Using The *Arabidopsis* Information Resource (TAIR) to Find Information About *Arabidopsis* Genes

Philippe Lamesch,[1] Kate Dreher,[1] David Swarbreck,[1] Rajkumar Sasidharan,[1] Leonore Reiser,[1] and Eva Huala[1]

[1]Carnegie Institution for Science, Stanford, California

ABSTRACT

The *Arabidopsis* Information Resource (TAIR; *http://arabidopsis.org*) is a comprehensive Web resource of *Arabidopsis* biology for plant scientists. TAIR curates and integrates information about genes, proteins, gene function, gene expression, mutant phenotypes, biological materials such as clones and seed stocks, genetic markers, genetic and physical maps, biochemical pathways, genome organization, images of mutant plants, protein sub-cellular localizations, publications, and the research community. The various data types are extensively interconnected and can be accessed through a variety of Web-based search and display tools. This unit primarily focuses on some basic methods for searching, browsing, visualizing, and analyzing information about *Arabidopsis* genes and describes several new tools such as a new TAIR genome browser (GBrowse), and the TAIR synteny viewer (GBrowse_syn). We also describe how to use AraCyc for mining plant metabolic pathways. *Curr. Protoc. Bioinform.* 30:1.11.1-1.11.51. © 2010 by John Wiley & Sons, Inc.

Keywords: Arabidopsis • databases • bioinformatics • data mining • genomics

INTRODUCTION

The *Arabidopsis* Information Resource (TAIR; *http://arabidopsis.org*) is a comprehensive Web resource for the biology of *Arabidopsis thaliana* (Huala et al., 2001; Garcia-Hernandez et al., 2002; Rhee et al., 2003; Weems et al., 2004; Swarbreck et al., 2008). The TAIR database contains information about genes, proteins, gene expression, mutant phenotypes, germplasms, clones, genetic markers, genetic and physical maps, biochemical pathways, genome organization, publications, and the research community. In addition, seed and DNA stocks from the *Arabidopsis* Biological Resource Center (ABRC) are integrated with genomic data, and can be ordered through TAIR.

The database content and other information relevant to plant scientists can be accessed through dynamic Web interfaces and static hypertext (HTML) pages. Users can perform simple searches of much of the database using names or keywords. Advanced search forms for different data types are used for more complex or specialized queries. Genomic data can be accessed through text-based queries, via the graphical genome browsers (SeqViewer; GBrowse, see *UNIT 9.9*), and with a variety of sequence similarity tools such as BLAST (see *UNITS 3.3, 3.4 & 3.11*) and FASTA (see *UNIT 3.9*). Data from TAIR can also be obtained in bulk from selected query tools and downloaded via file transfer protocol (FTP) from the Web site. TAIR has an extensive network of links from the database and Web site to other sources of *Arabidopsis* genomic data around the world.

TAIR is a curated database; data are processed by Ph.D.-level plant biologists who ensure their accuracy. Curation adds value to the large-scale genomic data by incorporating information from diverse sources and making accurate associations between related data. Data from manual literature curation, such as protein localization, biochemical function,

Current Protocols in Bioinformatics 1.11.1-1.11.51, June 2010
Published online June 2010 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471250953.bi0111s30

gene expression, and phenotypes, are added to the corpus of knowledge presented for each locus in the genome.

For *Arabidopsis* to be effectively used as a reference plant species, it is essential that researchers know what data are available and how to use the information they obtain. This unit includes several basic protocols for accessing the wealth of information about *Arabidopsis* genes that has been generated by the research community and made available through TAIR. The types of data and tools at TAIR are diverse and cannot all be described in one unit. Therefore, this unit focuses on the data and tools that are related to retrieving and mining information about genes. These protocols are based upon data and tools available as of April, 2010. As with any Web-based informatics resource, the data and tools will change over time.

The major data sections of TAIR are organized into eight categories, which appear on the navigation toolbar on all TAIR pages. Text-based query tools for performing simple and complex searches of specific types of data in TAIR, such as genes (see Basic Protocol 2), DNA, proteins, polymorphisms (including alleles), people, laboratories, and microarray experiments (see Basic Protocol 6) are found in the **Search** section. The **Browse** section allows the user, among others, to browse the ABRC stock catalog, the *Arabidopsis* transposon families, and the *Arabidopsis* gene families, as well as gene and plant ontology terms (see Basic Protocol 4). Within the **Tools** section are TAIR's graphical genome browsers (SeqViewer, GBrowse; see Basic Protocol 3), MapViewer for aligning physical and genetic maps, sequence similarity software (NCBI BLAST, WuBLAST, and FASTA), motif analyzer and patmatch (see Basic Protocol 7), the TAIR synteny viewer GBrowse_syn (see Commentary), the literature full-text search tool Textpresso (see Commentary), an *Arabidopsis* chromosome map tool (see Commentary), and the *Arabidopsis* metabolic pathway database AraCyc (see Basic Protocols 8 and 9), among other data analysis and visualization tools. Under the Tools section one will also find the TAIR bulk data retrieval tool for downloading sequences, protein data, and Gene Ontology assignments (see Basic Protocol 4 and Commentary) in bulk. The **Stocks** section contains links to the ABRC stock catalog (see Basic Protocol 5), DNA and germplasm searches (see Basic Protocol 5), and information about the stock center. The **Portal** section hosts pages with links to other databases and Web sites containing useful data and tools. The Portal also contains comprehensive lists of community resources generated by large-scale functional genomics projects, general information about *Arabidopsis* biology and history, and educational resources. The **Download** directory contains several logically organized directories containing large data sets related to gene, sequence, microarray, Gene Ontology and other data. All these files can be downloaded using the file transfer protocol (FTP). The **Submit** section contains forms and documentation for submitting data to TAIR. In the **News** section are links to the *Arabidopsis* community newsgroup, announcements from TAIR, meetings, and job postings.

BASIC PROTOCOL 1

TAIR HOMEPAGE, SITEMAP, AND NAVIGATION

The TAIR home page (*http://arabidopsis.org*) is the main entry point to the database and Web site (Fig. 1.11.1). To facilitate navigation of the TAIR Web site, a navigation toolbar is located at the top of all TAIR pages containing buttons such as Tools, Search, and Portals. When mousing over each item in the tool bar, a drop-down menu appears with clickable submenus that lead to a variety of datasets, tools, and external links. Several additional buttons are located above the main toolbar, including items such as Help, About Us, and Login. The Help section of the Web site (*http://arabidopsis.org/help/*) provides a quick guide to new users, frequently asked questions, a glossary of terms used on the Web site, tutorials, a search help function, and user guides for database searches, specific tools, and registration. Registered users can click on Login to order stocks and

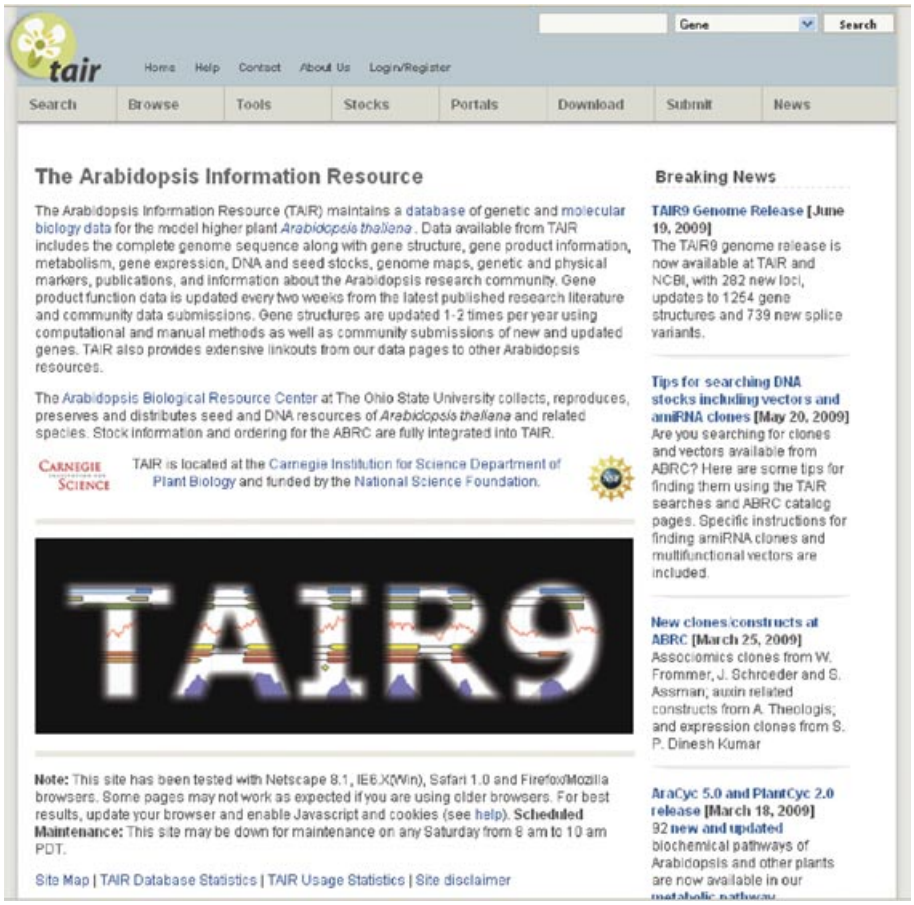

Figure 1.11.1 TAIR's home page (*http://arabidopsis.org*) is the main entry point to the database and Web site.

update personal information. The About Us section has information about the project, its goals, and its deliverables.

Necessary Resources

Hardware

Computer with Internet access

Software

Up-to-date Web browser. The browser must have cookies enabled to log in and process stock orders. TAIR makes extensive use of JavaScript; this feature must also be enabled. See *http://www.arabidopsis.org/help/index.jsp* for information on properly configuring one's browser.

Performing a quick search

1. Go to the TAIR home page (*http://www.arabidopsis.org*). Type the search term into the text box in the upper right corner of the page and choose a category from the drop-down menu (see Fig. 1.11.1). Click the Search button.

   The quick search performs a name search for most of the objects in the database (e.g., Genes, Clones, ESTs or BAC ends, People/Labs, Polymorphisms/Alleles, Germplasms, Ecotypes, Keywords, Genetic Markers, Proteins, Seed and DNA Stocks by stock name, and Vectors). By default, this is a "contains" search (a search for `aba1` *retrieves both ABA1 and ATRABA1A). It is also important to be aware that this search is not limited*

to the name field. For example, if the gene category is chosen, the gene description and keywords fields will be searched as well as the name. This is done to avoid missing any potentially relevant results, but sometimes too many results are returned. To perform an exact name search, choose the "exact name search" option from the drop-down menu to the right of the search box. This option will search the name field for all the data types listed in the drop-down menu. For additional search options, try the advanced searches listed under the Search header in the top navigation bar.

2. A list of all matching records is displayed for the data type chosen. Click on each record to access full details for that object, or download the current page of results using the download button at the top of the page. For gene search results, the additional option "download all" provides a way to download the entire result set at once, and "get all sequences" provides an option to download sequences for all the genes in the result set.

3. Alternatively, if the "Exact name search" option was chosen from the drop-down menu to the right of the box where the search term was typed in step 1, a list of data types with matching records and the number of records for each data type are displayed on a summary (TAIR Search Result) page. Click on any item in the list to display a summary of all the records retrieved for that data type. In this example, clicking on Proteins displays a list of the two ABA1 proteins encoded by different splice forms of the *ABA1* gene.

Performing a Web site search

4. In the event that a query returns no results, try either one of the Advanced Searches (see Basic Protocol 2) or the Web site search. To search the Web site, select "Google TAIR website" from the drop-down menu to the right of the box where the search term was typed in step 1 (see Fig. 1.11.1) and click the Quick Search button.

   The Web site search uses the Google search engine and can search for words as well as phrases. The Web site search looks for matches to each input word in the static Web pages. To force the search to treat the words as a phrase, use quotes. For example, `"disease resistance"` *will find all pages with the phrase "disease resistance" in the text. Left unquoted, the search engine finds all pages with the term disease OR resistance in the text.*

BASIC PROTOCOL 2

FINDING COMPREHENSIVE INFORMATION ABOUT *ARABIDOPSIS* GENES

The locus detail pages represent the most comprehensive starting point for a researcher interested in finding out what is known about a gene. The physical location of an annotated gene on the genome is called a locus in TAIR. The locus serves as a useful concept for grouping genes with other objects having the same genomic location. For convenience, genetically defined genes (i.e., those identified by linkage studies but which are not yet associated with a genomic sequence) are also included as loci that have a genetic, but no physical location. Each locus is associated with at least one gene model, which can be thought of as a version of a gene. Several gene models or splice-variants can be associated to a gene locus based on alternatively spliced mRNAs. The locus detail page collects information such as each model's exon and intron boundaries, experimentally determined or predicted function, gene expression data, mutant phenotypes, associated germplasms, polymorphisms, clones, and publications. Because data in TAIR are highly integrated, it is possible to access the locus detail page from almost every other detail page in the database. This protocol illustrates a commonly used way of finding genes using the Advanced Gene Search form.

Necessary Resources

See Basic Protocol 1

Searching for information about a specific gene or set of genes

1. Go to the TAIR home page (*http://www.arabidopsis.org*). In the top navigation bar click on the Search header (see Fig. 1.11.1) and select the Genes link to go to the TAIR Gene Search page (*http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene*).

2. To search by name, choose "Gene name" as the option from the Search Name drop-down menu (the options are "Gene name," "description," "phenotype," "GenBank accession," or "GenBank gi"). Using the drop-down menu to the right of this, set the search to an exact match or an inexact match (the options are "contains," "starts with," "ends with," or "exactly") and type the name in the text box on the right-hand side of the same line. For example, to find a set of related genes sharing a core symbolic name, such as ARF for Auxin Response Factor family members (Hagen and Guilfoyle, 2002), type in `ARF` as the name term and choose the "starts with" option to the left of this. Click the "submit query" button.

   Gene names include systematic names assigned based on chromosomal location (so called AGI locus identifiers such as AT1G01010) or gene symbols. For more information about Arabidopsis gene nomenclature, see the Arabidopsis Gene Nomenclature Guidelines (http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp).

3. All of the loci that match the query term will be displayed in a list of results (on a page titled TAIR Gene Search Results). Click on the locus name to view the locus detail page. A sample locus detail page obtained by using the search name `ABA1`, and then selecting the AT5G67030 locus from the TAIR Gene Search results page, is shown in Figure 1.11.2.

   The default search only retrieves genes that are active in the database. Checking the "include obsoleted genes" check box will retrieve both active and obsoleted genes, along with the history of their status in the database. Genes may become obsolete if they are merged with other genes—or if improved genome annotation methods find inadequate evidence for their existence. TAIR retains information about obsolete genes in order to maintain a record of their histories and associations.

Using the detail pages to find information about a locus

4. On a locus detail page (Fig. 1.11.2), data sections are displayed in alternating color bands; related data are generally grouped together. The following annotations (the red lettered items on the left side in Fig. 1.11.2) summarize the typical information displayed on a locus detail page. Definitions of each data type can be obtained by clicking on the question marks to display a pop-up definition window.

   a. Gene summary information (Fig. 1.11.2A, a items).

      *General information about the gene can be found in several places. The **Description** field is a short summary of the gene's function either manually composed by a curator or computationally generated. The latter is only shown if the locus has not yet been curated manually. The computation description contains the gene's full-name, GO and PO terms, best A. thaliana protein match, and the number of protein Blast hits in other species (NCBI BLink).**Other names** include all alternative forms of the gene name. The **Update history** indicates the changes made to the physical locus over time, such as being replaced or being made obsolete. The **Date last modified** provides a time stamp indicating when information about the locus was last added or updated.*

   b. Gene model information (Fig. 1.11.2A, b items).

      *The **Representative gene model** for a protein coding gene is the gene model with the longest coding sequence (CDS); for other gene types, the representative model is set as default to the .1 model.*

      *The **Map detail image** displays the exon-intron structures of all gene models of a locus. Clicking on the image directs the user to GBrowse.*

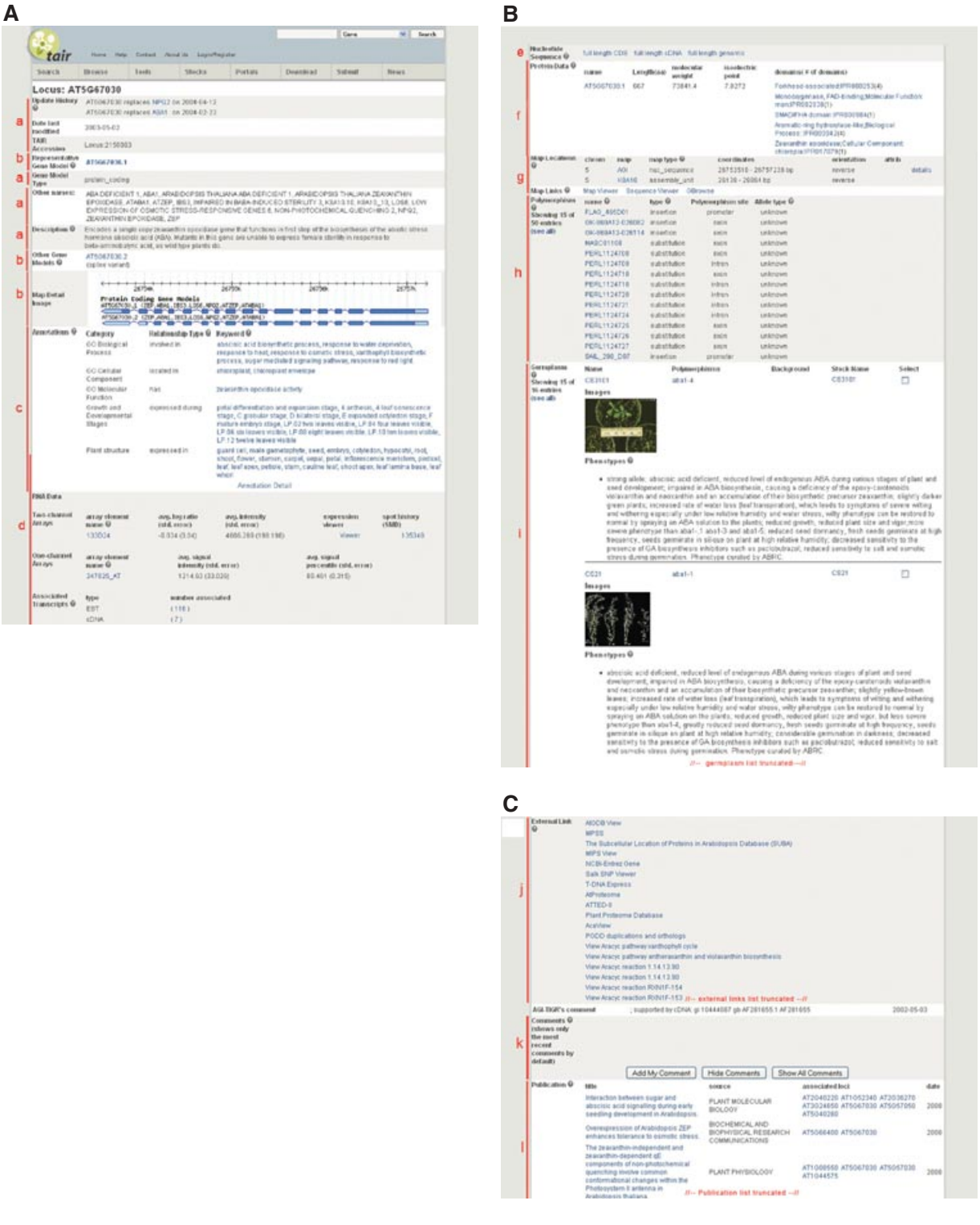

Figure 1.11.2 A sample of a locus page from TAIR showing the major data included in the detail page. A portion of the germplasm section has been deleted for simplicity. Each of the data types displayed in the alternating colored bands can be grouped into one or more the following categories: (**A**) (a) general descriptive locus information, (b) gene model information, (c) functional annotations, (d) gene expression data; (**B**) (e) nucleotide sequences, (f) protein data, (g) mapping data, (h) polymorphisms and alleles, (i) germplasm information; (**C**) (j) links to resources outside of TAIR, (k) comments about the locus, and (l) papers and abstracts.

c. Gene function, biological role, and localization (Fig. 1.11.2A, c item).

*The **Annotations** section contains all of the controlled vocabulary terms that describe the molecular function, biological role, subcellular localization, and expression of the gene product. The annotations are grouped according to the type of vocabulary and summarized on the locus page. Click on the **Annotation Detail** link (located at the bottom of the annotations section) to display the full annotation details, which include the type of evidence supporting the annotation and the corresponding references.*

d. Gene expression (Fig. 1.11.2A, d item).

*Information about the expression of the gene can be found in the **RNA Data** section and the lower part of the **Annotations section**. In the RNA Data section, array elements from one-channel and/or two-channel experiments that map to the locus are listed. Array element names are linked to detail pages where their expression behavior across a variety of experiments can be found (see Basic Protocol 6). For elements whose expression has been analyzed across all experiments, the average log ratio of expression values, along with standard error, are provided along with links to the Expression Viewer (for finding similarly expressed genes) and Spot History (only available for microarray elements from arrays in the Stanford Microarray Database). Lists of full-length cDNAs and expressed sequence tags (ESTs) can be found in the **Associated Transcripts** subsection within the RNA Data section. Click on the number next to the clone name to see a list of all the clone records. The clone records are linked to GenBank, where information about the cDNA libraries (and therefore expression) can be found. Finally, information about gene expression, curated from the literature, is shown in the Annotation band along with the Gene Ontology associations (see Fig. 1.11.2A, section "c": "expressed during", "expressed in").*

e. Nucleotide sequences (Fig. 1.11.2B, e item).

*Links to genomic sequence, full-length CDS, and full-length cDNA sequence are located in the **Nucleotide Sequence** section. Clicking on the sequence name will display a new window containing the sequence, which can be uploaded directly into TAIR's WU-BLAST tool. In addition to WU-BLAST, TAIR has two other local sequence alignment tools: NCBI-BLAST (Altschul et al., 1990; also see UNITS 3.3, 3.4 & 3.11) and FASTA (Pearson, 1995; see also UNIT 3.9). These tools use some specialized Arabidopsis sequence data sets such as T-DNA/transposon insertions, intergenic regions, upstream and downstream sequences, and UTRs (http://www.arabidopsis.org/help/helppages/BLAST_help.jsp#datasets). These tools can be accessed from the TAIR homepage under the **Tools** section.*

f. Protein data (Fig. 1.11.2B, f item).

Structural and physical characteristics of the protein encoded by the reference gene model, including molecular weight, conserved domains, and pI, are displayed in this section. Click on the AGI name in the protein section to open a new window displaying more detailed information and the amino acid sequence.

g. Map locations (Fig. 1.11.2B, g item).

*The **Map Locations** section displays the chromosome and coordinates of the locus for the maps on which it is found. The gene can be viewed in a whole-genome context by clicking on one of the three map options (Map Viewer, Sequence Viewer, GBrowse) in the Map Links section (See Basic Protocol 3).*

h. Alleles and polymorphisms (Fig. 1.11.2B, h item).

*All of the polymorphisms that map within the locus are shown in the **Polymorphisms** section, along with the type of variation. This section includes natural variations found in different ecotypes and induced mutations (e.g., T-DNA insertions) that have been mapped by sequence identity and alleles that have been curated from the literature. To find detailed information about a polymorphism, click on the name of the polymorphism.*

i. Germplasm information (Fig. 1.11.2B, i item).

*The **Germplasm** section provides information on all germplasms available for a locus, including phenotype descriptions and mutant images.*

j. External links (Fig. 1.11.2C, j item).

There are other Web sites that provide either alternate views or different information about a locus. In order to provide access to as much information about a locus as possible, TAIR has links to the corresponding locus pages in other databases and Web sites. Types of external links include other Arabidopsis genome annotation databases, gene expression databases, and functional genomics sites, as well as links to tools for further analysis. For example, all sequenced loci are linked to other Arabidopsis annotation databases including NCBI, MIPS, and AtGDB. TAIR also provides links to Uniprot and NCBI from the protein detail pages.

k. Comments (Fig. 1.11.2C, k item).

"Comments" contain additional data contributed by registered TAIR users, and are included in the display for nearly all of the TAIR detail pages. This function can be used to report new data, as well as errors or omissions related to the displayed object.

l. Publications (Fig. 1.11.2C, l item).

Papers and conference abstracts are shown at the bottom of the detail page in the section marked Publications. Publications include published literature imported from PubMed, Agricola, and BIOSIS, along with abstracts from the International Arabidopsis Meetings. Only the most recent ten papers are listed on the detail page; to retrieve the complete list, click on the Show All Publications link. Clicking on the title of the publication opens a new link to the detailed record where one can read the abstract, link to the PubMed citation, and find authors among TAIR's community.

Saving the results of a search to a file

5. Return to the list of results obtained by the query submitted in step 2 (page titled TAIR Gene Search Results). Check the box to the far left of the results summary. Each page of results must be saved separately. Only those results that are selected will be saved. Use the Check All function to save all of the results displayed on the page.

   Before downloading a large set of results, use the browser to go back to the Advanced Search page, make sure the number of records per page of results is set to the maximum (usually 200 records/page), and resubmit the query.

6. After selecting all of the desired results on a page, click on the Download Checked button (or Download All if you wish to export all results) in the upper right corner of the TAIR Gene Search Results page. The checked results will then be displayed in the browser window as tab-delimited text file. Use the Save As function under the File menu in the browser toolbar to save the results in a file on the local computer. This process must be repeated for each page of results.

7. In order to retrieve sequences for the selected results, click on the Get Checked Sequences button (or Get All Sequences if you wish to retrieve sequences for all results) on top of the TAIR Gene Search Results page. This will bring you to the Sequence Bulk Download and Analysis page from where you can retrieve different types of sequences for your list of genes. For more information about that tool, see the Commentary.

   The download feature is found on all of the search results pages. Each set of results includes different information in the downloadable file. See the help documents for the specific search to view a listing and description of the downloaded fields. The files contain tab-delimited text that can be opened using a text editor or spreadsheet software such as Microsoft Excel.

USING THE *ARABIDOPSIS* GENOME BROWSERS (SeqViewer AND GBrowse)

BASIC PROTOCOL 3

TAIR provides two alternative Web applications (SeqViewer and GBrowse) that allow users to explore the annotated *Arabidopsis* genome sequence. SeqViewer is a graphical genome browser developed by TAIR while GBrowse (Stein et al., 2002; see *UNIT 9.9*) was developed by the Generic Model Organism Database project (GMOD). Both tools allow the user to search for and display various sequence features such as genes,

polymorphisms, T-DNA insertions, and transcripts (ESTs/cDNAs), provide a mechanism for navigating around the genome, and allow individual users to customize the type of data displayed. As such, these tools are useful for a wide variety of tasks including positional cloning, identifying mutants in a gene of interest, finding cDNA and ESTs for a gene of interest, and finding and displaying the distribution of sequence features (e.g., polymorphisms, T-DNA insertions) in a whole-genome context. While both tools share some functionality, each tool has its own specific set of features. Additionally, GBrowse contains many data types not represented in SeqViewer.

Necessary Resources

See Basic Protocol 1

Exploring SeqViewer

Displaying a defined region of the genome

1. Go to the TAIR home page (*http://www.arabidopsis.org*). In the Tools drop-down menu of that page, click on the link to SeqViewer. Alternatively, go directly to the URL *http://www.arabidopsis.org/servlets/sv*.

   This will invoke the SeqViewer home page, which shows the five chromosomes of the nuclear genome sequence represented as five green bars, one for each chromosome. When using SeqViewer, it is a good idea to note the version number/date shown below the chromosome bars. Genome annotation changes over time; the versioning/time stamp provides a way of tracking annotations that may change or become obsolete.

2. To search by name, make sure that the "name" radio button is selected, then enter (by typing) or upload (using the Browse button) a file of up to 250 names into the text input box in the lower right section of the home page. For example, to search for the gene AT1G01040, enter `AT1G01040` in the text input box and choose "gene" as the name search option from the drop-down menu to the right of the "name" radio button. Submit the query by clicking the Submit button. The results of the search are shown in "whole-genome view" (Fig. 1.11.3). Matches to the genome are displayed as tick marks (red on screen) on the green chromosome bars in the whole-genome view.

3. The genome browser can also be searched using short sequences such as PCR primers used for genetic markers or highly conserved sequence motifs such as miRNA core binding sites. Paste in or upload up to four nucleotide sequences (each between 15 and 150 nucleotides long) in FASTA format and choose the radio button to search by "sequence." The sequence search finds only exact sequence matches; ambiguous matches are not allowed. Hits to the genome are displayed as red tick marks on the whole-genome view and in a special Match track in the Close-up view (see below).

Displaying a close-up view of a genomic region

4. To display an enlarged view of the genomic region centered on the gene or other object found in the search, click on the red tick mark corresponding to the match. This opens up a Close-up view of a 200-kb region, approximately centered on the gene highlighted in yellow (and similar to the view shown in Fig. 1.11.4).

   Click on any object to retrieve its detail page from the database. Mousing over the data will display a brief summary in a small pop-up window.

   In the example shown in Figure 1.11.4, the highlighted locus is displayed at a 10-kb resolution and centered in view. To obtain a similar view using the Close-up view controller, first zoom to 10 kb by selecting this option from the drop down-menu (shown in the close up view section of Fig. 1.11.4), then enter the name of the locus in the text entry box next to the Find button then click on the Find button (below the "Zoom to" controls in Fig. 1.11.4). The display will now show a 10-kb window centered on the selected locus, which is highlighted in yellow, as shown in Figure 1.11.4.

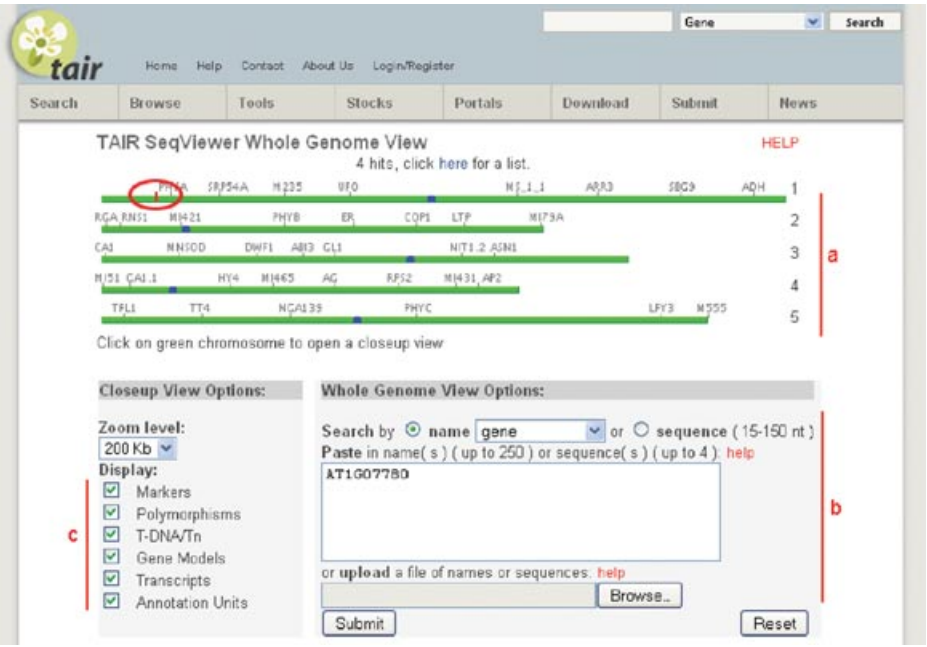

Figure 1.11.3 SeqViewer home page after submitting the gene name AT1G07780 as a query term. The five nuclear chromosomes are shown as green lines with blue boxes indicating the location of the centromeres (a), a few markers are included as landmarks for orientation. Queries can be typed, pasted, or uploaded into the text input box (b). The available options include searches by name or sequence. The number of matches is displayed above the chromosomes (in this example this number is 4) and is hyperlinked to a list of results. Each match to the genome is indicated with a red tick mark on the chromosomes; clicking on the mark will open a detailed Close-up view. The Close-up View options (c) are used to select the zoom level and types of objects to display in the detailed view. For the color version of this figure go to *http://currentprotocols.com/protocol/bi0111*.

Alternatively, use the cursor to center the view. Move the cursor along the centering bar at the top of the Close-up view between the left and right scroll arrows (letter "c" in Fig. 1.11.4). A yellow bar will appear above the cursor, indicating which region to select. When the yellow bar is over the desired region, click once in the centering box.

5. Create a custom view of any region of the genome by clicking on the appropriate chromosome in the whole-genome view (i.e., the screen illustrated in Fig. 1.11.3). A new Close-up display will appear centered on the selected region of the chromosome. In the Close-up view control panel (Fig. 1.11.4) enter the left and right coordinates into the Select Range box at the left of the screen and click the Go button.

   Each chromosome corresponds to a pseudomolecule that is a composite of all linked BAC sequences in the genome tiling path. BAC sequences may be trimmed or extended in regions of overlap to ensure a continuous sequence. Coordinates for each base pair are indicated by the following convention: numbers start from the top of the upper chromosome arm (to the left of the centromere on the SeqViewer Whole Genome View) and end at the bottom of the lower chromosome arm. When selecting coordinates to input for generating a custom view, the quickest way to find coordinates for an object is to obtain this data from the detail page. For example, to create a custom view between two loci, go to each locus detail page and find the coordinates shown for the AGI map in the Map Locations band. Pay close attention to the orientation of the locus: if the two objects are on opposite strands, the starting coordinate of one sequence will be flipped, relative to the other.

   The custom view feature is useful for positional cloning. A genetically defined region between two markers can be displayed to find new markers or polymorphisms in a region for fine mapping. Alternatively, when searching for candidate genes within the interval, a downloadable summary of the genes located in the displayed region can be obtained by clicking on the List Genes in Range button.

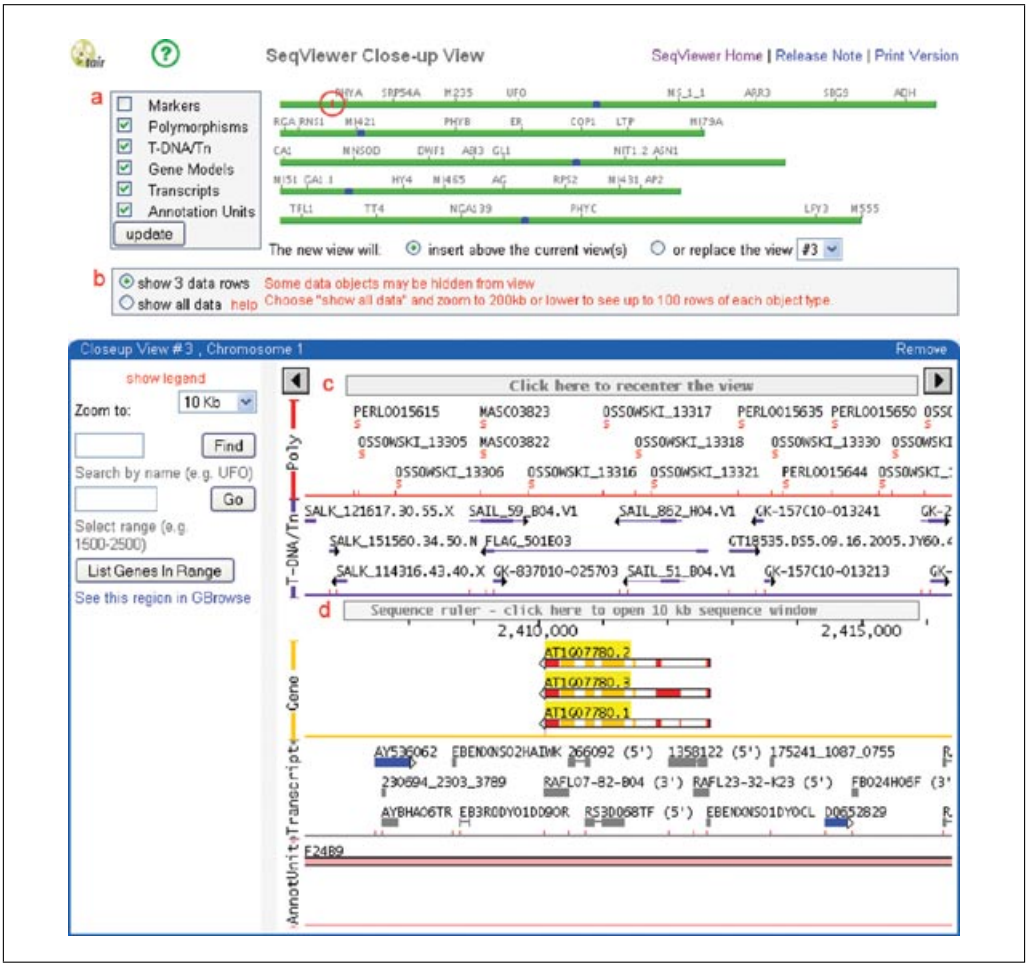

Figure 1.11.4 A 10-kb region of chromosome 1 centered on the AT1G07780 locus, which is highlighted in yellow. (a) The area of the genome shown in the Close-up view is indicated by the numbered box in the whole genome view. (b) The radio button for selecting three or all rows of data to display in the Close-up view. (c) The gray re-centering bar. (d) The gray bar between the T-DNA and gene bands is used for selecting a 10-kb region to display in the nucleotide sequence view. For the color version of this figure go to *http://currentprotocols.com/protocol/bi0111*.

Displaying selected bands of data in the Close-up view

6. The global controller on the left-hand side of the whole-genome view page can be used to select sequence features to display (letter "c" in Fig. 1.11.3). Each sequence feature (genes, transcripts, polymorphisms, T-DNA/Transposons, genetic markers, and annotation units) can be removed or added to the display in the Close-up view by checking or unchecking the box next to the feature name. These check boxes are also found on the Close-up view page, and therefore can be used before or after zooming in to a Close-up view (Fig. 1.11.4); the selection will affect all open Close-up views. To see a complete explanation of the bands and graphics used in the display, click "show legend" located just above the "zoom to" box in the Close-up view control panel (left side of Fig. 1.11.4).

Displaying all rows of data in the Close-up view

7. In order to simplify the display, the default Close-up view shows only three rows of data for all sequence features. Items not displayed are indicated by black tick marks; displayed items are indicated with red tick marks. To display all of the data for each sequence feature in the Close-up view, zoom to between 10 kb and 200 kb and click the radio button to display all data in the box below the green chromosome bars (indicated by the letter "b" in Fig. 1.11.4). This will expand the rows for all of the selected

data types to show all objects in the Close-up view. Before showing all rows it is a good idea to zoom in to a fairly high level of resolution, as large amounts of data can result in a very long page. Zoom levels of 1 Mb and higher will only display a maximum of six rows for each type of data. The six types of data available are as follows.

a. Markers.

Genetic markers shown in SeqViewer are mapped based on sequence identity and include the following types: simple sequence length polymorphisms (SSLPs); cleaved amplified polymorphisms (CAPS); amplified fragment length polymorphisms (AFLPs); restriction fragment length polymorphisms (RFLPs); and other markers detected by hybridization such as single nucleotide polymorphisms (SNPs) used on whole-genome mapping arrays (Borevitz and Nordborg, 2003). The marker type is indicated in the mouse-over pop-up window. Markers can be used for a variety of purposes such as genetic mapping or positional cloning and as linked markers for tracking specific alleles.

b. Polymorphisms.

Polymorphisms include substitutions, small insertions (less than 20 bp), deletions, and combination insertion-deletions (INDELs). All polymorphisms are mapped relative to the reference (Col-0) genome, although TAIR includes polymorphisms between a wide variety of other natural variants (ecotypes). Sequence variations between natural populations can be used as a starting point for generating genetic markers for mapping, designing allele specific primers for a given locus, quantitative trait analysis, and linkage disequilibrium studies. TAIR has incorporated and mapped hundreds of thousands of polymorphisms from several large-scale SNP identification projects including the recent Perlegen project (Clark et al., 2007), the Multinational Arabidopsis Steering Committee SNP Database (https://www.genomforschung.uni-bielefeld.de/GF-dataresources/masc/search_masc_snps.php), Nordberg Lab Genomic Survey and Linkage Disequilibrium project (http://walnut.usc.edu/2010), and the Stanford Genome Center (SGC). Cereon has also made a list of over 50,000 polymorphisms between the two most common laboratory strains (Columbia and Landsberg erecta) freely available to academic researchers and nonprofit institutions. The Cereon dataset has not been incorporated into the SeqViewer, but can be downloaded from the TAIR Web site (http://arabidopsis.org/Cereon/index.jsp).

c. T-DNAs/Transposons.

Plant genomic sequences flanking T-DNA or transposon insertions are used to map approximate insertion sites onto the genome. The arrowhead indicates the direction of the sequence and points away from the insertion site which lies at the opposite end. Thick lines indicate regions of the insertion flanking sequence that match to the genome and thinner lines indicate regions that do not match the genomic sequence. The mouse-over pop-up window shows the name of the insertion, the start position, and length of the flanking sequence. The positions of insertions are rough estimates and should be confirmed by amplifying the product using genomic and insertion sequence primers and resequencing the ends. The validated sequences can be used to update the records in TAIR.

d. Gene models.

For each locus, the representative gene model and all splice variants are displayed in this band. Each splice variant is indicated by a suffix following the locus identifier (e.g., .1, .2, .3). The direction of transcription is indicated by the arrowhead. Exons are shown in yellow, introns are white, and UTRs (if known) are shown in red.

e. Transcripts.

Full-length or partial cDNA transcripts are indicated in blue, and ESTs are gray. The exon-spanning regions are indicated with solid boxes and the introns by thin lines. The direction of transcription is indicated by the arrowheads. For genes without full-length cDNA support, the transcripts can be used to verify the gene model structure as well as to identify misannotated or unannotated genes. Some transcripts may map to intergenic regions and may indicate the presence of a gene that has not yet been annotated. Other transcripts may indicate the presence of alternatively spliced forms, or genes for which de novo methods of detection predicted incorrect products.

f. Annotation units.

Annotation units are units of sequence derived from large-insert clones that comprise the backbone of the whole-chromosome assembly. To simplify construction of the tiling path and annotation of genes in regions of overlap, some of the original genomic clone sequences were trimmed and others were extended based on neighboring clone sequences. Therefore, the annotation unit sequences no longer represent the original clone sequences, and the coordinates of genes and other features mapped on annotation units differ from the coordinates on the corresponding clone sequences found in GenBank. Genomic sequence corrections applied for TAIR genome releases require recalculation of the chromosome coordinates and assembled sequence. TAIR maintains a list (http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/agicomplete.jsp) of incompletely sequenced BACs and known gaps remaining from the genome sequencing project.

The Arabidopsis genome assembly did not change between TIGR5 and TAIR8, but has been updated for TAIR9 (June 19, 2009). To find out more details on the TAIR9 assembly update please go to ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release/readme_TAIR9.txt

TAIR provides a script that allows users to convert coordinates from TIGR5, TAIR6, TAIR7, or TAIR8 to the coordinates on the updated TAIR9 assembly. This script, and instructions on how to run it, can be found here: ftp://ftp.arabidopsis.org/home/tair/Software/UpdateCoord/

Using the SeqViewer Nucleotide View to view annotations

From the SeqViewer Close-up view, there are several ways to drill down to the 10 kb Nucleotide View.

8. From the Close-up view, use the sequence ruler (indicated as "d" in Fig. 1.11.4) to select a region of the genome to view. Point the cursor to the desired area in the ruler and click. A SeqViewer Nucleotide View window appears as shown in Figure 1.11.5.

   Alternatively, to view a 10-kb region centered on a specific object, position the cursor over the object (such as the locus AT3G02310) and click on the link to "nucleotide seq view" in the pop-up window that appears. The start of the locus will be on the fifth line of sequence in the 10-kb window. A third option is to go directly to the Nucleotide View from the list of matches to the genome. After submitting a query, click on the link to the list of matches above the whole-genome view in the SeqViewer. Click on the coordinates in the last column of the list (Location) to display the Nucleotide View.

Displaying selected features in the Nucleotide View

The Nucleotide View (Fig. 1.11.5) shows 10 kb of sequence at a time. The view can be scrolled 5 kb upstream or downstream using the arrows at the top and bottom of the view. The location of genes is shown on the far right; the direction of the arrow indicates the direction of transcription. The display can be set to show specific sequence features singly or in combination, on either DNA strand. An explanation of the display is shown in the legend at the top of the page. The procedure below describes how to view the location of the T-DNA insertion AL940283 in *SEP2* in the Nucleotide View.

9. In the drop-down menu selector labeled Choose Objects to be Highlighted (located in the upper right corner of the Nucleotide View, shown in Fig. 1.11.5) select Genes/T-DNA/Tn Insertions. The display shows the translation start and stop points of *SEP2* in blue highlighting; the UTRs are red, exons are uppercase and yellow, and introns are lowercase and purple. The positions matching insertion flanking sequences are shown underneath the corresponding genomic sequence. Matching regions are represented by a double dashed line (===); nonmatching regions are represented by a single dashed line (—), and the approximate point of insertion at the 5′ end of the flanking sequence is represented by a vertical line (|). The arrowhead (3′ end of the flanking sequence) shows the orientation of the flanking sequence relative to the chromosome.

Figure 1.11.5 A nucleotide sequence view centered on AT1G07810 showing annotated genes and T-DNA/transposon insertion flanking sequences. The drop-down menu (shown in upper right corner) was used to select the items to display in the nucleotide sequence view.

One of the many nice features of SeqViewer is the ability to copy and paste sequences directly from the nucleotide sequence view where the upper/lowercase formatting is retained. This can be useful when exporting sequences to primer design programs and selecting primers that span introns.

Exploring GBrowse

Viewing a gene or region of interest in GBrowse

10. Go to the TAIR home page (*http://www.arabidopsis.org*). In the Tools section of the menu bar, click on the link to GBrowse (Fig. 1.11.1). Alternatively, go directly to the URL *http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/*.

    The GBrowse (Fig. 1.11.6A,B) display is divided into five main sections: (1) Instructions, which provides sample data entry points and examples of GBrowse search queries; (2) Search, which allows you to input your query and select the data source; (3) Overview, which shows a graphical representation of the chromosome and region currently displayed; (4) Details, which provides a pictorial representation of the genomic features in the selected region; and (5) Tracks, which allows you to customize the display settings and select which features are displayed in the details section.

11. The names and position of genomic features such as genes or genetic markers can be entered in the search box (Fig. 1.11.6A). For genes, either the AGI code (e.g., `AT1G05460`) or gene symbol (e.g., `SDE3`) are valid search queries. Nucleotide

ranges can also be entered to allow specific regions of interest to be displayed. The chromosome and start and end coordinate of the desired region must be entered in the following format `Chr1:1504365..1514364`.

If a query returns multiple hits, GBrowse will display these as distinct rows with the position of each feature shown. Clicking the hyperlink will open the detail display for the selected region.

12. The assembly version or build can be selected from the data source drop-down menu. By default, the most recent version is displayed.

    Altering the assembly version changes the chromosome sequence, gene models, and other tracks to those of the specified release, allowing alternative versions to be compared. Note that gene models or other features present in a later release may potentially be absent, located at a different position, or otherwise altered in an earlier release. A description of the latest genome release can be found at http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp, details of earlier releases can be found on the TAIR FTP site under the respective release directory, ftp://ftp.arabidopsis.org/home/tair/Genes/.

13. Entering a feature name or region and clicking Search will update the overview and details display. The overview map shows the position of the region displayed in the detail view relative to the rest of the chromosome. The size of this region is shown in the Scroll/Zoom drop-down. As default, certain annotated data is displayed in the details panel; this includes gene models, annotation units (BACs), *Arabidopsis* cDNAs, and polymorphisms.

    If a specific feature was searched for by name (e.g., `AT1G05460.1`), the feature is highlighted in yellow. This highlighting provides a convenient way of maintaining position of the feature when the view is expanded to display a larger region. Highlighting can be turned off by clicking the Clear highlighting button directly below the details display panel.

14. The zoom feature can be used to adjust the viewing dimensions in order to display a larger-scale view of the genome. Select the desired region size from the drop-down menu and the display will automatically reload to a new detail map covering the region.

    The change in view scale is reflected in the increased size and position of the yellow box in the overview panel and the change in genomic coordinates in the search box.

15. To move along the chromosome, use the arrow buttons to shift the display to the left or right. The entire display of the chromosome can be flipped by checking the Flip box and clicking the Update Image button (at the lower right of the image). Flipping the display may be useful when viewing a minus strand gene.

    Moving the cursor over a feature will bring up a pop-up box which displays additional information specific to the feature. For genes, this includes known symbols and the functional description. Every feature is also hyperlinked; clicking on a feature will open a new data page specific to the feature. For example, clicking on a protein coding gene model will open the TAIR gene page, whereas clicking on a Brassica EST transcript will link out to the relevant GenBank entry at NCBI.

Figure 1.11.6 *(appears on next page)* Overview of the GBrowse tool. (**A**) The upper panel of this tool allows the user to search for landmarks in the genome such as gene names or chromosome positions, and to zoom in and out of a specific genomic region. The central panel displays a series of data tracks such as genes, cDNAs, polymorphisms, the VISTA sequence similarity track and many more. (**B**) Using the track menu on the bottom, the user can choose which tracks to display by selecting either whole data categories or specific types of data. For the color version of this figure go to *http://currentprotocols.com/protocol/bi0111*.

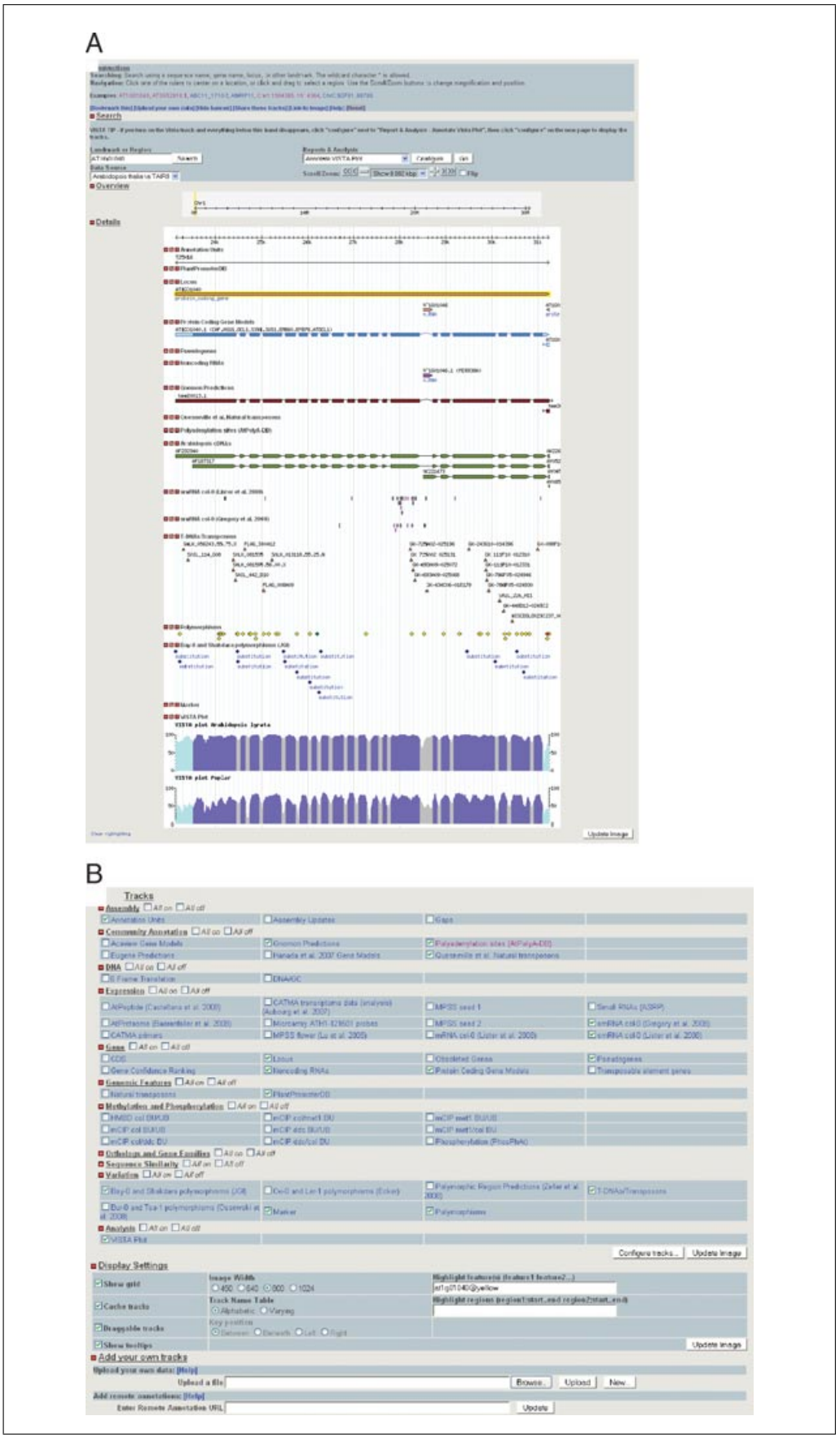

Figure 1.11.6 *(legend appears on previous page)*

Customizing the GBrowse display

16. TAIR GBrowse has 11 track categories: Assembly, "Community annotation," DNA, Expression, Gene, Genomic Features, "Methylation/phosphorylation," "Orthologs/gene families," "Sequence similarity," Variation, and Analysis. Each track category has multiple check boxes for different types of data (Fig. 1.11.6B).

    Further information about the track can be obtained by moving the cursor over the track name. In addition, clicking the track name opens a separate page describing all available tracks.

17. To add or remove tracks from the detail display simply check or uncheck the required tracks and click the Update Image button. The track order can be adjusted by simply clicking the track title in the details panel and dragging the track to a new position.

18. Each track can be individually configured by clicking the question mark box next to the track title. This allows the user to choose the shape and color of the glyphs, put a limit on the number of features displayed in any one region, and set preferences if a text label is displayed.

    Tracks can also be customized by clicking on the Configure tracks button. To revert to the default settings click Revert to default.

19. The display settings panel can be used to change certain features of the details view such as the image width, position of the key, or how the tracks are listed. The highlight feature boxes can be used to highlight specific features or regions which may be useful when giving presentations or showing images in publications.

Visualizing private data in GBrowse

20. To upload your own annotation data to GBrowse go to the "Add your own tracks" panel at the bottom of the GBrowse page (Fig. 1.11.6B). This feature allows you to view your own annotations such as primers and cDNA clones in the context of the *Arabidopsis* genome. GBrowse requires a text file describing the features you wish to view. The browser accepts the annotation file format described in the tutorial document available by clicking on the help hyperlink in the "Add your own tracks" panel. Alternatively the annotations can be uploaded in generic feature format version 3 (GFF3); for details see *http://song.sourceforge.net/gff3.shtml*.

21. Locate your created text file by clicking on the Browse button; when found, upload the file. Once uploaded, GBrowse will automatically incorporate your annotations into the details view. A track option will also be created below the previously existing tracks. Uploaded tracks can be configured in the same manner as tracks provided by TAIR.

22. The Annotation data text file can be edited or deleted by clicking on the respective buttons in the Add your own tracks panel.

    Uploaded annotations will persist until you delete them; these annotations are private and will not be seen by other individuals.

23. To upload remote annotations to GBrowse, paste the Web address into the Enter Remote Annotation URL box. This feature allows you to view annotations created by other groups in your own GBrowse. In addition, if you have access to a Web sever, you can publish your own tracks so that they are available to colleagues or collaborators. Further details about this process can be found on the hyperlinked help page.

Using the decorated FASTA function

GBrowse allows you to download a decorated FASTA file. (FASTA files are described in *APPENDIX 1B*). This option allows you to extract the sequence in a particular region and highlight specific features of interest. For example, coding regions can be marked in a different colored font and polymorphisms shown in bold or underlined, allowing you to easily identify which polymorphisms lie in coding regions.

24. Go to the Reports and Analysis features box and select Download Decorated FASTA file from the menu options (Fig. 1.11.6A). Clicking Configure opens the feature configuration page.

25. From the configuration page you can select which features you wish to highlight on the FASTA sequence file and choose from a variety of markup options such as caps, italics, bold, and alternative font and background colors.

26. Once satisfied with your selection click Go. The new Web page will display the FASTA sequence for the region displayed in the detail view with the selected features highlighted.

BASIC PROTOCOL 4

USING THE GENE ONTOLOGY ANNOTATIONS: FINDING GENES WITH SIMILAR FUNCTIONS

Annotations (associations of controlled vocabularies or keywords to data objects) provide a richer, more complex picture of a gene that is also more computationally accessible for the purpose of querying, classification, and making correlations among seemingly unrelated data. TAIR makes extensive use of controlled vocabularies for describing data in the database. The controlled vocabularies (ontologies) that are used by TAIR are also used by other model organism databases, thereby facilitating cross-species comparisons. All of the ontologies used by TAIR are included in the Open Biological Ontologies Project (*http://obo.sourceforge.net*) where they are freely accessible.

TAIR is member of the Gene Ontology (GO) Consortium (*http://www.geneontology.org*) and participates by developing and refining the ontologies and annotating gene products (The Gene Ontology Consortium, 2010). The GO controlled vocabularies describe three aspects of gene products: molecular function, biological process, and subcellular location. Both TAIR and The Institute for Genomic Research (TIGR) have used the GO ontologies for annotating *Arabidopsis* gene products (Wortman et al., 2003; Berardini et al., 2004). Both TIGR's and TAIR's annotations are displayed in TAIR and contributed independently to the GO database, where they are accessible through the AmiGO query tool for making cross-species queries. The other main ontology used at TAIR is developed by the Plant Ontology Consortium (POC; *http://www.plantontology.org*). The POC is using the GO model to develop controlled vocabularies for plant structures and developmental stages. In TAIR, both of these ontologies are used to annotate many additional types of data such as microarray experiments, gene expression, phenotypes, and publications.

Necessary Resources

Hardware

Computer with Internet access

Software

Up-to-date Web browser. The browser must have cookies enabled to log in and process stock orders. TAIR makes extensive use of JavaScript; this feature must also be enabled. See *http://www.arabidopsis.org/help/index.jsp* for information on properly configuring one's browser.

Files

`WRKYFamily.txt`

Using the Keyword Browser to find candidate genes

For researchers, finding candidate genes involved in a particular pathway typically involves a fishing expedition using a variety of genetic, molecular, and biochemical assays. The GO annotations can be useful in making educated guesses about what genes may act in a pathway or are members of transcriptional/signaling cascades. These predictions can then be tested experimentally. For example, *ERA1* (AT5G40280) encodes a protein farnesyltransferase; mutants have low prenylation levels and defects in meristem organization and abscisic acid–mediated responses (Cutler et al., 1996; Yalovsky et al., 2000; Ziegelhoffer et al., 2000). What other genes might be involved in prenylation, and do they act in the same or another pathway?

1. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select Keywords from the drop-down menu that appears. The page shown in Figure 1.11.7A is returned (TAIR Keyword Search and Browse; can be directly accessed at *http://www.arabidopsis.org/servlets/Search?action=new_search&type=keyword*). Enter term (keyword) `farnesyltransferase` in the text box and choose "contains" (an inexact search) from the drop-down menu to the left of the text box. From the group of check boxes for restricting the search, choose GO Molecular Function as the keyword type and click the Submit Query button.

   Many of the terms in GO exist as complex phrases. TAIR searches take the entire entered term or phrase as a complete phrase rather than a set of words. Consequently, an "exact match" search will often not retrieve any entries. Therefore, the authors recommend using the "contains" option for keyword searches.

2. On the Keyword Search Results page (Fig. 1.11.7A), each controlled vocabulary term is displayed along with a count of all data objects (e.g., loci, publications, annotations) annotated to that term. Click "loci" to display the genes annotated to "farnesyltransferase activity." In this example, the results list includes *ERA1* (AT5G40280) and 12 other genes (this number may change as annotations are constantly updated). One of these genes, *AT3G59380*, was originally predicted to encode a putative alpha subunit of a protein farnesylation complex. Thus, it was annotated to the term "farnesyltranseferase activity" with the corresponding evidence code, inferred by sequence similarity (ISS). Based on this information it would be reasonable to predict that *AT3G59380* may partner with or act in the same pathway as *ERA1*. As it turns out, *AT3G59380* was found to correspond to the *PLURIPETALA* (*PLP*) locus. *plp* mutants have a similar phenotype to *ERA1* mutants, and the locus is epistatic to *ERA1*, indicating that PLP functions in the same pathway (Running et al., 2004).

Finding genes annotated to related functions

3. On the Keyword Search Results page, find the listing for "protein farnesyltransferase activity," and click on the "treeview" link. This will open a window displaying the term in a hierarchical tree view (Fig. 1.11.7B).

   In the Gene Ontologies, terms have a parent-child relationship to one another. Parent terms are generally less specific than their child terms. A child term may be a part of the parent (as thylakoid is part of chloroplast) or a type of the parent (as chloroplast is a type of plastid). In contrast to simple hierarchies, a child term may have more than one parent. The ontologies are intended to be as biologically accurate as possible. Terms and their relationships are defined by what is known about the biology of the process, function, or cellular component. By examining the structure of the ontology to find related terms, related gene products can also be found via their annotations to the terms.

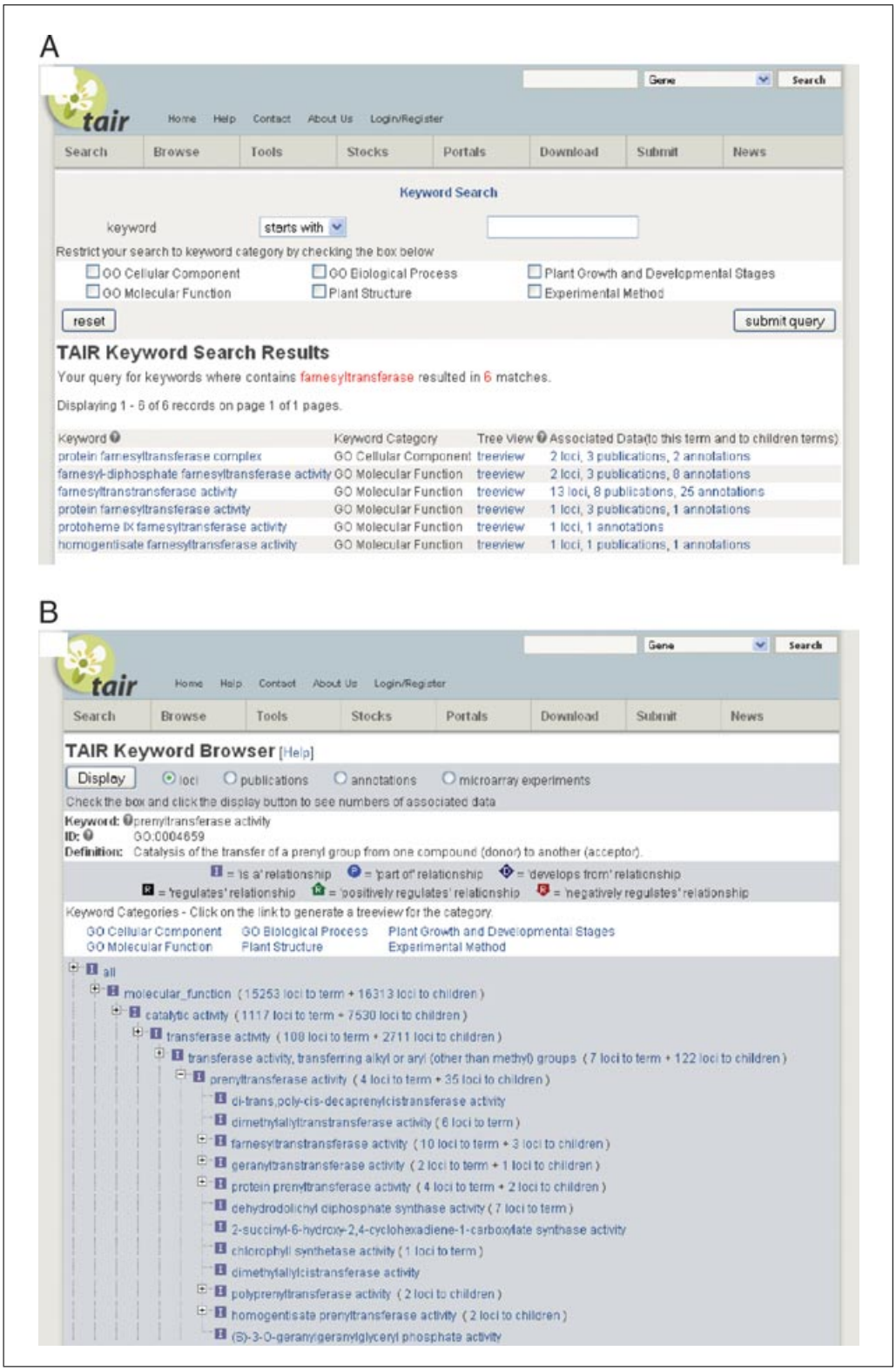

Figure 1.11.7 (**A**) Keyword search results after querying for the GO Molecular Function terms containing the word farnesyltransferase. (**B**) A tree view of the term "protein farnesyltransferase activity" and associated gene annotations.

4. Click on the plus sign next to the parent term ("protein prenyltransferase activity") to expand the node and display all of the child terms.

5. To display genes annotated to each of the parent and child terms, select the "loci" radio button from the top of the tree view page (Fig. 1.11.7B), then click the Display button. The display will be redrawn to show a count of the number of loci annotated to each term and the number of loci annotated to the children of each term. Click on the link to list loci annotated to the term "protein prenyltransferase activity" to find all proteins that are annotated to this term.

Using the GO annotations to group sets of genes

GO Annotations can also be used to rapidly classify related genes such as gene families or co-clustered genes revealed by analysis of microarray expression data.

6. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select GO annotations from the drop-down menu that appears. Alternatively, go to the URL *http://www.arabidopsis.org/tools/bulk/go/index.jsp*.

7. Upload a list of AGI locus identifiers using the sample data file `WRKYFamily.txt`. This file contains a list of 74 loci all belonging to the WRKY transcription factor family (Eulgem et al., 2000). Select the Text radio button under "Select output type"; it is desirable to save the results in a table format on one's own computer. Click on the "Get all GO Annotations" button. The output file contains a list of all the specified loci and their annotations to all three aspects of the GO ontology.

   The annotations include the evidence code and reference for the data supporting the annotation. The file can be saved onto a local computer as a tab-delimited text file. If the HTML option is chosen, the results are hyperlinked to TAIR detail pages for loci, keywords, and publications. The Web output also has links to the corresponding keyword entry in the GO database, where one can find annotations to genes from other organisms.

Classifying sets of genes into functional categories

8. Alternatively, instead of getting a list of all annotations, the genes can be grouped into broader categories based on their annotations. After uploading the gene list (step 7 above), choose "HTML output" and click the Functional Categorization button.

   For each aspect of the GO ontologies, a subset of terms have been selected to represent 10 to 20 major categories, called GO Slim categories. If a gene is annotated to a child term of one of the GO Slim terms, it is included in the category. The GO Slim is less specific, but presents a simpler classification. The results include gene annotations that are both experimentally supported and computationally predicted. To find sets of annotated genes based on evidence codes, use Search by Associated Keyword on the Gene Search page (http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene). GO Slim assignments are also included in the detailed GO annotation output (from step 7). See http://arabidopsis.org/help/helppages/go_slim_help.jsp for a list of all GO Slim terms and their definitions.

9. The database will return a functional categorization list showing all categories represented in the genes from the input file, along with the frequency of distribution of the genes within the set (Fig. 1.11.8A). To view a list of genes in each category, click on the number in the "Gene count" column.

   Only the categories represented by the genes in the list are included; the absence of any of the GO Slim categories means that there are no genes in the list that fall into that particular group. The default option displays the list grouped by keyword type and then by categories. The table can be re-sorted to list the most to least frequent categories. Frequency refers to the number of occurrences of a gene-keyword pair in the list. Multiple annotations to the same term are essentially compressed in this view, in contrast to the

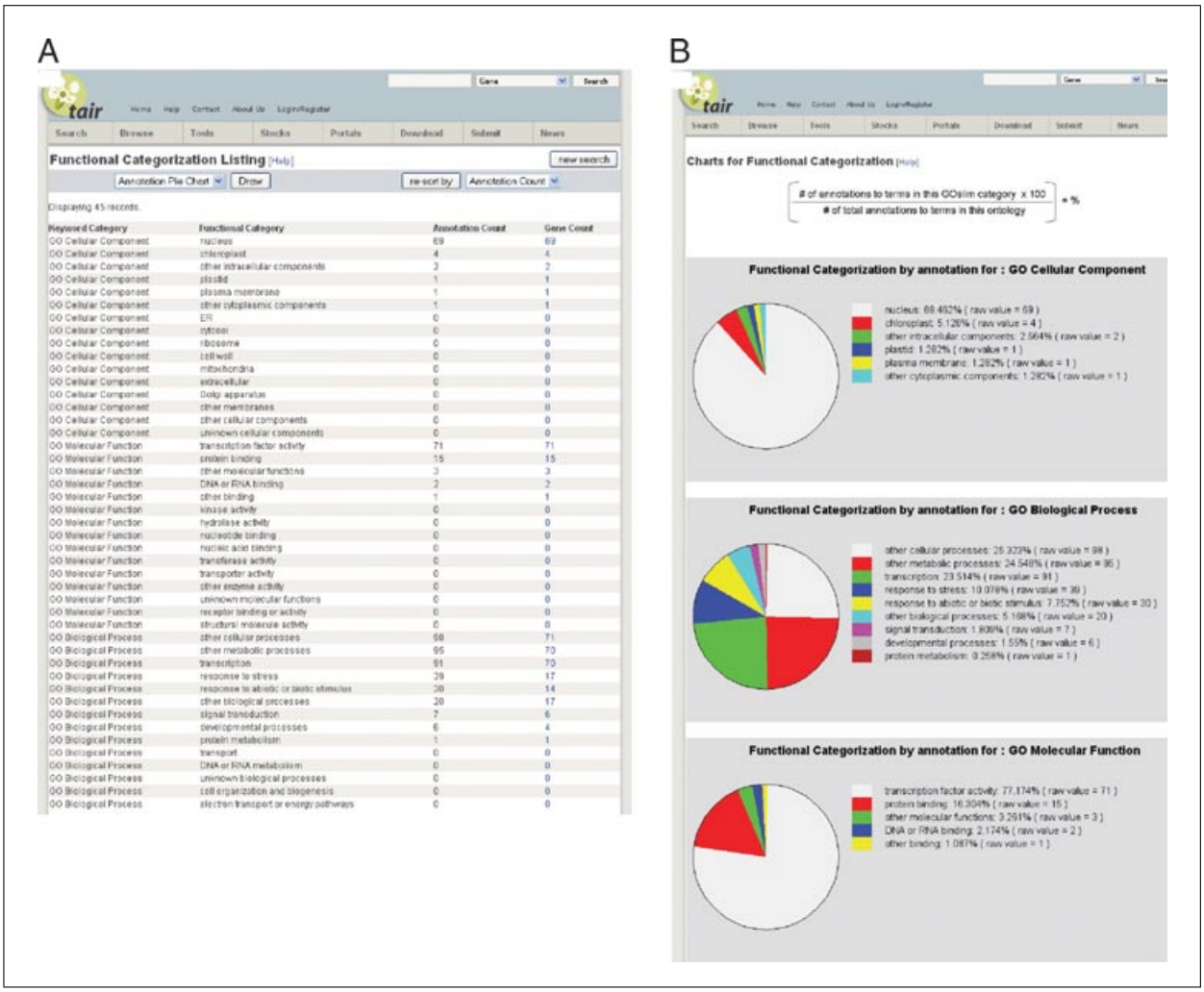

Figure 1.11.8 (**A**) Results display for functional categorization of *WRKY* genes. The members of this family fall into 22 different GO Slim categories based on their annotations to more granular GO terms. The list can be re-sorted by choosing Frequency from the "re-sort by:" drop-down menu and clicking on the "re-sort" button. The list of 22 categories is shown grouped by keyword category. The frequency of annotations to each category is listed in the last column; the number is linked to a list of genes annotated to the terms that are children of that category. (**B**) Clicking on the "create pie charts" button generates pie charts showing the distribution and frequency of annotations to each of the GO slim terms. A different pie chart is created for each aspect of the GO ontologies.

Get all GO annotations option. Genes that are annotated to multiple terms that fall into different categories will be included in each of the GO Slim bins. Therefore, the total number of annotations to each aspect of the GO ontologies may be greater than the total number of genes in the query list.

Displaying the functional classification as pie charts

10. Above the Functional Category column (Fig. 1.11.8A), click on "create pie charts." This will create a new page showing three separate pie charts, one for each aspect of the Gene Ontology (Fig. 1.11.8B). Depending on how the results are sorted, the sections can be displayed from most to least frequent category, or by related categories. The percentage of the total is shown in the color key for each graph.

11. To save the graph images, hold down the Ctrl key while clicking on the image, or right click the mouse if using a PC, and save the image to the clipboard or to a file. The images are in Graphic Interchange Format (GIF), which can be opened using a variety of graphics software.

FINDING AND ORDERING MUTANT SEEDS AND cDNA CLONES FROM THE STOCK CENTER

In an effort to define the function of all *Arabidopsis* genes, the research community, supported by funding agencies, has invested heavily in generating populations of *Arabidopsis* mutants. Different types of mutations provide different information about a gene's function. TAIR's database includes mutants generated by chemical mutagenesis and insertional mutagenesis using T-DNA or transposons to generate knockouts and enhancer trap/gene trap constructs that can both disrupt gene function and reveal expression patterns. Other insertional mutagenesis strategies produce overexpression or ectopic expression phenotypes, which are useful when the loss of gene function does not reveal any overt phenotype. TAIR's Germplasm search can be used to find *Arabidopsis* strains containing mutations in a gene of interest.

Necessary Resources

See Basic Protocol 1

Finding mutants in a gene

1. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select Seed/Germplasm from the drop-down menu that appears.. Alternatively, go directly to the URL *http://www.arabidopsis.org/servlets/Search?action=new_search&type=germplasm*.

2. The resulting Germplasm Search page (not shown in the figures) offers various search options. In the first section, called Search by Name, Phenotype or Stock Number, type in the name of the gene (in this example, `PIN1`). Choose to search by "gene name" (the default setting) and for an exact match by selecting "exactly." Submit the query by clicking the Submit Query button.

   The search allows for up to three name fields to be selected, and each box has a drop-down menu of attributes that can be chosen. When multiple terms are entered, the query is treated as an AND search. Searching with "gene symbol" set to `PIN1` *and "description" set to* `meristem` *will search for all Germplasms associated with PIN1 that contain the term "meristem" in the description. A common reason why searches fail is because too many restrictive options are selected. Click on Help next to the Germplasm Search page title to learn more about all of the search parameters.*

3. The resulting page displays a summary of the germplasms. If there are images associated to the germplasms in the database, a small camera icon appears next to the name. Click on the icon to view the image. If the germplasm is available as a stock from the ABRC, a check box will appear at the far right of each record. Detailed information about the stock, polymorphism, and germplasm can be found by clicking on the respective names of the data in the results summary, in this case the germplasm name CS8065.

Ordering stocks

To order stocks from the ABRC, one must be registered at TAIR and be affiliated with a laboratory as a member or principal investigator. Instructions on how to register online can be found in the help documents linked to the online community search and registration forms (*http://arabidopsis.org/servlets/Community?type=person&action=edit&new=true*).

4. Examine the germplasm records; click on the name of the germplasm to view specific details such as phenotype and polymorphism data. If the germplasm is an ABRC stock, the stock information will be included in the detail page. Select each stock to order by checking the Select box in the Stock Information section on the germplasm detail page. Next, click on the Order from ABRC button.

Users outside of North America should check the Order from NASC button. The Nottingham Arabidopsis Stock Center (NASC) provides seeds to users outside of North America.

5. Alternatively, stocks can be added directly to one's "cart" by clicking the Order check box for each stock on the results summary page. Once all of the selected stocks on a page have been chosen, click the "order checked stock" button in the upper right corner of the results page (not shown).

   If one is not logged in, one will be prompted to log in, and if one is not affiliated with a laboratory one will be asked to affiliate oneself to a laboratory. A laboratory affiliation is required for placing stock orders. If one does not have a login (i.e., are not registered) one will be prompted to register. Since registration typically takes less than 24 hr to activate, one should download and save one's search results and try placing the order again when notification has been received that the account has been activated. Another way to find mutants is to use the Polymorphism/Allele search (http://www.arabidopsis.org/servlets/Search?action=new_search&type=polyallele). The results may differ because there are many Polymorphisms without corresponding Germplasm entries. For example, polymorphisms due to natural allelic variations (e.g., ecotype differences) and T-DNA/transposon insertions are loaded into TAIR and associated to loci, but may not have links to Germplasm entries. This is generally the case for T-DNA/Transposon insertion lines that are not available from the ABRC. Other ways of finding mapped alleles and insertions is to use SeqViewer (Basic Protocol 3) or BLAST to search the Insertion Flanking sequence data set.

Finding full-length cDNA or EST clones for a gene of interest

TAIR's database contains hundreds of thousands of clones, including full-length cDNA and expressed sequence tag (EST) clones generated by functional genomics projects, which are available to the research community from the ABRC or other sources. ESTs generated by sequencing cDNA clone ends may also include the entire coding sequence and can be useful as probes for Northern blotting or in situ hybridization.

6. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select DNA/Clones from the drop-down menu that appears.. Alternatively, go directly to the URL *http://www.arabidopsis.org/servlets/Search?action=new_search&type=dna*.

7. From the drop-down menu in the Output Options section, choose "clone" for searching full-length cDNA clones or "clone end" for searching ESTs. In the "Search by" section, choose "GenBank accession" and "exactly" from the respective drop-down menus, and type the accession number (for example, `AY040062`) in the text box to the right. Click the Submit Query" button.

   The available search options depend on the type of DNA being searched (e.g., clone, clone end, pooled genomic DNA, library, vector, filter, stock, or host strain). The search options for clones and clone ends are the same, but the results differ. Consult the DNA search help pages (http://arabidopsis.org/help/helppages/dnasearch.jsp) for information on the specific parameters and how to use them.

8. If the clone is available as a stock from the ABRC, check the order box in the last column and then click on the Order Checked Stock. If the clone is not available as a stock or is not in stock, click on the link from GenBank accession number to go to the GenBank record. This record should have information for the donor of the sequence, who can be contacted to obtain the clone directly.

   If not sure of the exact stock to look for, or simply to find out what is available, one can browse the ABRC stock catalog (http://www.arabidopsis.org/servlets/Order?state=catalog). The catalog is organized into different categories of DNA and Seed stocks. Catalog entries are linked either to the relevant stock pages or to information about the class of stocks and how to order them.

USING PUBLIC MICROARRAY DATA IN TAIR

BASIC PROTOCOL 6

TAIR provides access to experimental results from both cDNA- and Affymetrix-based platforms of microarray data that TAIR received before June 2005. More recent *Arabidopsis* microarray data can be found in ArrayExpress, GEO, and NSCArrays. Experimental data come from large-scale functional genomics projects such as the *Arabidopsis* Functional Genomics Consortium (AFGC) and the European-initiated *Arabidopsis* Functional Genomics Network (AFGN), as well as individual submissions from researchers. Curation of the experimental data includes associations to controlled vocabularies and classification to facilitate searching. The microarray data that TAIR accepted includes both raw and user normalized data values; in addition, TAIR applied a standard normalization method to all submissions to facilitate comparison of different experiments. For a description of TAIR's normalization methods, see the documentation on the TAIR Web site *http://arabidopsis.org/tools/bulk/microarray/analysis/index.jsp*.

Necessary Resources

See Basic Protocol 1

Finding sets of genes that are similarly expressed from microarray experiments

A common use of public microarray data mining is to identify genes that behave similarly in all experiments. First it is necessary to search among the available sets of experiments to find relevant ones. The normalized data can then be exported into clustering software to identify sets of genes which show similar behavior. This protocol describes how to access and download public expression data for a subset of experiments preformed on a single platform (e.g., Affymetrix).

1. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select Microarray Experiment from the drop-down menu that appears. Alternatively, go directly to the URL *http://arabidopsis.org/servlets/Search?type=expr&search_action=new_search*.

2. Choose Affymetrix from the drop-down menu under Search by Array Manufacturer to limit the results to experiments using this platform. This will include both the earlier 8K and newer 25K ATH1 (whole-genome) chips. To narrow down the results to a specific subset of related experiments, choose "tissue comparison" from the Experiment Category drop-down menu. Click the "submit query" button at the top or bottom of the form.

Finding information about the experiment

3. The experiments matching the submitted query (Affymetrix chip, tissue-comparison experiments) are summarized in the results page that is returned. To find information about the experiment, click on the name of the experiment, in this case Tissue Type Arrays of Columbia-0. This opens a new page showing a summary of the experiment (Fig. 1.11.9A). At the top of the page are a series of tabs that can be used to navigate to specific subsections of the microarray experiment data. Click the tab marked Samples.

4. Each of the samples used in the experiment are listed in alternating sections on the Samples detail page. The data include links to the germplasms that were the source of tissue used in the experiment, a description of the tissue used to prepare the RNA, and links to the RNA extraction method. Environmental conditions employed for each of the samples, including any treatments, are displayed in a table.

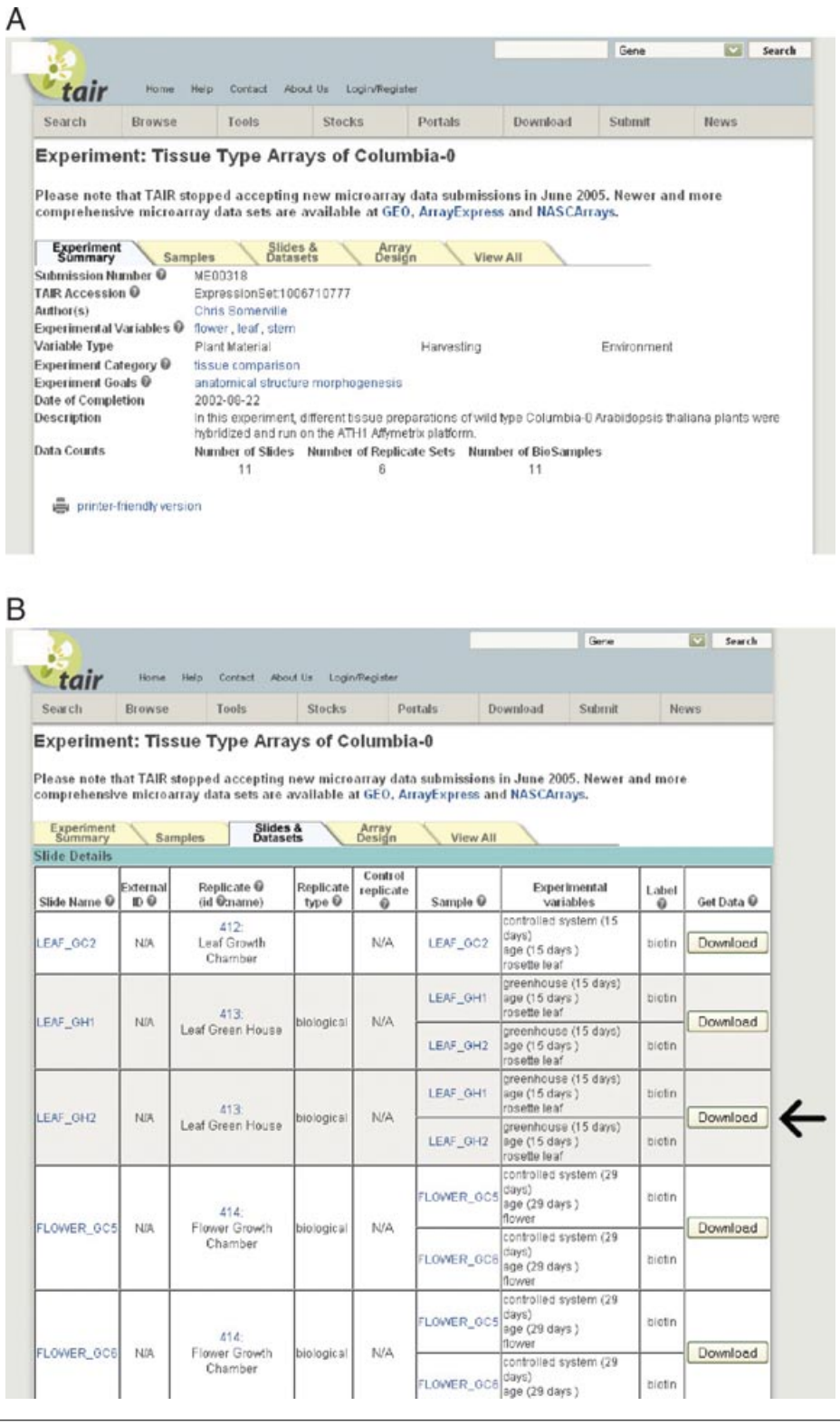

Figure 1.11.9 Subsections of a detail page for a microarray experiment. The tabs are used to navigate to different subsets of information about the experiment. (**A**) The experiment summary page, which is linked from the experiment search results. (**B**) The Slides & Datasets subsection for the experiment. The download button (arrow) links to an automatic download of tables containing the experiment summary along with raw and normalized values for each slide in the set.

Downloading microarray data

5. Click the tab labeled Slides & Datasets. Each slide (hybridization) that is a part of the experiment is displayed in this section (Fig. 1.11.9B). Replicate sets are grouped together and shown in alternating color bands. Replicates may be technical, where the same RNA sample was used for both hybridizations, or biological, where different RNA samples were obtained from identically treated plants and hybridized. For each hybridization in the experiment, the raw and normalized data can be downloaded by clicking on the Download button in the last column. The data files are compressed. Use a utility such as Stuffit Expander (Macintosh) or WinZip (PC) to expand the tab-delimited text files before opening them in Excel or other spreadsheet programs. Alternatively, entire data sets (all hybridizations in the experiment) can be downloaded from the TAIR FTP site (*ftp://ftp.arabidopsis.org/Microarrays/Datasets*) using the ExpressionSet identifier found on the Experiment Summary page to locate the appropriate experiments. Consult the `README` file in the `ftp` directory for information and a description of the data columns in the files.

Analyzing microarray data

TAIR's emphasis with microarray data has been to provide long-term storage and access to publicly available data. There are many other groups that have focused on developing analysis tools (see Suggestions for Further Analysis; also see Internet Resources at the end of this unit); however, TAIR provides limited resources for analyzing microarray expression data. AraCyc's Omics Viewer can be used to visualize changes in expression of genes involved in metabolic pathways (see Basic Protocol 9). For analyzing clustered data for all of the hybridizations in the database, TAIR has two tools that can be used to find genes with similar patterns of expression across all experiments (see Advanced Parameters in the Commentary).

Finding the expression pattern of a gene of interest

The expression pattern of a gene or set of genes across all or a subset of microarray experiments in TAIR's database can be queried and visualized using the Microarray Experiment Search.

6. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select Microarray Expression from the drop-down menu that appears. Alternatively, go directly to the URL *http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression*. This search can be used to find expression data for up to 100 genes (or array elements) using gene names, locus identifiers, microarray element names, or GenBank accession numbers. The search can be further restricted by expanding additional search options such as expression values and experiment type. The following steps show how to find all Affymetrix chip–based experiments in which the expression of a locus (AT1G55020, which encodes a protein with lipogenase activity) was detected.

7. Choose "locus" from the Search by Name or GenBank Accession drop-down menu and enter the `AT1G55020` in the text input box to the right of that menu (see Fig. 1.11.10).

8. To restrict the search to data from specific types of arrays, go to the section marked Select Array Type/Design, and choose Affymetrix GeneChips from the Array Type drop-down menu.

   The default Any selection, for the Array Design drop-down menu (on the right of the Array Type menu), includes data from both 8K and 25K arrays. If not sure what array contains the gene of interest, keep this default selection. To further narrow the results, it is possible to choose one or the other array type in order to compare the same array element.

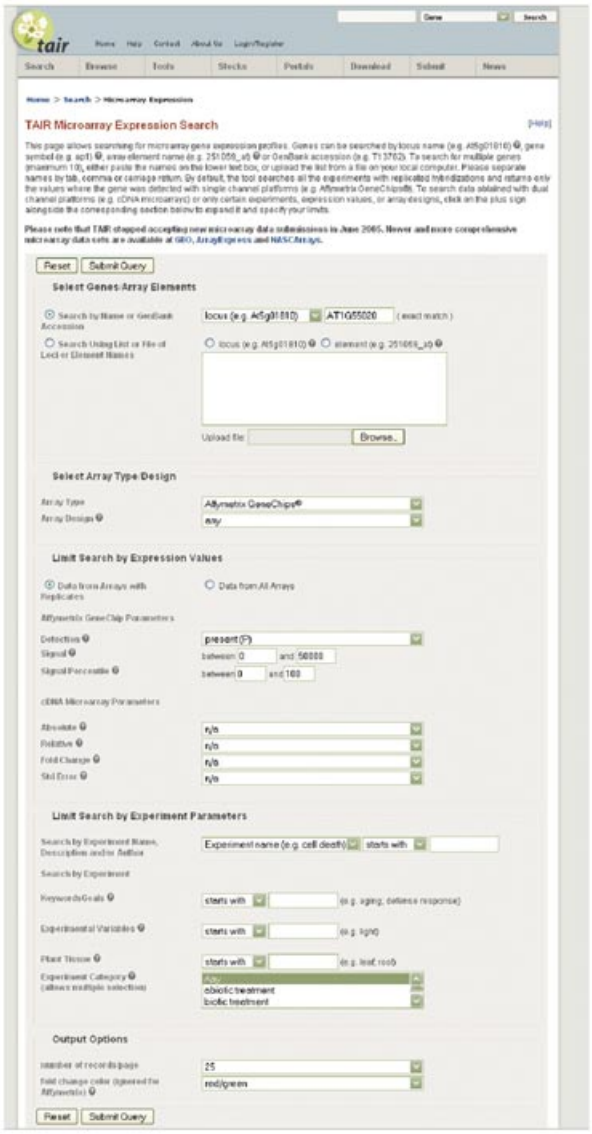

Figure 1.11.10 Web interface for searching gene expression data from microarray experiments in TAIR.

9. The Limit Search by Expression Values section allows one to adjust expression value parameters for either Affymetrix arrays or cDNA arrays. Since this example involves Affymetrix data, use the parameter selections for this type of array. In the section marked Detection, choose Present, which will only include data from hybridizations where the transcript was detected.

   Another option available in this section is to include expression data from all experiments, not only those with replicate hybridizations. For the maximum number of results, set this option to include "Data from all arrays."

10. Click the Submit Query button. The results are displayed in a summary table specific for the Affymetrix data (Fig. 1.11.11). A different summary page is used for cDNA array data; see the Microarray Expression Help pages for information on this format (*http://arabidopsis.org/help/helppages/expression_search.jsp*). The data can be sorted by different fields, such as experiment name for grouping related data sets, or by locus identifier (for sorting results from multiple locus queries).

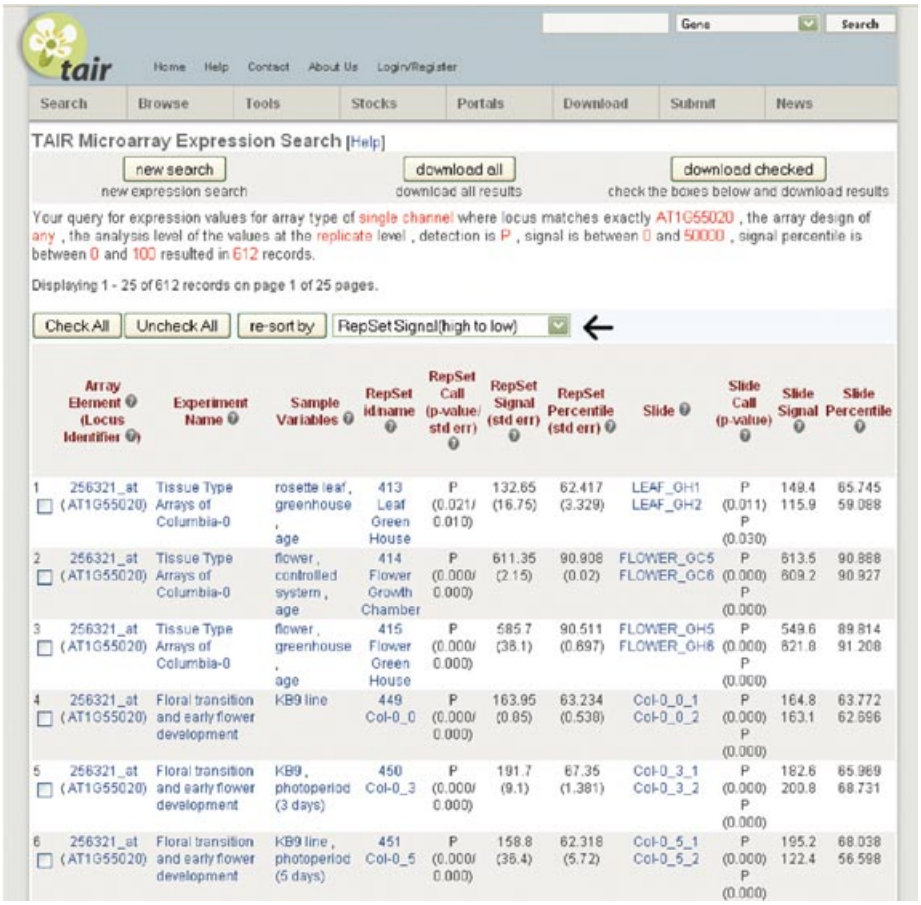

Figure 1.11.11 Sample result set for the Microarray Expression search using AT1G55020 locus. The drop-down menu (arrow) and resort button are used to change the order of the results display.

Using the results summary

11. The summarized results include information about expression values for the element and other data that are hyperlinked to the corresponding detail pages.

12. The first column lists the array element/locus names linked to their respective detail pages. The next two columns (Experiment Name, Sample Variables) can be used to find information about the experimental design. To find specific data for each experiment, click on the values in the column RepSet/id/name, which will jump to the Slides & Datasets subsection of the Microarray Experiment Details (see above), or go directly to the slide data.

13. Information about the expression value from replicate hybridizations is summarized for each element in (Fig. 1.11.11, columns 5, 6, and 7). The detection call indicates if the element was present/absent or marginal, and includes a *p*-value that reflects the confidence in the call. A *p*-value closer to zero represents a greater certainty. The replicate set signal represents the mean value (and standard error) for the element in replicate hybridizations. The signal value is a quantitative value calculated for each probe set and represents the relative level of expression of a transcript. The replicate set percentile compares the signal value for the element against all other elements in the replicate hybridization. Therefore, an element in the 90th percentile has an average signal greater than 90% of the other elements in the hybridization. If these columns are empty, no replicate set data is available.

14. For each individual slide (hybridization), the expression values (detection or slide call, slide signal, and slide percentile) are indicated in the last three columns.

SUPPORT PROTOCOL

MAPPING ARRAY ELEMENTS TO ANNOTATED LOCI

For both oligomer and cDNA arrays, it is essential to know what genes being assayed correspond to which probes (array elements). This protocol describes how to map array elements to the corresponding locus and how to obtain short gene summaries (annotations) that describe the gene products.

Necessary Resources

See Basic Protocol 1

Finding corresponding loci for a subset of array elements

1. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the upper menu bar (Fig. 1.11.1), and select Microarray Elements from the drop-down menu that appears. Alternatively, go directly to the URL *http://www.arabidopsis.org/tools/bulk/microarray/index.jsp*.
2. Enter (by typing) or upload (using the Browse button) a list of array element names in one of the appropriate boxes.
3. From among the Search Against radio buttons, choose the array platform (e.g., AFGC, Affymetrix 8K or 25K array).
4. Select the Text radio button under "Output type:" and click Get Microarray Elements to submit the query.

   Alternatively, checking the HTML option displays the results in the Web browser and includes links to TAIR detail pages (e.g., locus, array element). For the AFGC array data, the HTML results include links to the Spot History in the Stanford Microarray Database (SMD) and to the Expression Viewer tool. The Expression Viewer displays precalculated cluster data from all of the AFGC experiments and can be used to find genes that are similarly expressed across all related experiments.

5. Save the list on the local computer. The list can be opened using a text editor or a spreadsheet program.

   The maximum number of results is 1000 records. To download all annotations for an array, follow steps 6 to 7, below. The annotations provided in this file include simple summary description. Users can also download GO annotations (Basic Protocol 4) for co-clustered genes, which can also be useful in classifying data from microarrays and for preparing figures and tables for publications.

Finding corresponding loci for all elements on an array

6. Complete lists of loci corresponding to array elements can be found on the TAIR FTP site (*ftp://ftp.arabidopsis.org/home/tair/Microarrays/*). Mapping files are available for some common array designs. Choose the appropriate manufacturer and click on the directory name.

   The assignment of locus identifiers to array elements is based upon BLAST analysis of sequence similarity using the AGI transcripts data set (http://arabidopsis.org/help/helppages/microarray_readme.jsp). If an array element (oligo or cDNA) matches to more than one locus, the element is flagged as "ambiguous" and all of the matched loci are included in the results set. In some cases, an element does not map to a locus. A likely reason is that the element resides in a region that has not yet been annotated as a locus. Another possible reason is that not all loci are represented in the transcript data set, e.g., non-coding RNAs.

7. The current files containing array element mappings are listed in the directory along with the date the mapping file was generated. Click on the file name to view the contents. Use the Save As option in the File menu of the browser to save the file to the local computer.

   The downloaded fields are: "array element name," "array element type," "organism," "control (yes or no)," "locus identifier," "one line locus description," "is ambiguous" (if yes, the element potentially maps to more than one locus), "chromosome," "start position," and "end position." The AFGC data also includes summarized expression values. Older versions of the mapping files are located in the "old" directories. The old mapping files are maintained because assignments of array elements to loci may change. These files are maintained in order to be able to trace the history and clarify any discrepancies due to conflicting annotations.

8. While the above data correspond to gene expression arrays, users of Affymetrix GeneChip *Arabidopsis* Tiling 1.0R Array data can find the TAIR9 remapped BPMAP or library file at the FTP site *ftp://ftp.arabidopsis.org/home/tair/Microarrays/Affymetrix/TilingArray1.0R/*. The original probe set was designed for TIGR5 version of the genome build and this file accommodates assembly changes from TIGR5 to TAIR9. It must be noted that except for substitutions, there is no change (insertions/deletions) in genome build from TIGR5 through TAIR8 genome releases. However, the assembly update from TAIR8 to TAIR9 included insertions and deletions in addition to substitutions. The README on the above FTP site explains the relevant files and its contents. This library file will be useful for expression analysis using TAIR9 annotation.

USING THE MOTIF ANALYSIS TOOL FOR IDENTIFYING POTENTIAL *cis*-REGULATORY MOTIFS IN UPSTREAM SEQUENCES

BASIC PROTOCOL 7

The Motif Finder identifies six-oligomer nucleotide sequences that are statistically overrepresented in a set of input sequences when compared to the whole genome. The most common application of this tool is for identifying potential *cis*-regulatory elements in genes whose expression patterns correlate into a cluster. Consensus sequences for putative transcription factor binding sites can be used to identify additional genes having the element in the promoter using the PatMatch program (see Commentary).

Necessary Resources

See Basic Protocol 1

Entering the search parameters

1. On the TAIR home page (*http://www.arabidopsis.org*) select Motif Analysis from the Tools drop-down menu. Alternatively, go directly to the URL *http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp*.

2. On the data entry form enter the locus identifiers of your genes of interest. This can be done manually or by uploading a list of identifiers from a file. In the example shown in Figure 1.11.12A, we queried for motifs in the 500 bp upstream region of 100 genes that are members of the Type I MADS-box gene family. Note that a minimum number of 3 locus identifiers has to be entered.

3. Select length (500, 1000, or 3000 bp) of upstream sequence to be queried. Submit the query.

   The sequence data sets are either 500-, 1000-, or 3000-base-pair sequences upstream of the translation start site of each gene in the genome. The 3000-base-pair option was added recently. The program will search for 6-mer words that are overrepresented in the upstream regions of the set of queried genes compared to upstream sequences in the entire genome. Both forward and reverse strands are queried.

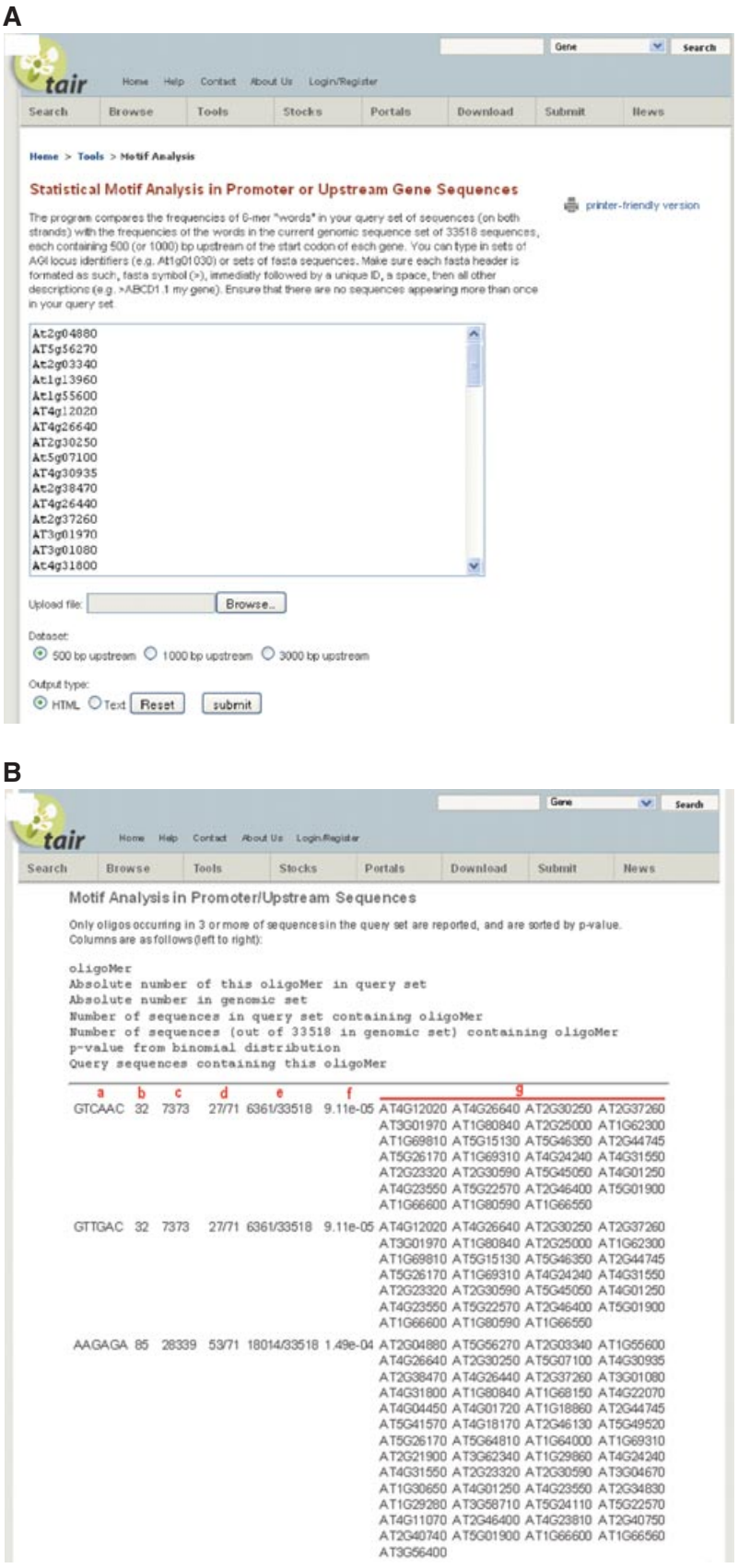

Figure 1.11.12 Motif Finder tool. (**A**) Users can type in or upload a list of genes and select the promoter length to be analyzed. (**B**) The resulting motifs are listed with the corresponding genes in which they are overrepresented.

Evaluating the results

4. The results are displayed in a table as shown in Figure 1.11.12B. The columns, denoted "a" through "g" in Figure 1.11.12B are as follows.

   a. Oligomer.

      Each over-represented six-oligomer sequence is listed in the first column of the results table.

   b. Absolute number of oligos in the query set.

      Number of times the oligo appears in the upstream regions (of chosen length) of the query genes. This number can be higher than the number of query sequences, as some sequences contain multiple occurrences of the motif.

   c. Absolute number of oligos in the genomic set.

      Number of times the oligo appears in the upstream sequences (of chosen length) of all genes in the genome.

   d. Number of sequences in query set containing oligomer.

      Shows the ratio of the number of queried sequences containing the oligomer over the total number of queried sequences.

   e. Number of sequences (out of 33,518 in genomic set) containing oligomer.

      Shows the ratio of the number of genome sequences containing the oligomer over the total number of sequences in the genome.

   f. *p*-value.

      This score reflects the probability of the six-oligomer sequence occurring in the selected query set by chance. The lower the score (closer to zero) the greater the likelihood the match is significant.

   g. Query sequences containing this oligoMer.

      All of the query genes containing the oligomer are listed here. The PatMatch tool can be used to locate other genes that contain the oligomer in the upstream sequence (see Commentary).

OBTAINING INFORMATION ABOUT *ARABIDOPSIS* METABOLIC PATHWAYS

BASIC PROTOCOL 8

AraCyc is a metabolic pathway database specifically for *Arabidopsis thaliana*. The database was initially built using the Pathologic module in the Pathway Tools software developed for MetaCyc (Karp et al., 2002; Mueller et al., 2003). Pathologic predicts possible metabolic pathways based upon the set of annotated enzymes available for a particular species. Following the initial computational build of AraCyc, pathways were manually validated and some were supplemented with additional experimental evidence. The AraCyc database has continued to grow and improve through additional rounds of computational pathway prediction, pathway validation, and manual pathway creation and curation (Zhang, 2005). New releases of the database are generated semiannually (*http://www.plantcyc.org/release_notes/aracyc/aracyc_release_notes.faces*). The latest version, AraCyc 6.0, was released in October 2009. AraCyc includes data concerning pathways, enzymes, reactions, compounds, and genes. Unification links to MetaCyc and PlantCyc facilitate comparison of pathways from different organisms. This protocol describes how to find and interpret pathways using galactose degradation as an example.

Necessary Resources

See Basic Protocol 1

Finding a pathway

1. From any TAIR page, go to the Quick Search tool in the header and select Metabolic Pathways from the drop-down menu. Then enter the name of a pathway (e.g., `galactose degradation`) in the text box and click the Search button.

   Other items, such as enzyme names and compounds, can also be entered in the Quick Search tool. The text entered will be used to match strings for all of the classes of objects in AraCyc. For example, entering "galactose" will return the pathways "UDP-galactose biosynthesis" and "galactose degradation I," the enzyme "galactose-1-phosphate uridylyltransferase," the compound "beta-d-galactose," any reactions that include "UDP-galactose" as a product or reactant, etc.

2. Alternatively, on any TAIR page, select AraCyc Metabolic Pathways from the Tools drop-down menu (see Fig. 1.11.1), or go directly to the URL h*ttp://www.arabidopsis.org/biocyc/index.jsp*.

3. On the main AraCyc page, click on the Search AraCyc link that is found in the main page text. Search AraCyc also appears as the first link on the sidebar. These links go to *http://www.plantcyc.org:1555/ARA/server.html*. In the Query drop-down menu on this page, choose "Pathway (by name)" from the drop-down menu and enter a pathway name (e.g., `galactose degradation`) in the text input box to the right. Click the Submit button.

   Some people have reported problems trying to access AraCyc from behind a firewall. AraCyc uses a nonstandard port that is sometimes blocked. If problems are encountered accessing AraCyc, users should ask the systems administrator at their own institution to unblock port 1555.

4. A results page appears that shows two galactose degradation pathway variants. Click on the desired pathway, such as "galactose degradation I," to view the pathway display page where the compounds shown are red, the reactions are blue, the enzymes are gold, and the genes are purple (Fig. 1.11.13). Enzymes and associated genes that are linked to reactions based on experimental evidence are shown in bold. Enzymes and associated genes that are only linked to reactions based on computational predictions are not shown in bold. Regulatory interactions, such as feedback inhibition of enzymes, are depicted using gray arrows. Yellow plus and minus signs located in gray circles in front of enzymes indicate their activators and inhibitors, respectively. If there are other pathways that feed into or emerge out of a depicted pathway, they are shown using green arrows and links.

   Additional information concerning the pathway is shown at the bottom of the page. This includes alternative names, a brief summary, references, and a hierarchical classification of the pathway. An evidence code icon in the upper right corner of the pathway display page indicates whether the pathway has computational, experimental, or other evidence to support its existence in Arabidopsis thaliana. Click on the icon to display more detailed evidence code information and to access the supporting reference(s).

Examining pathway details

5. When a pathway page is initially opened, it may not display all of the available information. For instance, enzymes, genes, and compound structures may not be shown in the initial view. Use the More Detail button at the top of the page to zoom in on the pathway until the desired level of detail is displayed. At the highest detail level it will become impossible to zoom in any further and the More Detail button will no longer appear.

   Additional parameters for pathway display can be defined by clicking the Customize Diagram button or by electing to Show or Hide predicted enzymes.

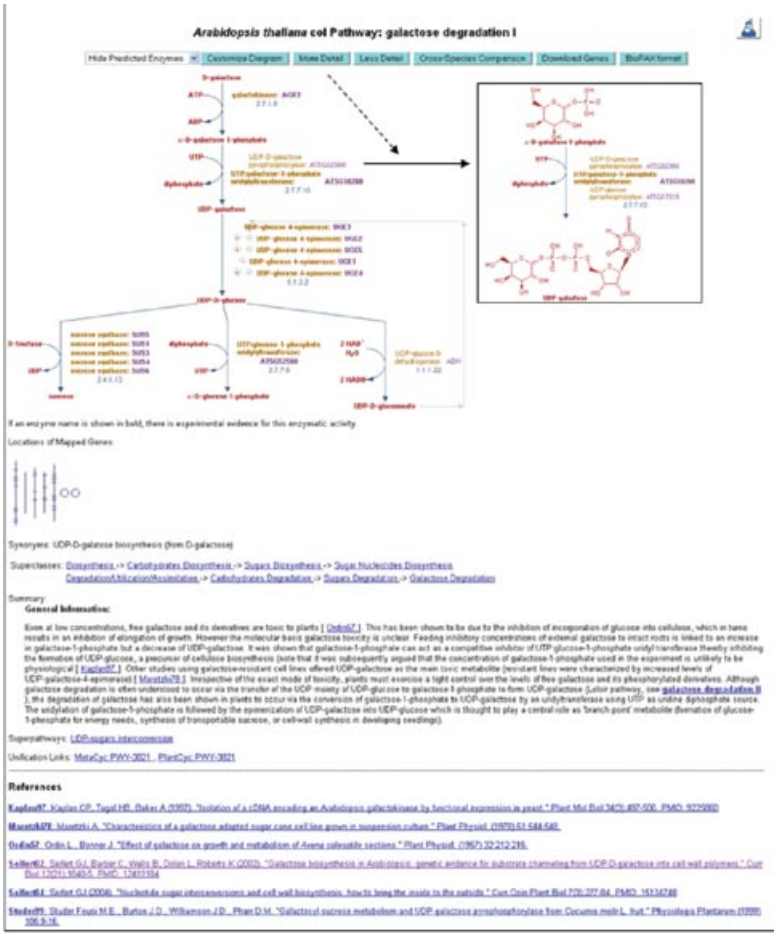

Figure 1.11.13 AraCyc pathway details for galactose degradation I. A pathway schematic shows the reactions (blue), compounds (red), enzymes (yellow), and genes (purple). At the highest zoom level (inset box) compound structures are displayed. Each enzyme linked to a reaction based on experimental evidence is presented in bold-face type. A pathway evidence icon appears in the upper right corner of the diagram. Additional links, a pathway summary, and references are available below the pathway schematic. For the color version of this figure go to *http://currentprotocols.com/protocol/bi0111*.

6. AraCyc pathway diagrams currently do not allow users to visibly distinguish enzymes that act in specific tissue(s) or organelle(s) or to identify where and when reactions or compounds occur. Similarly, intracellular and intercellular transport of compounds is not depicted in pathway pages even though many pathways do involve more than one organelle or tissue. To gain further information about the subcellular localization of enzymes, click on the yellow enzyme name on the pathway page and find the Locations annotation on the enzyme detail page. However, the most up-to-date information about enzyme expression patterns and subcellular localizations can be obtained from the locus page on the TAIR Web site (see Basic Protocol 2). Click on any purple gene name on the pathway diagram to directly open the corresponding TAIR locus page. Alternatively, click on the appropriate locus on the Location of Mapped Genes map displayed below the pathway diagram (Fig. 1.11.13).

   The biochemical activity of some enzymes has been identified even if the underlying gene has not been cloned. In these cases, clicking on the purple gene name will take the user to an AraCyc gene page.

Selecting an aspect of the pathway to display in detail

7. Click on any reaction, compound, enzyme, or gene to find more information. Each data type has a unique detail page.

   On every detail page, the data in AraCyc are extensively interwoven through a network of hyperlinks to other data types. For example, the Gene-Reaction schematic on the enzyme detail page indicates if an enzyme is a monomer or a polymer and provides links to all the subunits of a multimeric complex. The compound detail page includes links to all of the reactions and pathways in which the compound is a substrate or product. Many additional types of data are present and interconnected on the enzyme, reaction, and compound pages.

Finding similar pathways in other organisms

8. Go to the main pathway page (see step 4, Fig. 1.11.13) and find the section called Unification Links under the pathway summary. Click on the PlantCyc link to see a similar pathway diagram that may include data from other plants, including poplar, tomato, rice, and *Medicago truncatula*. Click on the MetaCyc link to access a comparable pathway diagram that may contain enzyme information for a variety of prokaryotic and eukaryotic organisms.

   Depending on the object described on the detail page, other types of links may appear in the Unification Links section. For instance, most enzyme detail pages provide a link to Phytozome (http://www.phytozome.net) to identify potential orthologs in 13 additional plant species. Meanwhile, reaction detail pages may have links to external databases such as BRENDA (http://www.brenda-enzymes.info), and compound pages may have links to resources such as KEGG (http://www.genome.jp/kegg).

9. On the main pathway display page (see step 4, Fig. 1.11.13), click on the Cross Species Comparison button to directly compare the individual reactions and associated enzymes present for each pathway in the different databases hosted by the Plant Metabolic Network.

   AraCyc can currently be directly compared to PlantCyc and PoplarCyc but additional plant databases should be available in the future.

Downloading pathway data files

10. On the main AraCyc page (see step 3), click on AraCyc Pathways to go to an FTP directory or go directly to *ftp://ftp.arabidopsis.org/home/tair/Pathways/*. Click on the file name that begins with `aracyc_dump` to download a file that lists all of the pathways found in AraCyc along with their associated reactions and enzymes.

11. From any pathway page (see step 4, Fig. 1.11.13), download a tab-delimited list of genes, reactions, enzymes and evidence codes by clicking the Download Genes button. Click on the "BioPAX format" button to download an XML BioPax file for the pathway.

Performing advanced pathway searches and data retrievals

12. To search for a specific subset of pathways that meet various criteria and to obtain specific data elements related to those pathways, go to the main AraCyc page (see step 3). Click on the link to the "Advanced Query Page" described in the text section entitled "Other AraCyc Tools." Alternatively, go to the bottom of any pathway page (see step 4) and click on the "Advanced Query Page" button. In the section entitled "Enter your query here" make sure that the database selected is *Arabidopsis thaliana* col.

    Complex queries can be constructed using the dynamic form, and specific data output fields can be selected. Describing the use of the Advanced Query page is beyond the scope of this protocol, but helpful instructions are available by clicking on the "Advanced Query Documentation" link present at the top of the Structured Advanced Query page.

Necessary Resources

Hardware

Computer with Internet access

Software

Up-to-date Web browser
Spreadsheet program (e.g., Microsoft Excel) or text editor

Files

`sample.dat` (available at *http://www.plantcyc.org:1555/expr-examples/sample.dat*)

Preparation of sample files

1. Generate a text file that contains the object name in the first column (referred to as "zeroth" column on the OMICS viewer input page) and numeric values in the subsequent column(s), (referred to as columns 1 to *n* on the input page).

   a. Several types of objects can be entered into the zeroth column:

   i. *Gene names and/or identifiers:* The unique AGI locus identifiers (e.g., AT1G01010) are the best choice. The use of gene names or symbols, e.g., TAR1, is also acceptable but will result in many fewer data points being displayed, because not all gene symbols are currently included in AraCyc.
   ii. *Protein names and/or identifiers:* AGI locus identifiers (e.g., AT1G01010) are the best choice.
   iii. *Compound names and/or identifiers:* All common chemical names are acceptable, but chemical compound names can be extremely difficult to match, even though efforts are made to enter synonyms in the database. A file listing all of compounds found in AraCyc pathways can be used to pre-screen the input file to identify name discrepancies (*http://www.plantcyc.org/downloads/data_downloads.faces*). AraCyc compound IDs such as CPD-714 can also be used in the input file.
   iv. *Reaction identifiers and/or EC numbers:* AraCyc reaction IDs such as RXN-1024 can be obtained by downloading the full database (*http://www.plantcyc.org/downloads/data_downloads.faces*). For EC numbers, only full EC numbers are allowed and no data will be displayed if the same EC number is associated with more than one reaction.

   b. The numeric values in the subsequent columns may represent:

   i. absolute values of transcript expression levels, metabolite abundance, etc.
   ii. relative values (e.g., log ratio of two values) of transcript expression levels, proteomic fragment expression levels, etc.
   iii. qualitative numbers used to represent different categories, e.g., all enzymes with EV-EXP (experimental) evidence = 1 and all enzymes with EV-COMP (computational) evidence = 0.

2. If the file is generated using Microsoft Excel, save the file as Text (Tab delimited), with the extension `.txt` using the Save As option under the File menu. If the file is generated using other software, make sure that the resulting file is a tab-delimited text file with the suffix `.txt`. However, files with the extension `.dat` are also accepted.

Displaying gene expression, proteomic or metabolomic data

3. On any TAIR page, select AraCyc Metabolic Pathways from the Tools drop-down menu (see Fig. 1.11.1), or go directly to the URL h*ttp://www.arabidopsis.org/biocyc/index.jsp*. On the main AraCyc page, click on the OMICS Viewer link that is found in the main page text. "OMICS Viewer for experimental data overlay" also appears as the second link on the sidebar of the AraCyc main page. These links go to *http://www.plantcyc.org:1555/ARA/expression.html.*

4. On the OMICS Viewer page, to overlay data from a file on an *Arabidopsis* metabolic map, find the field labeled "File containing expression data (NOT a URL)." Click the Browse button and navigate to the text file prepared in steps 1 to 2. Click "Open." Alternatively, use the sample file provided.

   The sample file can be obtained by clicking on the link in the yellow text box on the left side of the page that says "For an example data file, see here" (http://www.plantcyc.org:1555/expr-examples/sample.dat).

5. Next, select the parameter to properly classify the type of data being displayed.

   a. To display a static image or an animation of absolute expression levels, choose Absolute and "a single data column."

   b. To display a static image or an animation of relative expression levels, choose Relative.

   i. If the relative expression level for each object has been pre-computed by the user and entered as a single value, select "a single data column."

      Example: Gene A has absolute levels of expression of 100 under condition Y and 200 under condition Z. The user calculates the ratio of expression of Z/Y and enters `2` *in a single data column*

   ii. If the absolute expression levels for each object are listed in different columns and the relative value needs to be calculated by the OMICS tool, select "the ratio of two data columns."

      Example: Gene A has absolute levels of expression of 100 under condition Y and 200 under condition Z. The user enters `100` *in column 1 and* `200` *in column 2, and directs the OMICS Viewer to calculate the ratio using the Data column boxes (see step 8 below).*

   iii. "Relative" and "a single data column" are the appropriate settings for the practice file.

      Although there are multiple columns of data, each one already has a computed ratio.

6. Next, choose the correct scale for displaying the data.

   a. Use the "0-centered scale" if positive and negative numbers appear in the data set (e.g., if a 2-fold up-regulation is represented as 2 and a 2-fold down-regulation is represented as -2). This option can commonly be used to display log scale-based data.

   b. Use the "1-centered scale" if there are no negative numbers and reduced levels are represented using decimal values (e.g., if a 2-fold up-regulation is represented as "2" and a 2-fold down-regulation is represented as "0.5").

      Please note that any negative values entered will be discarded if "1-centered scale" is selected.

   c. "0-centered scale" is the appropriate scale for the data set in the sample file that uses negative numbers to represent reductions in expression level.

7. Use the drop-down menu next to "The items in the first (zeroth) column of your data file are" to identify the type of data input entered.

   As mentioned in step 1b, the OMICS Viewer can be used to display data for genes, proteins, compounds, reactions, or "Any of the above." If "Any of the above" is selected, it is possible to combine, for example, gene expression and metabolomics data into a single display. There are some dangers inherent in this approach. Some names may be ambiguous if the program cannot determine what type of object they represent. In addition, data values from different kinds of experiments may not be directly comparable, so the resulting diagram may be misleading in some important ways.

   It is particularly difficult to display both gene and protein data on the same OMICS viewer display simultaneously. The OMICS viewer only shows the names and specific data values attributed to genes and not to proteins. Any protein values entered will only be displayed as color-coded, but unlabeled lines if they are entered using their FRAME-IDs (e.g., `AT1G01010-MONOMER`*). However, they will be displayed using the gene name if they are entered just using their AGI locus identifier. Please note that it is not possible to simultaneously enter data for a gene and a protein using the same AGI locus identifier.*

8. Next, use the two text boxes in the section entitled "Single Experiment Time Step or Animated Time Series" to direct the program to use data from the correct columns in the input file.

   IMPORTANT NOTE: *the first column adjacent to the identifiers is called column 1 by the program. So, data entered into column B on an Excel sheet are considered to be present in column 1.*

   a. To display a static image of absolute data points or precomputed relative values, enter the number of the single desired data column from the input file into the "Data column (numerator in ratios)" box.

      For example, enter `1` *to display the first column of data or* `2` *to display the second.*

   b. To display an animation of absolute data points or precomputed relative values, enter a list of numbers corresponding to the desired columns from the input file into the "Data column (numerator in ratios)" box.

      For example, using the sample data file that has data for six time points in successive columns, enter `1`, `2`, `3`, `4`, `5`, *and* `6` *on separate lines to display all six relative data values in a time series animation.*

      The animation feature can be used to display time series data, but it can also be used to show data associated with a series of different mutants collected in a series of different tissues, etc.

   c. To display a static image of relative values that are computed by the program, enter the number of the single data column that should serve as the numerator into the box on the left called the "Data column (numerator in ratios)" box. Enter the number of the column that should serve as the denominator in the box to the left titled "If using two columns, denominator data column."

      For example, if time point 1 is in column one and time point 2 is in column two, enter `2` *in the box to the left and* `1` *in the box to the right to show the relative increase or decrease in data values from time 1 to time 2.*

   d. To display an animation of relative values that are computed by the program, enter the number of all the data columns that should serve as numerators in the box on the left called the "Data column (numerator in ratios)" box. If all of the columns will be divided by the same column, enter the number of that single column that should serve as the denominator in the box to the left titled: "If using two columns, denominator data column." However, if any of the numerators should be divided by different denominators, then, enter a series of paired values of numerator and dominator using the pair of boxes.

For example, enter `2 3 4 5 6` *in the box on the left and* `1` *in the box on the right to see a 5-step animation of the following ratios: "2/1" "3/1" "4/1" "5/1" "6/1." In contrast, enter* `1 2 3 4` *in the box on the left and* `5 5 6 6` *in the box on the right to see a 4-time-point animation of the following ratios: "1/5" "2/5" "3/6" "4/6."*

9. Adjust the parameters at the bottom of the Color Scheme section to change the display of the results.

   a. To use the full color spectrum to represent all of the values in the input file, use the default option ("Full color spectrum," computed from data provided).

      This option makes it difficult to compare directly between files that have very different quantitative data ranges. For example, if the maximum value of induction of a transcript is 30-fold in one experiment and 3-fold in another, both would be displayed using the same peak color of red.

   b. To display only data that appears below a specified threshold choose one of the two latter options.

      The "Full color spectrum with a maximum cutoff" option allows users to directly compare across experiments with different data ranges. The "Three color display with specified threshold" option uses three primary colors to represent high (red), medium (blue), and low (green or yellow) expression in bins based on the range of the values in the input file. Therefore, the color schemes can change for each input file if the range of the values is different. If too many values fall above the threshold and are displayed in red, the parameters can easily be changed.

10. Use the final set of parameters to select the data display options.

    a. Keep the default "Paint data on cellular overview chart" option selected to view the data in the context of metabolic pathway diagrams.

       To see a report of pathways that contain at least one element that is above (or below the inverse/negative of) a specified threshold, click on the "Generate a table of individual pathways exceeding threshold" option and enter the threshold value.

       If a value of 4 is selected, then any pathways that contain at least one object with a data value above "4" or below "1/4 (0.25)" will be listed if a 1-centered scale is used. Any pathways with at least one data value above "4" or below "-4" will be listed if a 0-centered scale is used.

    b. Select the "Paint data on cellular overview chart" to display the quantitative data based on the relative positions of the genes in the genome.

       This option may be less informative in Arabidopsis, where metabolic operons appear to be a very rare occurrence (Field and Osbourn, 2008), than in prokaryotes.

Displaying the results

11. Click the Submit button. A new image is returned displaying four types of data. The sample data set, `sample.dat`, has multiple time points. Data values from the initial time point are shown in Figure 1.11.14. The various parts of the display are as follows:

    a. The overview diagram.

       This overview depicts the full set of metabolic pathways and orphan reactions present in the AraCyc database. Related pathways are grouped within a lighter shaded box. Compounds are represented using shapes, and genes, proteins, and reactions are represented using the lines between the compounds. Quantitative data are overlaid on these elements using the color scheme selected in step 9.

       In many cases, several genes or enzymes, each with their own expression level, will map to a single reaction. This is because the reaction might be catalyzed by an enzyme complex made up of several gene products, or by several isozymes encoded by different genes.

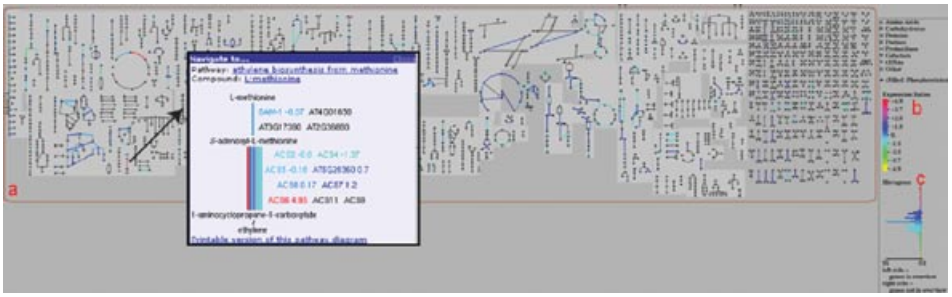

Figure 1.11.14 AraCyc OMICS Viewer Pathway Overview. Basic Protocol 9, step 11 provides details about (a) the overview diagram, (b) the color key, the basic data statistics (not shown), and (d) the histogram. Pop-up windows displaying more detailed information and links to AraCyc pathway and compound pages can be opened by clicking on any compound shown on the overview diagram. A pop-up window of the ethylene biosynthesis pathway showing data values for specific isozymes is superimposed on the overview diagram. For the color version of this figure go to *http://currentprotocols.com/protocol/bi0111*.

However, to simplify the display, only a single line with one color is depicted for each reaction on the overview diagram. For absolute expression levels, the maximum is selected. For relative expression levels, the value whose log has the greatest deviation from zero is selected, under the assumption that a user is primarily interested in identifying the genes whose expression levels differ most between the two data sets. But all of the individual values for the isozymes can be seen by clicking on any of the pathways displayed in the overview. A pop-up window will appear that shows multiple lines for different enzymes performing the same reaction (see Fig. 1.11.14). The quantitative data associated with each isozyme is also listed next to each gene.

b. The color key for the overview.

This shows the correspondence between colors and numeric values.

c. Basic statistics from the input file.

These include, e.g., (1) gene, compound, protein names, etc., that could not be resolved or, that are not included in AraCyc; (2) total number of data values displayed; and (3) minimum, maximum, and median values, and mean and standard deviation of the natural logs of the values of the input data. These statistics are not computed when generating animations, and therefore are not shown in Figure 1.11.15.

d. A histogram showing the distribution of values in the data set.

This is displayed directly beneath the color key. The data value range is divided into 50 intervals, using the same criteria used for assigning colors. The number of data values in each interval is shown on the histogram, colored appropriately. To the left of the vertical axis is the histogram for the genes that are represented in the overview. To the right of the axis is the histogram for all other genes.

Understanding the color scale

12. If the "Full color spectrum, computed from data provided" option was chosen in step 9, a maximum cutoff value is computed from the data supplied in the uploaded file. The minimum cutoff value is determined based on the maximum cutoff value and the other parameters. For absolute expression levels, the minimum cutoff value is zero. For relative expression levels that are not logs, the cutoff is the inverse of the maximum cutoff. For relative expression levels that are logs, the minimum is the negative of the maximum cutoff. The color spectrum is then mapped evenly along a log scale between the maximum cutoff and the minimum cutoff.

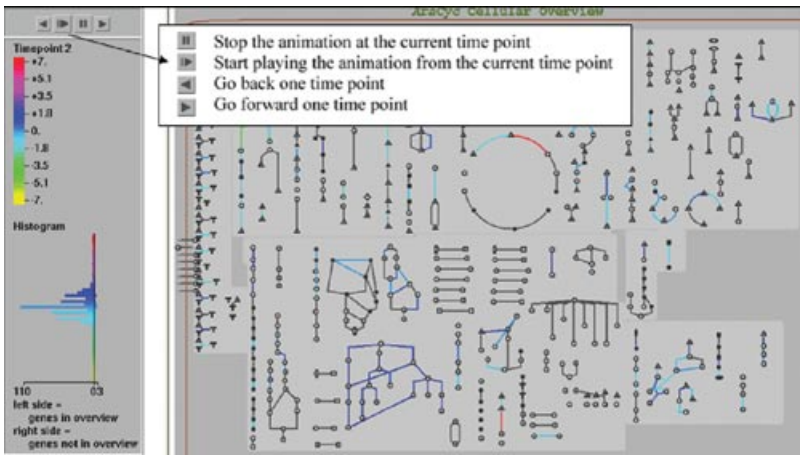

Figure 1.11.15 Four buttons control the display of an OMICS viewer animation. The buttons can be used to stop and restart the animation, and step through the individual "slides."

Using animation controls

13. An expression time series can be displayed as an animation by specifying multiple data column numbers (see step 8). The result is a dynamic HTML page that initially plays the animation in a continuous loop, showing how the expression values and histogram change with each experiment. Four buttons control the animation; their functions are listed in Figure 1.11.15.

    Older browsers that do not support dynamic HTML will not be able to run the animation. The basic statistics (step 11c) shown on static OMICS viewer data displays are not shown for animations; however, a histogram (step 11d) is shown for each different "slide" in the animation series.

Finding pathways with elements that are induced/increased or repressed/reduced

14. Select the "Generate a table of individual pathways exceeding threshold" during the data input phase (see step 10a to i, above) to identify pathways containing elements with associated data points that meet a specific quantitative criterion.

    The pathways and all of their associated genes will be displayed below the basic statistics information. This option is only available when displaying a static data set.

15. Alternatively, use the color key information and scroll over the page to identify pathways that exceed or fall below desired thresholds.

Navigating from the OMICS viewer to other elements in the database

16. To obtain more information about each pathway, click on a compound from the pathway on the OMICS viewer overview. A pop-up box displaying the compound names and the names of all of the isozymes catalyzing each reaction will be superimposed over the OMICS viewer display (Fig. 1.11.14). The pop-up window has links to pages with additional information about the pathway, the specific compound selected, and, if applicable, any superpathways that contain the selected subpathway.

GUIDELINES FOR UNDERSTANDING RESULTS

General Considerations for Using TAIR

As with any Web-based resource, some general guidelines should be observed when interpreting results. Databases are constantly changing; new information is incorporated and interfaces can also change often.

Revisions to the Data in the Database

Over the course of genome annotation, many new genes have been added and existing genes have been made obsolete or updated (split or merged) to reflect new information (Haas et al., 2003; Swarbreck et al., 2008). In 2004, TAIR inherited the responsibility of maintaining the genome sequence and annotation from the former Institute for Genomic Research (TIGR), now the J. Craig Ventner Institute (JCVI), which provided the genome sequence and annotation from 2000 to 2004. The naming convention agreed upon by the AGI for adding new loci and updating existing loci (*http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp*) has been followed by TAIR. Users are encouraged to submit structural annotation updates to TAIR and to provide the relevant sequence to GenBank or TAIR (*http://www.arabidopsis.org/submit/gene_annotation.submission.jsp*). Changes in sequence annotation may affect the association of genes to related data such as protein domains, polymorphisms, and homologies. For example, domains associated to a locus that was subsequently split may then be associated to only one of the two resulting loci. The locus history, shown on the top of the locus page (Fig. 1.11.14), summarizes all of the changes that had been made to the locus. The locus history can also be searched independently by locus name using the Locus History Search (*http://www.arabidopsis.org/tools/bulk/locushistory/index.jsp*). For many data sets in the FTP site, TAIR maintains older versions of the data. Users should always note the date or version information associated with any data files, such as BLAST data sets or GO annotations.

Evidence Codes in GO Annotations

When interpreting Gene Ontology annotations, it is essential to understand the process of annotation and the importance of evidence codes in interpreting the annotations. The GO Consortium has developed a set of evidence codes (The Gene Ontology Consortium, 2010) as a way of quickly assessing the strength of the assertion made in the annotation. In TAIR, annotations include an evidence description, in addition to the evidence code (Berardini et al., 2004). The evidence description is a set of controlled vocabularies that describe the type of experimental or computational evidence in greater detail. For example, an annotation having the evidence code "inferred from mutant phenotype" (IMP) may be further specified by including more specific information about the type of experiment done such as "RNAi experiments." Since more than one gene may be affected by RNA interference, the phenotype may be due to changes in expression of multiple loci. Thus the GO annotation should be viewed with the understanding that the phenotype may be due to the loss of function of more than one homologous locus. When no information is found in the available published literature, annotations are made to the terms "unknown process," "unknown function," or "unknown cellular component." Such "unknown" annotations indicate that at the time of annotation no information was available for the associated gene. In contrast, a gene lacking annotations altogether might have available data but has not yet been curated. At TAIR, GO annotation is an ongoing process; annotations are updated as new information about genes is published (Berardini et al., 2004). Each annotation has an associated date, which refers to the date the annotation was made. Registered users can also submit their own functional annotations to TAIR (*http://www.arabidopsis.org/submit/functional_annotation.submission.jsp*). The attribution for these annotations is linked to the submitter's personal profile in TAIR.

Evidence Codes in AraCyc

AraCyc uses a similar system of evidence codes to indicate the evidence supporting pathways and reactions (*http://bioinformatics.ai.sri.com/evidence-ontology/*). The evidence codes for the existence of a pathway in AraCyc are indicated on the pathway overview page. One or more evidence icons are shown, in the upper right corner, depending on

the amount and type of evidence; the icons are hyperlinked to a new page describing the evidence and include any appropriate citations. The galactose degradation pathway described in Basic Protocol 8 (Fig. 1.11.13) has an icon of a flask, which indicates that the pathway was experimentally determined. The source of the evidence (namely, the papers containing experiments that were used to describe the pathway) is linked to the evidence detail page. Each pathway diagram also has a color-coded pathway evidence glyph displayed on the bottom of the page that displays more information about the evidence supporting individual reactions in the pathway. Green lines indicate reactions in which the enzyme is present in *Arabidopsis*. A black line signifies that a reaction is known to occur in other organisms but the enzyme that catalyzes the reaction in *Arabidopsis* has not been identified. If the enzyme is present in *Arabidopsis* and catalyzes a reaction that is unique to a pathway, the line is orange. Spontaneous reactions, such as those occurring in polymerization pathways, are indicated with magenta lines. If one's favorite pathway is not yet in AraCyc, one is encouraged to contact the curators to request the addition of the pathway. A list of updates and planned additions is posted on the AraCyc home page (*http://www.arabidopsis.org/tools/aracyc/*). Users are also encouraged to report errors or omissions in the database.

COMMENTARY

Background Information

TAIR was originally a collaborative project between biologists at the Carnegie Institution, Department of Plant Biology, and computer scientists at the National Center for Genome Resources, initiated in 1999. In 2004, the Carnegie Institution resumed all responsibilities for TAIR. TAIR is the third incarnation of an *Arabidopsis* community database after AAtDB (An *Arabidopsis thaliana* Database, which continued from 1991 to 1994) and AtDB (*Arabidopsis thaliana* Database, which continued from 1994 to 1999; Flanders et al., 1998; Rhee et al., 1999). TAIR arose out of the need to accommodate genomic data such as the genome sequence, gene annotations, and integration of physical and genetic maps, in the context of the experimentally verified data in the literature. In 2000, TAIR merged with the *Arabidopsis* Information Management System (AIMS), which was a database for the *Arabidopsis* stock center (*Arabidopsis* Biological Resource Center; ABRC; Scholl et al., 2000, 2003). This resulted in an increase in the number of registered users of TAIR, currently 19,722, making TAIR arguably the largest model organism database to date. From 2003 to 2006, TAIR accommodated genome-wide expression data, starting with cDNA microarray data from the first functional genomics effort in *Arabidopsis*, the *Arabidopsis* Functional Genomics Consortium (*http://www.arabidopsis.org/portals/masc/AFGC/index.jsp*). Therefore, TAIR is the primary source for finding experimentally derived information about genes and proteins from the literature, seed and DNA stocks, genome sequence and annotation, and gene expression data from microarrays. Consult the on-line documentation for up-to-date statistics of the database content (*http://www.arabidopsis.org/servlets/processor?type=tablestats*), sources of data in TAIR (*http://www.arabidopsis.org/about/datasources.jsp*), and usage statistics (*http://www.arabidopsis.org/usage/*).

Design principles and current limitations

TAIR has been designed and built as a Web tool to allow researchers to access all of the data housed in TAIR using a standard Web browser such as Internet Explorer or Mozilla Firefox. It is built upon industry standards for database management systems, software architecture, and software design (Weems et al., 2004). TAIR is primarily designed by biologists and, although the authors try their best to create their interfaces with biologists in mind, it has not always been possible to arrive at solutions that meet every user's requirements. A certain amount of familiarity with *Arabidopsis* and with basic concepts of molecular genetics and plant biology is assumed. Consequently, the breadth of information on the home page and myriad options on the search interfaces can be daunting to a novice user. More experienced users and developers may be frustrated by the difficulty in obtaining the entire database for retrieving specialized, custom data sets.

Keeping up to date with TAIR and Arabidopsis research

There are two ways to keep updated on the *Arabidopsis* research scene. Registered users can choose to receive a quarterly e-mail newsletter from TAIR that describes significant new or updated data and tools. In addition, researchers that use *Arabidopsis* or *Arabidopsis* data frequently can subscribe to the newsgroup (*http://arabidopsis.org/news/newsgroup.jsp*), which is a moderated mailing list through which meetings, jobs, and new data/tools are announced and where researchers post problems and get feedback.

TAIR News and Job Postings are now also available as RSS feeds. Subscribe by clicking on the RSS icon on the Home, Breaking News, and Job Postings pages. An alternative way of staying connected with TAIR is to become a fan of TAIR—via the *Arabidopsis* Information Resource on Facebook (*http://www.facebook.com*)—or receive tair_news twitter feeds (*http://twitter.com/tair_news*).

Alternatives to TAIR

While there are no complete alternatives to TAIR, there are several Web sites that provide significant amount of *Arabidopsis* data and alternative ways of viewing them. They can be grouped into two categories: sites that provide alternative views of the genome and sites that specialize in subsets of data that are either not yet accommodated or covered in less depth in TAIR. All of these sites are linked extensively from TAIR, whereby the sites in the former category are linked from each locus detail pages and sites in both categories are listed and updated in the TAIR Portal pages (*http://www.arabidopsis.org/portals/index.jsp*). There are five major sites that offer alternative views of the *Arabidopsis* genome. First, Nottingham *Arabidopsis* Stock Center (NASC, *http://arabidopsis.info*) provides information about seed stocks and the genome using the Ensembl viewer (Birney et al., 2004). In addition, NASC provides over 1000 Affymetrix chip data and a number of visualization and analysis tools for exploring the gene expression data (Craigon et al., 2004). SIGnAL (*http://signal.salk.edu/*) from the Salk Institute offers a genome viewer (T-DNA Express, *http://signal.salk.edu/cgi-bin/tdnaexpress*) that is decorated with all of the T-DNA and transcript data that are generated from Salk and other laboratories around the world. Often, SIGnAL displays data that are not yet displayed at TAIR; therefore, it is a good idea to check this site to get the latest mapping of T-DNA insertions and cDNA clones. This site also offers gene expression data from the recent work on genome tiling arrays (Yamada et al., 2003). The former Institute for Genome Research (TIGR), now JCVI (*http://www.jcvi.org* and Munich Information of Protein Sequences (MIPS, *http://mips.helmholtz-muenchen.de/plant/athal/index.jsp*) contributed significantly to the *Arabidopsis* genome annotation and both offer views of the genome. In addition, AtGDB (*Arabidopsis thaliana* Genome Database, *http://www.plantgdb.org/AtGDB/*) offers another view of the genome that has been annotated with their gene-prediction algorithms. For most genes, the description and exon-intron structures of genes in these sites are identical. However, in a small number of cases, there are genes that have different descriptions and/or structures because of the differences in the methods of annotation and interpretation of the evidence. Users should pay attention to the evidence shown for the gene structures and make their own interpretation of the structure and function of the genes. There are also many sites that provide detailed information about a subset of genes of *Arabidopsis* such as chromatin remodeling factors, transcription factors, and small RNAs. TAIR tries to maintain up-to-date links to these resources from the TAIR Portal pages. Please contact TAIR by e-mail (*curator@arabidopsis.org*) if there are missing or nonfunctional links.

Alternatives to AraCyc

There are additional resources available on-line that provide access to *Arabidopsis* metabolic pathway information. One of the most widely used is the KEGG Pathway database (*http://www.genome.jp/kegg/pathway.html*). Although KEGG catalogs information for many different species, it is possible to select *Arabidopsis*-specific information highlighted on the reference metabolic pathways (*http://www.genome.jp/kegg-bin/show_pathway?org_name=ath&mapno=01100&mapscale=1.0&show_description=show*). A more comprehensive set of links to external databases, data sets, and tools that pertain to metabolic pathways, enzymes, and metabolites is available through the Metabolomics portal at TAIR (*http://www.arabidopsis.org/portals/metabolome/index.jsp*).

Additional tools at TAIR

In addition to the tools discussed in the protocols, TAIR hosts several other useful analysis tools which are briefly described below.

N-Browse: A molecular interaction viewer

N-Browse (*UNIT 9.11*), a molecular interaction viewer, is now available at TAIR and allows users to visualize protein interactions of their *Arabidopsis* genes of interest. All protein interactions shown in TAIR N-Browse are based on experimental methods and have been reviewed by curators from TAIR, BioGrid, and/or IntAct. Interactions are color coded based on their methods of detection, and links to the original papers that reported the interactions are provided. N-Browse was developed and implemented for TAIR by the Kris Gunsalus Lab at New York University. N-Browse can be accessed at *http://www.arabidopsis.org/tools/nbrowse.jsp*.

Synteny viewer: GBrowse_Syn

GBrowse_syn is a GBrowse-based synteny browser designed to display multiple genomes, with a central reference species compared to several additional species. It is included with the standard GBrowse package (version 1.69 and later). GBrowse_syn was built to help researchers study and analyze syntenic regions, homologous genes, and other conserved elements between sequences. It can also be used to study genome duplication and evolution. By comparing newly sequenced or less studied genomes to the well annotated *Arabidopsis* genome in GBrowse_syn (*http://gbrowse.arabidopsis.org/cgi-bin/gbrowse_syn/arabidopsis/)* scientists can identify novel genes and putative regulatory elements.

The first version of the GBrowse_syn tool at TAIR includes the genomes of *A. thaliana*, *A. lyrata*, and *P. trichocarpa*. Additional plant genomes will be added to this synteny browser in the near future. The *A. lyrata* and *P.trichocarpa* alignment data were provided to us by Pedro Pattyn from the University of Ghent.

Textpresso

Textpresso is an information extracting and processing package for biological literature. Textpresso for *Arabidopsis* (*http://www.textpresso.org/arabidopsis*) allows users to search all abstracts and over 15,700 full-text publications in TAIR. Keyword searches can be narrowed by searching in specific categories. Textpresso was initially developed by Hans-Michael Muller, Eimear Kenny, and Paul W. Sternberg, with contributions from Juan-Carlos Chan and David Chen. This new version, Textpresso 2.0, was developed by Hans-Michael Muller with contributions from Arun Rangarajan and Tracy K. Teal.

Bulk data retrieval tool

The bulk data retrieval tool allows the user to retrieve specific types of data in bulk for a list of genes (*http://www.arabidopsis.org/tools/bulk/index.jsp*). Data types that can be obtained using this tool include gene descriptions, GO annotations, DNA and protein sequences, protein information, locus histories and microarray elements.

Restriction analysis tool

The restriction analysis tool allows the user to perform a restriction analysis based on the arbitrary DNA sequence submitted (*http://www.arabidopsis.org/cgi-bin/patmatch/RestrictionMapper.pl*). This program was written and made available to TAIR by Dr. Shuai Weng at AtDB.

Chromosomal map tool

This tool (*http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp*) allows the user to map genes on top of the five *Arabidopsis* chromosomes using a list of locus names (e.g., At1g01010). The list should contain one locus name per line. The resulting image, which displays the location of the queried list of genes on the five chromosomes, can be saved in a variety of formats.

Patmatch

PatMatch (*http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl)* was designed for identifying patterns in a selected dataset (e.g., TAIR9 genes, TAIR9 proteins, upstream sequences, etc.) that match regular expressions. PatMatch can be useful for finding short nucleotide patterns such as *cis*-elements, Massively Parallel Signature Sequence (MPSS), Serial Analysis of Gene Expression (SAGE) tags, or small RNA binding sites. Patmatch can also be used to search for motifs in protein sequences. Users can input regular expressions that include mismatches, insertions, and deletions, and apply standard IUPAC notation to indicate ambiguous sequences. Other options of this tool include the selection of a target data set, strand to be queried (in case of nucleotide search), and number of results to retrieve. The program uses the same target data sets as TAIR's BLAST and FASTA software. PatMatch does not generate alignments or provide scores for best hits. The output is provided

as a summary table of the results followed by a detailed list of all the hits. The summary table indicates the number of hits, sequences with hits, and queried sequences, and provides an option to generate a text-formatted file of all the hits. The detailed list of all the hits displays the following data:

Sequence name: Name of the gene or sequence for which a hit was found.

# of hits: Number of times the query pattern was found in that specific sequence.

Hit pattern: Pattern used for the query.

Matching positions: Start and end position of the hit. These coordinates are always relative to the sequence (e.g., gene, upstream region, intergenic region. . .).

Hit sequence: Hyperlink to the sequence for which a hit was found. The pattern match is highlighted in red letters. For nucleotide searches, coordinates shown here are always relative to the chromosome.

If one needs to process large amounts of data or increase the number of results to be included, it is possible to download the PatMatch1.1 program *(ftp://ftp.arabidopsis.org/home/tair/Software/Patmatch/)* and run it locally on a Unix-based system. The BLAST data sets used by PatMatch can also be downloaded from the TAIR FTP site (*ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR9_blastsets/*).

Critical Parameters and Troubleshooting

No data found

A frequently reported problem is that searches do not retrieve any results. There are many reasons why this occurs. The simplest explanation is that the data are not in the database. To find out if the sought after data type(s) are in TAIR, consult the database statistics page (*http://arabidopsis.org/servlets/processor?type=tablestats*) linked from the bottom of the home page. Often the data are in TAIR but not found because of problems arising from poorly formed queries or improper use of the search forms. The temptation to fill out all of the optional fields in the advanced searches can generate too many restrictions that limit the scope of the data retrieved. Use fewer, rather than more options. Another reason why searches fail is that the data are not accessible through the existing search interfaces. The categories under the Advanced Search section of the Web site (*http://arabidopsis.org/servlets/Search?type=general&action=new_search*) list data types that can be searched. If interested in obtaining data that are in TAIR (from the DB statistics list), but not accessible through any of the advanced searches, send an e-mail to the curators to request the data.

Too much data found

While "no data found" is probably the most common problem encountered, retrieving too many results can also be a problem. There are two ways to handle this problem: (1) using the advanced searches and restricting parameters to retrieve a subset of the results, or (2) manipulating the results set to select a subset of data. Restricting the search parameters can be done on all the Advanced Search pages and detailed help on using these parameters is available (*http://arabidopsis.org/help/helpcontents.jsp*). Large results sets can be downloaded and reformatted to explore the data more efficiently. All of the search results can be downloaded as tab-delimited text files (see Basic Protocol 2). The results can be imported into software like Excel that allows manipulations such as sorting, reordering, reformatting columns, and graphing the results. For some search tools, such as the Expression Search, reordering the results by different columns can be performed directly on the Web site.

Layers of connected data that are hidden

TAIR's database structure exploits the relational database design and each data type has a high degree of association to other data types. This network of associated data is not easily represented in a two-dimensional, tabular format via hyperlinks. Consequently, associated data may be separated by two or more hyperlinks. For example, currently, polymorphisms that are associated to a locus are listed on the locus detail page, but in order to see any pictures associated with mutants that carry these polymorphisms, it is necessary to first click on the polymorphism name to see these details, and then click on the germplasm band on the polymorphism detail page. The associated images of the plant's phenotypes are displayed on the germplasm detail page.

Reporting problems and requests to curators

Perhaps the most important thing to know about troubleshooting TAIR is that users are encouraged to e-mail curators (*curator@arabidopsis.org*) and report problems or request data. Often, users that want a particular set of customized data contact the curators, who generate and put the data in the public FTP directory. Also, some of the

problems reported reflect problems and can change how data are curated or imported into TAIR, and thus contribute to the overall enhancement of the resource.

Advanced parameters

Despite the extensive content of this unit, it still does not cover all of the functionalities and tools that are offered at TAIR. Users familiar with the basic functionalities and who are interested in using more sophisticated tools and exploring the data are encouraged to try some of the advanced tools. These are listed under the Analysis Tools and Help Central sections. One of these tools and data sets is highlighted below.

VxInsight and TreeView

All of the cDNA-based microarray data at TAIR have been filtered, normalized using the Lowess method (Dudoit et al., 2002), and clustered using the VxInsight's topography-based clustering as well as hierarchical clustering. The clustered data can be explored using VxInsight's visualization software and TreeView (*UNIT 6.2*) software, respectively. Both pieces of software are stand-alone programs that need to be installed locally. More information about the data and instructions on how to download the software is found online (*http://arabidopsis.org/tools/bulk/microarray/analysis/index.jsp*). One of the obvious uses of this clustered data is to find genes of interest and look for other genes that have significantly correlated gene expression profiles. TAIR will provide a similar data set of all high-density oligomer arrays in the near future.

Suggestions for Further Analysis

Microarrays

TAIR's integration of microarray data has focused on providing ways to search and download the data. As such, development of tools for mining and analyzing data from the many experiments that are publicly available has not been emphasized. There already are a number of existing tools that have been developed for mining public *Arabidopsis* data. One notable tool is GeneVesitgator (*https://www.genevestigator.ethz.ch/*), which contains most of the publicly available high-density array data from AtGenExpress (*http://arabidopsis.org/info/expression/ATGenExpress.jsp*) and other laboratories, and allows searching and displaying of the data (Zimmermann et al., 2004). Users must log in, after which they can search for genes that are expressed in specific conditions, growth stages, or organs, or for genes of particular interest to them, and get a comprehensive view of the expression profiles in the different environmental conditions, growth stages, and organs. NASCArrays (*http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl*) contains high-density oligomer array data generated by the NASC community Affymetrix array service (*http://affymetrix.arabidopsis.info/*). In addition to the database, NASC Arrays has tools for data mining, including a graphical viewer to display expression levels of each array element from all hybridized arrays (Spot History), a two-gene scatter plot for comparing expression of two genes across all hybridizations, and GeneSwinger, a tool for sorting of experiments based on the extent of expression variability for genes of interest. Finally, MapMan (Thimm et al., 2004; *http://mapman.mpimp-golm.mpg.de/*) is a software system for generating sets of biological domains (bins) for either a metabolic pathway or signal transduction pathway. It allows researchers to overlay gene expression or proteomic data over the user-driven biological domains. GenMapp (Dahlquist et al., 2002; Salomonis et al. 2007; *http://www.genmapp.org/*) is a similar program that is also available for microarray data analysis.

Working with a desktop version of AraCyc

The complete set of AraCyc data can be obtained through a free license agreement (*http://www.plantcyc.org/downloads/license_agreement.faces*) and can be loaded into a stand-alone desktop version of the Pathway Tools software (*http://biocyc.org/download.shtml*). Some additional features related to searching, grouping items, tracing metabolites, creating and altering pathways, and displaying OMICS viewer data are available in the desktop version of Pathway Tools program.

Submitting data to TAIR

One of the most fundamental aspects of science is sharing data and results with the research community. The fruits of research drive new areas of discovery, and funding agencies, such as the National Science Foundation (NSF), have invested heavily in developing community resources. Web sites and databases such as TAIR make these data accessible to anyone connected to the Internet. The long-term sustainability of databases will increasingly rely upon

contributions by the research community (Rhee, 2004).

TAIR encourages feedback and data submission and provides several ways for researchers to contribute their expertise and data. Instructions for submitting various types of data including gene function, interaction partners, expression patterns, markers, phenotypes, and several others, are available on the Submit Overview page (*http://arabidopsis.org/submit/index.jsp*), accessible from the Submit drop-down menu in the top navigation bar. Users can prepare data formatted according to the guidelines or download and use the preformatted Excel spreadsheets. The spreadsheets contain macros that ensure that the proper data formats are used. To use the spreadsheets, macros must be enabled. TAIR will also accept direct submissions by email to *curator@arabidopsis.org* for small datasets and corrections to existing data, as well as very large datasets and those requiring special formats. Please contact us with any questions about data submission.

In addition, each data detail page includes a Comments section where additional information can be added by community members; click on the comment text to view the entire comment. Registered users can submit comments that are then immediately displayed in the Comments section of the detail page. On-line instructions for submitting comments are available at *http://arabidopsis.org/help/helppages/addcomment.jsp*.

Submitting data to AraCyc

AraCyc gratefully accepts new and updated information from the research community concerning metabolic pathways, compounds, enzymes, and their associated genes. Users are also encouraged to report errors in the database. Data submissions and corrections are recognized on the Contributors page (*http://www.plantcyc.org/about/contributors.faces*).

Several avenues for contacting the curators are available:

a. Send an e-mail to *curator@plantcyc.org*.

b. Click on the Pathway, Enzyme, and Compound Data Submission Forms link on the main AraCyc page or select Data Submission under the Submit Data menu bar item on every AraCyc page.

c. Click on the Feedback menu item at the top of each AraCyc page or the Report Errors or Provide Feedback button on the bottom of each data page.

Acknowledgments

The authors of this unit are grateful for the continued support of members of the research community who share their expertise, ideas, and criticisms—all of which improve TAIR and PMN immensely. Drs. Tanya Berardini, Donghui Li, and Peifen Zhang are to be thanked for their helpful comments, along with all of the curators and programmers who make TAIR work. This work is supported by the National Science Foundation (TAIR grant DBI-0850219 and PMN grant DBI-0640769), and the National Institutes of Health (Gene Ontology grant 5P41HG002273-0955000650). This unit is Carnegie Institution for Science—Department of Plant Biology publication no. 1754.

Literature Cited

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S.Y. 2004. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* 135:745-755.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyras, E., Fernandez-Suarez, X.M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M., and Hubbard, T. 2004. Ensembl 2004. *Nucleic Acids Res.* 32:D468-D470.

Borevitz, J.O. and Nordborg, M. 2003. The impact of genomics on the study of natural variation in *Arabidopsis*. *Plant Physiol.* 132:718-725.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., Chen, H., Frazer, K.A., Huson, D.H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J.R., and Weigel, D. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338-342.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. 2004. NASCArrays: A repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.* 32:D575-D577.

Cutler, S., Ghassemian, M., Bonetta, D., Cooney, S., and McCourt, P. 1996. A protein farnesyl transferase involved in abscisic acid signal transduction in *Arabidopsis*. *Science* 273:1239-1241.

Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., and Conklin, B.R. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31:19-20.

Dudoit, S., Yang, Y.H., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:E15.

Eulgem, T., Rushton, P.J., Robatzek, S., and Somssich, I.E.. 2000. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* 5:199-206.

Field, B. and Osbourn, A.E. 2008. Metabolic diversification: Independent assembly of operon-like gene clusters in different plants. *Science* 320:543-547.

Flanders, D.J., Weng, S., Petel, F.X., and Cherry, J.M. 1998. AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the *Arabidopsis* Genome Initiative. *Nucleic Acids Res.* 26:80-84.

Garcia-Hernandez, M., Berardini, T.Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Rhee, S.Y., Scholl, R., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. 2002. TAIR: A resource for integrated *Arabidopsis* data. *Funct. Integr. Genomics* 2:239-253.

The Gene Ontology Consortium. 2010. The gene ontology in 2010: Extensions and refinements. *Nucleic Acids Res.* In press.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., and White, O. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31:5654-5666.

Hagen, G. and Guilfoyle, T. 2002. Auxin-responsive gene expression: Genes, promoters and regulatory factors. *Plant Mol. Biol.* 49:373-385.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., and Rhee, S.Y. 2001. The *Arabidopsis* Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29:102-105.

Karp, P.D., Paley, S., and Romero, P. 2002. The pathway tools software. *Bioinformatics* 18:S225-S232.

Mueller, L.A., Zhang, P., and Rhee, S.Y. 2003. AraCyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiol.* 132:453-460.

Pearson, W.R. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* 4:1145-1160.

Rhee, S.Y. 2004. Carpe diem: Retooling the publish or perish model into the share and survive model. *Plant Physiol.* 134:543-547.

Rhee, S.Y., Weng, S., Bongard-Pierce, D.K., Garcia-Hernandez, M., Malekian, A., Flanders, D.J., and Cherry, J.M. 1999. Unified display of *Arabidopsis thaliana* physical maps from AtDB, the *A.thaliana* database. *Nucleic Acids Res.* 27:79-84.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. 2003. The *Arabidopsis* Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31:224-228.

Running, M.P., Lavy, M., Sternberg, H., Galichet, A., Gruissem, W., Hake, S., Ori, N., and Yalovsky, S. 2004. Enlarged meristems and delayed growth in plp mutants result from lack of CaaX prenyltransferases. *Proc. Natl. Acad. Sci. U.S.A.* 101:7815-7820.

Salomonis, K., Hanspers, A.C., Zambon, K., Vranizan, S.C., Lawlor, K.D., Dahlquist, S.W., Doniger, J., Stuart, B.R., and Pico, A.R. 2007. GenMAPP 2: New features and resources for pathway analysis. *BMC Bioinformatics* 8:217.

Scholl, R.L., May, S.T., and Ware, D.H. 2000. Seed and molecular resources for *Arabidopsis*. *Plant Physiol.* 124:1477-1480.

Scholl, R., Sachs, M.M., and Ware, D. 2003. Maintaining collections of mutants for plant functional genomics. *Methods Mol. Biol.* 236:311-326.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* 12:1599-1610.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. 2008. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* 36:D1009-D1014.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y., and Stitt, M. 2004. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37:914-939.

Weems, D., Miller, N., Garcia-Hernandez, M., Huala, E., and Rhee, S.Y. 2004. Design, implementation, and maintenance of a model organism database for *Arabidopsis thaliana*. *Comp. Funct. Genomics* 5:362-369.

Wortman, J.R., Haas, B.J., Hannick, L.I., Smith, R.K. Jr., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., White, O.R., and Town, C.D. 2003. Annotation of the *Arabidopsis* genome. *Plant Physiol.* 132:461-468.

Yalovsky, S., Kulukian, A., Rodriguez-Concepcion, M., Young, C.A., and Gruissem, W. 2000. Functional requirement of plant farnesyltransferase during development in *Arabidopsis*. *Plant Cell* 12:1267-1278.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S.X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H.L., Tripp, M., Chang, C.H., Lee, J.M., Toriumi, M., Chan, M.M., Tang, C.C., Onodera, C.S., Deng, J.M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A.D., Gurjal, M., Hansen, N.F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V.W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P.X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E.K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R.W., Theologis, A., and Ecker, J.R. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302:842-846.

Zhang, P., Foerster, H., Tissier, C., Mueller, L., Paley, S., Karp, P., and Rhee, S.Y. 2005. MetaCyc and AraCyc: Metabolic pathway databases for plant research. *Plant Physiology* 138:27-37.

Ziegelhoffer, E.C., Medrano, L.J., and Meyerowitz, E.M. 2000. Cloning of the *Arabidopsis* WIGGUM gene identifies a role for farnesylation in meristem development. *Proc. Natl. Acad. Sci. U.S.A.* 97:7633-7638.

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