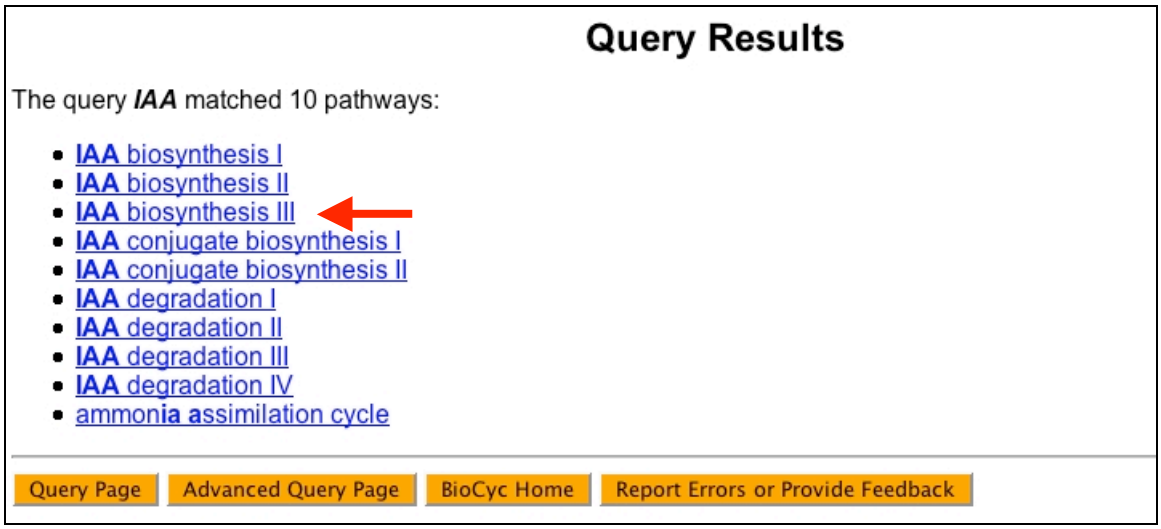
You can find variations of pathways from different organisms using MetaCyc. The query tools and displays are identical to AraCyc, however, there is also an optional quick search box. MetaCyc contains ONLY CURATED pathway information from a plants, humans and microbes. This tool can be used to compare pathways in different organisms.



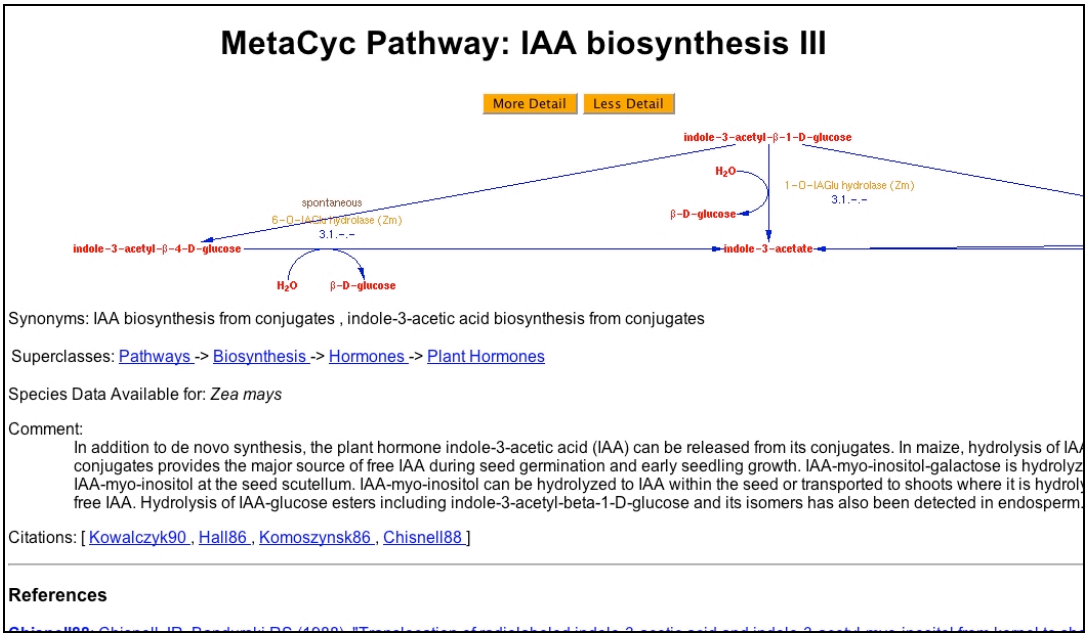
Performing a search for pathways

1. Go to the MetaCyc home page – enter the URL: http://metacyc.org/
2. Select Database search from menu.
3. Ensure the selected dataset is MetaCyc.
4. Choose query by pathway name.
5. Enter IAA.
6. Click submit.





Ten pathways that contain the term IAA will be shown in the results. Each variation of a pathway has a roman numeral appended to the name. For example, there are three variations of the IAA biosynthetic pathway (I,II and III). You can click on the link to each of the pathways and open them in separate windows to compare the pathways to each other.



MetaCyc Pathway details for variation III if the IAA biosynthetic pathway.
The pathway data came from maize. The MetaCyc detail pages are identical to the ones in AraCyc.

Displaying expression or other large scale data using the 'Omics' viewer.

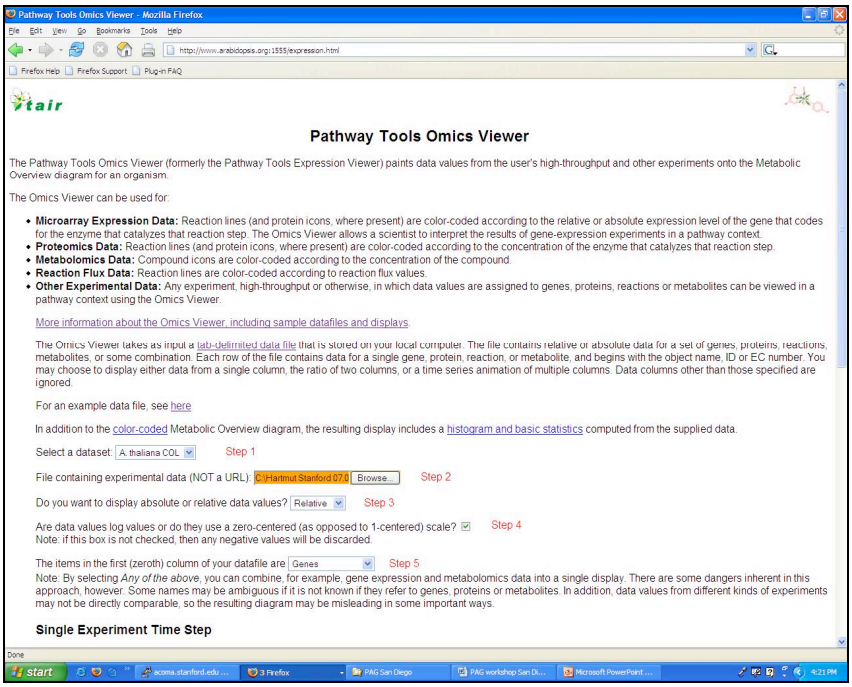
Exercise 1: Displaying microarray expression data.

Sample files for this exercise

We have prepared two sample files for you to use in this exercises. Both files can be accessed from the TAIR FTP site (ftp://ftp.arabidopsis.org/home/tair/tmp/).

ExpressionSample.txt: This file contains analyzed data from a cDNA microarray experiment which assayed gene expression at several time points. The first column (column zero) has a list of Arabidopsis locus names. The remaining columns are the log ratio normalized values in terms of fold change for each time point in the experiment.

ExpressionMetabolomicsSample.txt: This file contains all of the data included in the first file and additional measurements for compound concentrations. The compound names are included in the first column (column zero) along with the locus names. NOTE: the metabolomics data supplied is only for illustrative purposes and does not correspond to any experimental dataset.



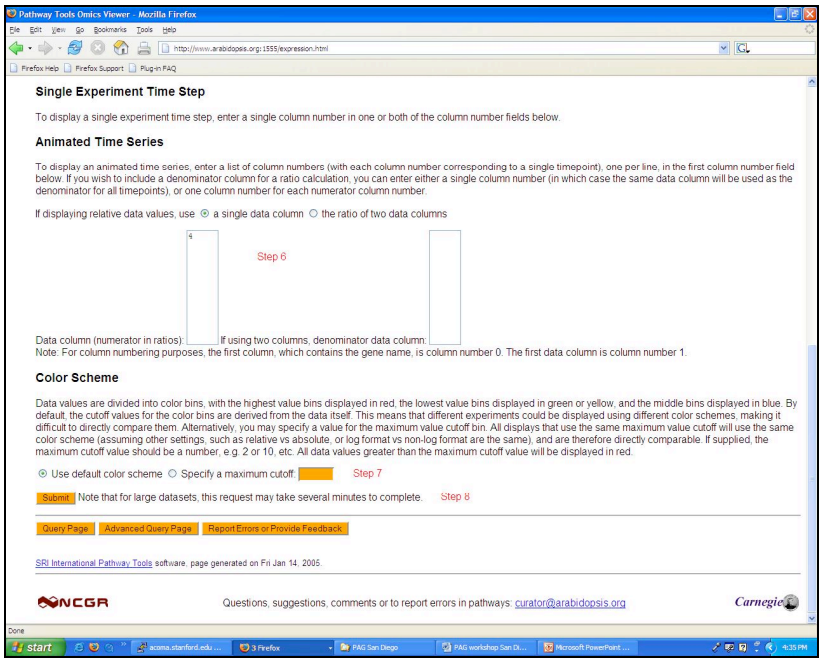
1. Select an organism database-A.thaliana COL

2. Upload your data file from your personal computer – the file must be prepared in a tab-delimited format, use the first example data set listed above (ExpressionSample.txt)

3. Set your data values (absolute or relative)

4. Check the box to display all data values, including the negative ones

5. Choose the type of data you would like to display (genes, compounds …. all of the above)



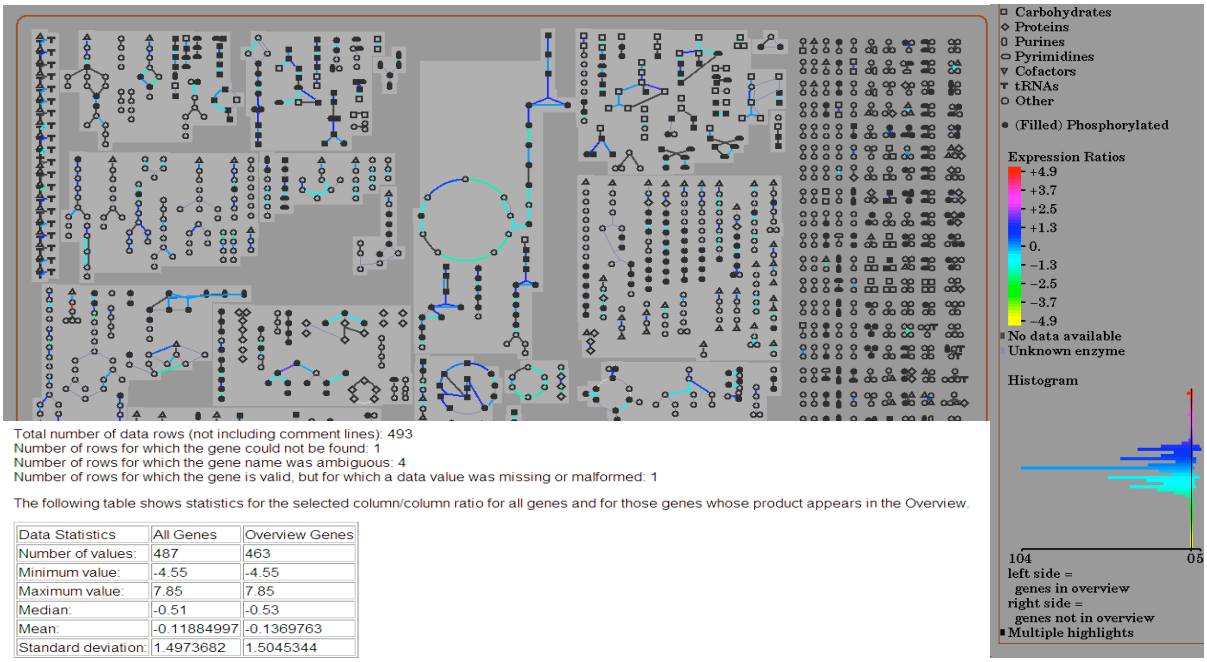
6. You can set up a single time experiment or an animated time series, the latter provides access to either displaying a time series or a comparison/ratio of chosen time points to each other. According to your dataset assign the columns and number of data points you would like to see.

7. Choose the color scheme expressing the values of your data either by default or a certain cutoff. The specified cutoff is used when you want to compare different expression experiments maintaining the given color-coding or when only one or very few genes are dramatically over expressed, which will reduce the spreading (and color-coded visibility) of the other genes in the experiment.

8. Submit your data

The metabolic map will show you, according to your data set, to what amount genes, compounds and pathways are expressed or have changed over time. You can display the pathway in question from the AraCyc database, if you click on the compound and you can see through the color setting what particular genes, compounds and reactions have been influenced by your experiment in the bigger picture of the overall *A. thaliana* metabolism. At the bottom of this page you will find some statistics (only for single time experiments), with details about the genes of your experiment expressed in the metabolic map of *Arabidopsis.*

The statistics will also list all the genes which could either not be found (e.g. not assigned to the metabolism of the map), genes which are ambiguous or genes with missing of malformed data (e.g. no expression value assigned to the gene).



Total number of data rows (not including comment lines): 493
Number of rows for which the gene could not be found: 1
Number of rows for which the gene name was ambiguous: 4
Number of rows for which the gene is valid, but for which a data value was missing or malformed: 1

The following table shows statistics for the selected column/column ratio for all genes and for those genes whose product appears in the Overview.

Data Statistics	All Genes	Overview Genes
Number of values:	487	463
Minimum value:	-4.55	-4.55
Maximum value:	7.85	7.85
Median:	-0.51	-0.53
Mean:	-0.11884997	-0.1369763
Standard deviation:	1.4973682	1.5045344

Objects that could not be found: At2g18440
Ambiguous object names: is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0 is ambiguous for G1F4, G1F0
Objects with missing or malformed data: At1g06460

Additional Resources

AraCyc tutorials (Quicktime movies)
Metabolic map tutorial
http://www.arabidopsis.org/help/tutorials/aracycmap.mov
Omics viewer tutorial
http://www.arabidopsis.org/help/tutorials/aracycexpr.mov
MetaCyc users guide
http://metacyc.org/MetaCycUserGuide.shtml
MetaCyc tutorials
http://metacyc.org/MetaCycExamples1.shtml

Using SeqViewer: TAIR's Arabidopsis Genome Browser/