16

Using Information From Public *Arabidopsis* Databases to Aid Research

Margarita Garcia-Hernandez and Leonore Reiser

Summary

The volume of *Arabidopsis* information has increased enormously in recent years as a result of the sequencing of the genome and other large-scale genomic projects. Much of the data are stored in public databases, where data are organized, analyzed, and made freely accessible to the research community. These databases are resources that researchers can utilize for making predictions and developing testable hypotheses. The methods in this chapter describe ways to access and utilize *Arabidopsis* data and genomic resources found in databases.

Key Words: Data mining; database; genomics; gene expression; bioinformatics; computational biology; *Arabidopsis*.

1. Introduction

Technological advances have fostered a new era in *Arabidopsis* research, giving us a well-annotated, sequenced genome ***(1)*** complementing a rich body of literature. More recently, international functional genomics initiatives (http://arabidopsis.org/info/workshop2010.jsp) have ignited a new wave of research to decipher the functions of every gene in the genome, and eventually, to understand what it takes to make a flowering plant. Large amounts of data about gene expression, metabolism, and protein and gene interactions are being generated by these projects. To accomplish the task of organizing and managing the data, groups and individual labs have created databases to store the information generated and make it available to the research community.

Scientists doing research in this "postgenomic" area are compelled to know how to make use of databases to extract the relevant information needed to further their research. The protocols in this chapter describe how to use databases to find what is known about *Arabidopsis* and to make inferences and predictions that can later be tested experimentally. Each protocol includes a summary of the rationale, a brief description of the database/tool(s), and the specific steps for querying, retrieving, and interpreting the data. The protocols, along with the corresponding databases and tools, are outlined in **Table 1**. This table of contents can be used to find specific protocols of interest within the chapter. When the methods described can be applied to databases other than the ones described below, the database is shown in the table along with a brief description. The databases described here represent a small portion of the vast collection of databases and bioinformatics resources available on the Internet. The protocols described here use tools and data available in databases as of the spring of 2004.

2. Materials

Programming experience is an asset to a scientist who wishes to analyze and manipulate complex, large datasets, but it is not essential to effectively mine databases. Anyone with access

From: *Methods in Molecular Biology, vol. 323:* Arabidopsis *Protocols, Second Edition*
Edited by: J. Salinas and J. J. Sanchez-Serrano © Humana Press Inc., Totowa, NJ

Table 1
Databases and Tools that Can Be Used to Perform Methods Described in This Chapter

Database: Tool	URL	Protocol/Description
Searching literature databases		
NCBI PubMed Database	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed	Finding relevant articles in the NCBI PubMed database
TAIR Publication Search	http://arabidopsis.org/servlets/Search?action=new_search&type=publication	Finding Arabidopsis publications using TAIR's Publication Search
Finding information about genes in Arabidopsis genome databases		
TAIR: Gene Search	www.arabidopsis.org/servlets/Search?action=new_search&type=gene	Finding gene information by name
TAIR GO Annotation Bulk Download and Analysis	www.arabidopsis.org/tools/bulk/go/index.jsp	Finding functional information about genes
TIGR: Gene Search	www.tigr.org/tdb/e2k1/ath1/ath1.shtml	Gene search by locus identifier, description, name
MatDB: Gene Search	http://mips.gsf.de/proj/thal/db/search/search_frame.html	Gene search using locus identifiers, BAC or cosmid clone names, keyword, MIPS codes, and BAC-based name or description
NCBI: Gene Search	www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene and www.ncbi.nlm.nih.gov:80/mapview/map_search.cgi?taxid=3702	Gene search can be limited by name, locus identifier, and description and species. Or search and browse the Arabidopsis reference genome database.
MIPS FuncCat Database	http://mips.gsf.de/proj/funcatDB/search_main_frame.html	Find genes with similar functions using MIPS functional categories
Gene Ontology Consortium	www.geneontology.org	Find genes with similar functions in Arabidopsis and other organisms
Searching DNA microarray data		
TAIR: Microarray Expression Search	www.arabidopsis.org/servlets/Search?action=new_search&type=expression	Finding the expression patterns of genes in different microarray experiments

(continued)

NASCArrays: Spot History	http://affymetrix.arabidopsis.info/narrays/spothistory.pl	Finding the expression history of a gene across all microarray experiments
TAIR: Microarray Experiment Search	http://arabidopsis.org/servlets/Search?type=expr&search_action=new_search	Search experiments and download data
TAIR: Microarray Element Search	http://arabidopsis.org/tools/bulk/microarray/index.jsp	Search array elements.
		View precomputed gene clusters
NASC Arrays	http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl	Search experiments and download data from experiments conducted at the NASC microarray facility. Several data mining tools
NCBI Gene Expression Omnibus	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo	Find microarray and SAGE experiments and gene expression profiles. Clustering analysis, and visualization tools
Array Express	http://www.ebi.ac.uk/arrayexpress/	Search microarray experiments, protocols and arrays
Stanford Microarray Database	http://genome-www.stanford.edu/microarray/	Find, analyze, and download microarray data
Genevestigator	www.genevestigator.ethz.ch/	Displays patterns of expression of genes from Arabidopsis Affymetrix datasets in the context of plant organ, growth stage, or stress response
MapMan	http://gabi.rzpd.de/projects/MapMan/	Displays Arabidopsis Affymetrix datasets onto diagrams of metabolic pathways or other processes
Searching massively parallel signature sequencing data		
MPSS: Simple Search	http://mpss.udel.edu/at/	Searching massively parallel signature sequencing data
Searching expressed sequence tags		
NCBI UniGene Search	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene	Using NCBI's UniGene to find ESTs for a sequenced gene
TIGR Gene Index Database	www.tigr.org/tigr-scripts/tgi/libtc.pl?db=atest	Using ESTs to find genes that are differentially expressed
AtGDB	www.plantgdb.org/AtGDB/prj/ZSB03PP/	EST mapping and clustering
Arabidopsis metabolic pathways		
AraCyc	www.arabidopsis.org/tools/aracyc	Finding metabolic pathways, reactions, enzymes, and compounds

(continued)

Table 1 *(Continued)*
Databases and Tools that Can Be Used to Perform Methods Described in This Chapter

Database: Tool	URL	Protocol/Description
AraCyc Expression Viewer	www.arabidopsis.org:1555/expression.html	Detecting changes in the expression of genes involved in metabolism
Kegg	www.genome.ad.jp/kegg/pathway.html	Search and browse metabolic pathways, regulatory pathways, and molecular complexes
Finding related protein sequences in Arabidopsis		
TAIR: WU-BLAST	www.arabidopsis.org/wublast/index2.jsp	Finding similar protein sequences
TAIR: Bulk Protein Download	http://tigrblast.tigr.org/er-blast/	Finding similar sequences
MIPS: BLAST	http://mips.gsf.de/proj/thal/db/search/search_frame.html	Finding similar sequences
NCBI: BLAST	www.ncbi.nlm.nih.gov/BLAST/	Finding similar sequences
NCBI: Blink	www.ncbi.nlm.nih.gov/	Precomputed pairwise similarity search of Arabidopsis proteins against all proteins in GenBank
From genotype—finding mutants for analysis of gene function		
SIGnAL T-DNA Express	http://signal.salk.edu/cgi-bin/tdnaexpress	Finding knockout mutations in a gene
TAIR: Germplasm Search	www.arabidopsis.org/servlets/Search?action=new_search&type=germplasm	Finding other types of mutations and mutants with similar phenotype
NASC: Catalogue Search	http://nasc.nott.ac.uk/catalogue.html	Finding mutants in a gene of interest
TAIR: WU-BLAST	www.arabidopsis.org/wublast/index2.jsp	Using BLAST to find T-DNA/transposon mutations in a specific gene or genes of interest
NASC: Insert BLAST	http://atensemble.arabidopsis.info/Multi/blastview	Using BLAST to find T-DNA/transposon mutations in a specific gene or genes of interest
AtIDB	http://atidb.org/cgi-perl/blast	Using BLAST to find T-DNA/transposon mutations in a specific gene or genes of interest
From phenotype to genotype—databases and tools for map-based cloning		
TAIR: Genetic Marker Search	www.arabidopsis.org/servlets/Search?action=new_search&type=marker	Finding and downloading sets genetic markers for mapping
TAIR: Polymorphism/Allele Search	www.arabidopsis.org/servlets/Search?action=new_search&type=polyallele	Finding polymorphisms between two ecotypes for generating new markers
TAIR: SeqViewer	www.arabidopsis.org/servlets/sv	Finding candidate genes in a genetically defined interval
TAIR: Monsanto Polymorphism Collection	http://arabidopsis.org/Cereon/index.jsp	SNPs and In Dels between Col and Ler Ler sequence
MASC SNP database	www.mpiz-koeln.mpg.de/masc/	SNP data for 12 ecotypes

The third column contains the subheading for the exact's protocol name described in detail in the text, or a brief description of other alternative tools that can be used to perform similar tasks.

to the Internet and a reasonably up-to-date computer should be able to perform all the steps in the protocols. A basic familiarity with computers, Internet browsers, and commonly used bioinformatics tools such as BLAST is assumed. There are a wide variety of textbooks, manuals and Web-based tutorials available for learning the basics of bioinformatics.

2.1. Computer Hardware and Software for Database Mining

The minimum requirements for database mining are a personal computer (PC), an Internet connection, and Web browsing software. Internet connection speeds are usually the rate-limiting step in web browsing. Because the amount of data being accessed from databases may be very large, it is desirable to have access through a high-speed network such as Ethernet, cable, or digital subscriber line (DSL) connection. To access databases, Web browser software, such as Internet Explorer, Netscape, or Safari, is required. Database interfaces should behave the same regardless of what operating system or browser is used. However, some functions may not work properly on older browsers. If possible, you should upgrade your browser to the most recent version available that can run on your operating system.

2.2. Databases

Databases are information storage and retrieval software systems. Typically, databases have three components: the database software for storing data, software that translates and executes requests (queries), and software applications that allow users to view data. *Arabidopsis* researchers have access to myriad databases that are either entirely devoted to *Arabidopsis* ("*Arabidopsis*-specific") or include *Arabidopsis* data along with information about other organisms. **Table 2** lists some of the main *Arabidopsis* databases and the types of information they contain. These databases are described briefly in the next section (**Subheading 2.2.1.**). There are many more *Arabidopsis* databases containing a variety of data types that are not included in the table. Links to these resources can be found at The Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org/info/2010_projects/Resources.jsp and http://www.arabidopsis.org/links/index.jsp). Some of these databases are about specific types of information, such as *cis*-regulatory elements or gene and enhancer traps, whereas others focus on specific classes of genes or disseminate data from a functional genomics project. A significant amount of *Arabidopsis* data can also be found in databases that contain information on other organisms, such as the National Center for Biotechnology Information's (NCBI) GenBank (http://www.ncbi.nlm.nih.gov/), the European Bioinformatics Institute's (EBI) InterPro (http://www.ebi.ac.uk/interpro/), PlantGDB (http://www.plantgdb.org/), and UK CropNet (http://ukcrop.net/).

2.2.1. Arabidopsis *Databases*

The databases described below are some of the primary sources of information about *Arabidopsis* and seed and DNA stocks. All these databases contain gene and protein data; however, they differ significantly in the breadth of information stored, the annotation methods employed, and the search tools and formats for displaying data. **Table 2** summarizes the main types of data found in each of these databases. Choosing which database to use depends on the task at hand and how well the database meets these needs in terms of content and accessibility. It can be both valuable and frustrating to have multiple places and ways to access data, particularly when different sources provide overlapping and sometimes inconsistent or conflicting information (*see* **Subheading 3.1.2.**).

2.2.1.1. Munich Information Center for Protein Sequences Arabidopsis thaliana Database (MatDB)

MatDB (www.mips.biochem.mpg.de/proj/thal/db/index.html) ***(2)*** contains structural and functional data about genes and proteins. Genes can be searched by locus identifier, Open Reading Frame (ORF) name (a nomenclature based on BAC or cosmid names), key words, and sequence similarity (BLAST). Results can be displayed on the graphical genome browser or in

Table 2
Types of Data and Tools Found at Some Primary Public Arabidopsis Resources

Types of Data/Databases	TAIR/ ABRC	TIGR	MIPS MatDB	AtGDB	AtIDB	SIGnAL	NASC
Alleles and polymorphisms	X						
Clones	X						
DNA stocks	X						
Full-length cDNAs	X	X	X	X		X	X
Gene expression—microarrays	X	X	X	X		X	X
Gene expression—Northerns, *in situ* localization	X						
Gene families	X	X	X	X			X
Gene structural annotations	X	X	X	X	X	X	X
Gene ontology annotations	X	X			X	X	
Genetic maps	X						X
Genetic markers	X						X
Genome browser	X		X	X	X	X	X
Metabolic pathways	X						
Nucleotide sequences	X	X	X	X	X	X	X
People and labs	X						
Protein localization	X	X	X				
Protein sequences	X	X	X	X		X	X
Protein structure annotation	X	X	X	X			X
Publications	X						
Seed stocks	X						X
Sequence analysis tools	X	X	X	X	X	X	X
T-DNA/transposon insertions	X				X	X	X

The left column shows a list of some major data types and tools that can be found in the public *Arabidopsis* databases listed in the top row.

table format. MatDB uses Functional Catalogue (FunCat) controlled vocabulary to classify gene functions. Protein families can be found using the MIPS in-house algorithm, SESAM, which can detect distantly related proteins.

2.2.1.2. The Institute for Genome Research (TIGR) *Arabidopsis thaliana* Annotation Database

TIGR's Arabidopsis Annotation Database (www.tigr.org/tdb/e2k1/ath1/) includes structural and functional annotations of all sequenced genes. Genes can be searched by keyword (description), locus name, location on a BAC or cosmid clone, or by sequence similarity (Wu-BLAST). Gene data are summarized in text and graphical formats. TIGR presents alignments supporting gene structures (primarily full-length cDNAs and expressed sequence tags [ESTs]) along with the annotation for validating gene structures. Along with TAIR, TIGR has been annotating gene functions using controlled vocabularies developed by Gene Ontology (GO) Consortium. TIGR has also analyzed and classified the Arabidopsis proteome into families (TIGRFams) that include newly identified groupings that may be specific to Arabidopsis or plants in general ***(3)*** (www.tigr.org/TIGRFAMs/index.shtml).

2.2.1.3. The *Arabidopsis* Information Resource and the *Arabidopsis* Biological Resource Center (ABRC)

TAIR strives to be a comprehensive resource for *Arabidopsis* ***(4–6)***. Like TIGR and MIPS, TAIR includes data for sequenced genes and also includes information about nonsequenced genes (genetic loci) and a wider variety of other data types (**Table 2**). As part of the GO Con-

sortium, TAIR collaborates in developing the vocabularies to accommodate plant genes and annotates *Arabidopsis* gene products ***(7)***. Data can be accessed through a variety of text-based searches and visualization tools such as the SeqViewer Genome browser. Sequence analysis tools are also provided. TAIR is also the gateway for finding and ordering seed and DNA stocks from the ABRC (www.biosci.ohio-state.edu/~plantbio/Facilities/abrc/abrchome.htm).

The ABRC distributes seed in North America and DNA stocks worldwide. Stocks can be searched and ordered using the DNA and Germplasm search tools, or browsed through the Web catalog (http://arabidopsis.org/servlets/Order?state=catalog).

2.2.1.4. Nottingham *Arabidopsis* Stock Center (NASC)

NASC (http://nasc.nott.ac.uk/) provides seeds within the European community. NASC also maintains the recombinant inbred (RI) map for the Lister and Dean Columbia-Landsberg *erecta* mapping population, which is used for genetic mapping of mutations. Seed stocks are accessible via text searching or through a catalog browser (http://nasc.nott.ac.uk/catalogue.html). NASC has recently expanded the scope of its database to include a wide variety of genomic data (*see* **Table 2**).

2.2.1.5. Salk Institute Genomic Analysis Laboratory (SIGnAL) T-DNA Express Gene Mapping Tool

SIGnAL T-DNA Express Gene Mapping Tool (http://signal.salk.edu/) ***(8)*** was initially developed as a tool for researchers to quickly find genes with T-DNA insertions in transgenic lines generated by the SSP T-DNA project. SIGnAL has become a more general resource for reverse genetics and now incorporates flanking sequences from insertion lines developed by other projects, full-length cDNAs and other data (**Table 2**). Genes can be searched by name, sequence, description, or GO annotation, and are displayed in a whole-genome view based on TIGR's tiling path. Gene structural annotations are imported from TIGR. Links for ordering T-DNA line stocks are also provided, as well as tools for designing primers for sequencing the T-DNA insertions.

3. Methods

A primary objective of database mining for most researchers is to find out everything that is known about a specific gene or set of genes. Some of the basic questions are: What does my gene encode? In what biological processes is it involved? With what other genes/proteins does it interact? In what tissues is it located and how is it regulated? In order to generate a testable hypothesis and design meaningful experiments, the current available knowledge must be obtained and analyzed.

The published literature continues to be the primary medium for reporting experimental data and is the most comprehensive resource. The peer review process employed by journals and the inclusion of experimental methods along with the results ensures that data quality can be easily assessed and the experiments can be reproduced if necessary. Publishing in journals is a slow process, articles are typically accessible only to subscribers, not all data produced can be published, and the sheer volume of information makes it difficult to synthesize. Public databases fill an important niche in the chain of data and hypothesis-driven experimentation. Databases can provide large amounts of data quickly and without restrictions. Many include data that would otherwise not be published. The nature of the Internet and databases also facilitates the integration of information from multiple sources through hypertext links. A disadvantage of databases is that data may be incomplete and data accuracy and quality can sometimes be difficult to assess. The distinction between information that is well supported by experimental evidence and what is inferred or predicted is not always clearly presented in genomic databases. To be maximally useful, databases must always include enough information so that a researcher can evaluate the quality and methods behind the data. Likewise, researchers need to approach the data with a healthy skepticism and consider the methods employed in order to evaluate the data with the appropriate confidence.

3.1. Searching Literature Databases

Researchers have published a wealth of data about all aspects of *Arabidopsis* physiology, biochemistry and development. Databases such as PubMed, Agricola, or BIOSIS index articles from a wide variety of journals and can be used to find citations and articles in electronic or print format.

The National Center for Biotechnology Information (NCBI's) PubMed (www4.ncbi.nlm.nih.gov/PubMed/) is the primary database for life-science literature. At the beginning of 2004 the number of *Arabidopsis* publications in PubMed totaled 12,798. PubMed has a powerful search interface and links to the rest of databases within the NCBI system, such as sequence and expression databases. Other useful features include MyNCBI (Cubby), which can store searches that users run regularly to check for new items retrieved. Full-text copies can be ordered, provided one is a registered user. PubMed records are linked to publishers' sites for access to the full text of the article. For help using the resource refer to the PubMed tutorial (http://www.nlm.nih.gov/bds/pubmed_tutorial/m1001.html).

TAIR compiles bibliographic records about *Arabidopsis* from PubMed, BIOSIS, and Agricola. In addition, TAIR includes publications not found in these databases, such as abstracts from the *Arabidopsis* Conferences, defunct *Arabidopsis* electronic journals (The *Arabidopsis* Information Service and Weeds World), books, and dissertations. In early 2004, publications in TAIR totaled 21,982 records. Publication records may be linked to authors' community profiles at TAIR, which facilitates finding contact information.

The following protocols describe how to find *Arabidopsis* publications in PubMed and TAIR.

3.1.1. Finding Relevant Articles in the NCBI PubMed Database

1. Start at the PubMed search page (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed).
2. Enter the desired term(s) in the text input box. Searches can be restricted using the Boolean operators AND, OR, and NOT to combine terms. To search for a phrase, it must be enclosed in quotes (e.g., "transcription regulation"), or with a special flag "[tw]" (e.g., "transcription factor [tw]"). Use wild-card characters (*) for inexact matching. For example, to find all the articles about all the Agamous-like genes, type in "AGL*." For more refined searching, the Preview/Index function provides a guided search that can be used to match each term to specific fields such as author, title, or abstract. After making your selections, click on "Preview" to view the number of articles that match your query. If too many results are found, additional limits can be imposed. To complete the search click on the "Go" button.
3. Finding the article text and saving citations: The default display format is a summary of the citation. The complete citation, including available abstracts, can be viewed by clicking on the author's names. Articles that are available online are linked to the publisher's Web sites, which may be freely accessible or require a subscription. To modify the display of results, select the appropriate option from the display menu. For example, to import a citation into reference management software, choose MEDLINE format. References can be saved into a file for downloading or sent to an e-mail address. After selecting the articles by clicking on the checkboxes alongside the citations, choose the desired option under the "Send to" menu and click on the "Send to" button.

3.1.2. Finding Arabidopsis Publications Using TAIR's Publication Search

1. Start at TAIR's Publication Search page (http://arabidopsis.org/servlets/Search?action=new_search&type=publication).
2. To search with a specific name or phrase, enter the desired terms in the text query boxes and choose the field to search from the drop-down menu (author, title, abstract or title, URL for electronic publications, journal, or book title). For example, to search for all publications about oxidative stress, type the phrase into the text box and select "Title/Abstract" in the drop-down menu. Unlike the PubMed search, quotes are not required; all text in a single box is treated as a phrase. To restrict the search by publication dates or publication type, fill in the corresponding boxes.
3. Click on the "submit" button to start the search.

Results are displayed in a summary format including the title, journal, authors, and year. The title is hyperlinked to a page containing the complete citation, links to authors' TAIR profiles, the abstract, if available, and a list of associated key words and genes.

3.2. Finding Information About Genes in Arabidopsis Genome Databases

When searching for genes it is helpful to be aware of the different types of gene names and problems associated with *Arabidopsis* nomenclature. In *Arabidopsis*, experimentally defined genes have been named using gene symbols based on the gene product's function (e.g., DFR for dihydroflavanol 4-reductase) or mutant phenotype (e.g., TT3 for Transparent Testa3). In addition, all the genes identified by the genome sequencing effort have been assigned a systematic name based on the chromosomal location (e.g., AT2G27150; www.arabidopsis.org/info/guidelines.jsp). These AGI codes, or locus identifiers, uniquely define a locus. One source of frustration is that different databases can sometimes have inconsistent or conflicting information about the same gene. After the *Arabidopsis* genome sequence was completed, TIGR was funded to reannotate the complete genome. This effort resulted in the consecutive release of five genome versions. In each release new loci have been added and many existing loci have been modified (e.g., gene models have been merged or split into two gene models) ***(3,9,10)***. TAIR has incorporated each different genome version released by TIGR, and has now replaced TIGR in the task of maintaining and updating the annotation of the genome. MIPS has also been involved in the annotation of the *Arabidopsis* genome independently using its own computational methods for genome and proteome analysis (PEDANT) ***(2,11,12)***. As a result, there have been some gene structure and naming discrepancies among TIGR, TAIR, and MIPS. To help follow the history of each gene identified by TIGR and MIPS use the Locus History tool available at TAIR (http://arabidopsis.org/tools/bulk/locushistory/index.jsp).

3.2.1. Finding Gene Information by Name

This protocol describes how to use TAIR's gene search to find genes by name, GenBank accession or description.

1. Start at TAIR's Gene Search: www.arabidopsis.org/servlets/Search?action= new_search&type=gene.
2. Define the name search criteria. Select name from the "Search Name" drop-down menu. This option is used to search by symbolic names (e.g., ABI3), full names (e.g., Abscisic Acid Insensitive 3) or locus identifier (e.g., AT3G24560). The name search also includes aliases.
3. Choose an exact or inexact search mode. When searching with a gene symbol choosing the "starts with" option is a way to find similarly named genes, such as members of a gene family. When searching with a GenBank accession, it is better to use an exact match in order to avoid retrieving spurious results. To search for a word or phrase within a gene description, choose the "contains" option.
4. Select the output format. The default values are 25 records, sorted by name. The position option can be used when finding genes by location.
5. Click "Submit Query" to start your search.
6. Finding information about the gene: The results display locus and gene model names that link to detailed information pages. The locus detail page collects and displays all the information associated to each locus and should be the starting point for finding comprehensive information about a gene.

Locus information includes a description, list of alternative names (aliases) for the locus, the chromosomal coordinates, and the date the record was last updated. The last item is useful as a means to assess the currency of the data. Locus information includes sequences and gene/protein structure and function. Clicking on the link from any of the sequences in the "Nucleotide Sequence" section will retrieve the sequence. Functional information about the gene may include Gene Ontology annotations, metabolic pathways, gene expression, and protein localization data. Other associated information may include genetic markers, alleles, polymorphisms,

transcript clones that map to the locus, publications, and people working on the gene or contributing data. The structure of the gene, along with any objects that map to the locus, such as transcripts, T-DNA/transposon insertions, markers, and polymorphisms can be displayed graphically by clicking on the link to the SeqViewer genome browser. Detailed information about all associated data can be obtained by clicking on the name of the object (e.g., polymorphism/allele name).

3.2.2. Finding Functional Information About Genes

To make data about a gene's function more amenable to computational methods of querying and analysis, many databases use structured controlled vocabularies for annotating gene products. As described above, both TIGR and TAIR have actively annotated *Arabidopsis* genes using vocabularies developed by the Gene Ontology (GO) Consortium (www.geneontology.org). The GO vocabularies describe three aspects of a gene product: molecular function (the biochemical activity of a gene product), biological process (the ordered assembly of more than one molecular function), and cellular component (location within the cell) ***(13)***. These vocabularies have been widely adopted by many model organism databases and are considered to be the standard for functional annotation. At TAIR, each annotation to a GO term includes a description of the evidence for the association and links to reference(s) supporting the annotation ***(7)***. Genes for which no experimental or computational data are available are annotated to the term "unknown" to distinguish them from genes that have not yet been annotated. As of January 2004, more than 27,000 loci (approx 78% of the genome) have been annotated to one or more GO terms ***(7)***. Of those, only about 3600 genes have been annotated based on experimental information extracted from the literature. The rest of the annotations are based on sequence similarity to known genes or from inferences based on computational predictions.

GO annotations provide an efficient mechanism for finding functional information for a gene or to find relationships within groups of genes, such as members of a gene family or clusters of genes with similar expression profiles. The following protocols describe how to retrieve GO annotations to classify sets of genes and find genes with related functions.

1. Start at the TAIR GO Annotation Bulk Download and Analysis tool (www.arabidopsis.org/tools/bulk/go/index.jsp)
2. Input the locus identifiers in the query box. Type, paste, or upload a file containing your list of locus identifiers.
3. Define the output options. Select HTML to view the results hyperlinked. Choose text for saving the results as a file.
4. To obtain a list of annotations click on the "Get all GO Annotations" button at the bottom of the page. From this page it is possible to find proteins annotated similarly in other organisms. To find other *Arabidopsis* genes with similar annotations and to see a term's definition click on the name of the locus, scroll to the Annotation section, and click on the term's name. The GO vocabularies are extremely detailed and include well over 12,000 terms that are organized into hierarchies. A simpler ontology (GO Slim), consisting only of the higher-level terms that places genes into broader categories, is also available.
5. To retrieve a group of genes classified into broader categories, click on the "Functional Categorization" button. Genes with specific functions, such as negative regulation of transcription, will be grouped according to broader category of the process of transcription. Keyword category refers to the type of GO term (Biological Process, Cellular Component, and Molecular Function). The second column corresponds to the GO Slim term that defines the broad category. Frequency refers to the total number of annotations (associations of genes to terms) that appear in the set of genes.
6. Visualizing the data as a graph: The functional classification table data can be transformed into a pie chart by clicking on the "Pie Chart" button at the top of the results page. The results are displayed as three separate pie charts, each grouping annotations for one of the three GO categories. The chart includes a key that shows the individual GO Slim terms, the percentage of total annotations represented by the term, and the total number of annotations to the term (raw value).

3.3. Finding Information About Gene Expression

An important tool for finding functional information comes from the analysis of gene expression. There are many reasons to analyze expression data, such as finding the pattern of expression of a gene in an organism, determining the effect of the environment on the expression of particular genes, or understanding how the expression of one gene affects the expression of other genes.

In *Arabidopsis*, several methods have been applied to study gene expression. Besides the traditional gene-by-gene approach using Northern, reverse transcription-polymerase chain reaction (RT-PCR), *in situ* hybridization, and β-glucoronidase/green fluorescent protein (GUS/GFP) reporter methods, the *Arabidopsis* community has invested substantially in applying large-scale methods that allow monitoring the expression of thousands of genes at once. Among those methods are DNA microarray, massively parallel signature sequencing (MPSS), and serial analysis of gene expression (SAGE). In addition, a large collection of expressed sequence tag (EST) sequences and gene trap lines are available, which can also be used in the analysis of gene expression. Expression data obtained by the use of traditional methods can be found mainly in the literature, whereas data from high-throughput methods are for the most part stored in databases. In this section we describe how to find expression information using public repositories that contain DNA microarray, MPSS, and EST data.

3.3.1. Searching DNA Microarray Data

DNA microarrays are one of the most powerful tools for investigating the expression patterns of thousands of genes in parallel, and it is now a common technique in many *Arabidopsis* labs. The widespread use of this technology has been facilitated by the existence of several public and commercial array designs for *Arabidopsis* and by the establishment of publicly funded projects, such as Arabidopsis Functional Genomics Consortium (AFGC) ***(14)*** and NASCArrays ***(15)***, which offer microarray services at affordable prices. As a result, a vast amount of microarray data has been generated, and a considerable portion of it can be found in public repositories. Some of the databases containing *Arabidopsis* microarray data are shown in **Table 2**. Public *Arabidopsis*-specific microarray databases are found in NASC and TAIR. In addition, some laboratories or projects have developed private databases to store the data they produce, and in some cases have made their data available on the Web. For links to some of those projects, and related microarray information, see http://arabidopsis.org/info/expression/index.jsp. Other microarray public repositories, such as NCBI's Gene Expression Omnibus (GEO), the Stanford Microarray Database (SMD), or ArrayExpress, include microarray data from several organisms, including *Arabidopsis*.

The TAIR microarray database supports both spotted and chip array technologies. TAIR currently houses two-channel cDNA-based arrays from the AFGC project ***(14)*** (115 experiment sets with 515 slides), and is adding Affymetrix-based arrays from AtGenExpress (http://arabidopsis.info/info/masc_annual_june03.pdf) and NASCArrays ***(15)*** (around 3000 chips total). In the near future TAIR will integrate *Arabidopsis* data from ArrayExpress (www.ebi. ac.uk/arrayexpress/) and individual users. TAIR microarray data are subjected to extensive curation and annotation. The result is a consistent representation of experiment and sample information. Also, raw data are renormalized at TAIR to filter out low-quality spots and to facilitate data comparison. TAIR provides the arithmetic mean and standard error measures of gene expression for each gene from replicated spots per array (if available), replicated hybridizations (if available), and across all arrays. Also, mapping of array elements to genes based on BLAST analysis is also provided (ftp://ftp.arabidopsis.org/home/tair/home/tair/Microarrays/AFGC/; ftp://ftp.arabidopsis.org/home/tair/home/tair/Microarrays/Affymetrix/). The renormalized data are accessible through text searching, and both the original and the TAIR-normalized datasets are available for download. Moreover, the AFGC dataset has been subjected to two clustering analyses at TAIR. One analysis clustered all slides against all slides, and the whole dataset is available for download, along with the accompanying visualization software VxInsight (ftp://ftp.arabidopsis.org/home/tair/

home/tair/Software/VxInsight/). In the second analysis, each slide was clustered against all the slides belonging to the same experimental category. These data can be visualized using the Expression Viewer tool.

NASCArrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) is a public repository for Affymetrix-based microarray data generated by the NASC facility as a public service ***(15)*** to the community. It provides access to experiment information and complete datasets are available for download. Query for specific gene expression profiles are not possible, but they can be viewed through a series of data mining tools. "Spot history" allows one to see the pattern of gene expression for each gene over all slides in the database. "Two-gene Scatter" allows the user to see the pattern of gene expression over all slides for two genes as a scatter plot. "Gene Swinger" looks at the experiments in which the expression of a given gene varied most. Samples and experiments are described in detail. At the time of writing, March 2004, NASC Arrays contained 58 experiments, and it is anticipated to produce up to 1000 chips.

3.3.1.1. Finding Expression Patterns of Genes in Different Microarray Experiments

This protocol shows how to use TAIR microarray database to find the expression profiles of a gene or genes in specific experiments.

1. Start at the TAIR Microarray Expression Search (http://arabidopsis.org/servlets/Search?action=new_search&type=expression).
2. Enter the gene name of interest in the query text box and select the appropriate type from the tab. Gene symbols (e.g., ap2); locus identifiers (e.g., At5g01810); array element names (e.g., 39B5T7 or at_34561), or GenBank accessions (e.g., T13762) can be used. To search for more than one gene at a time, upload or paste a list of locus identifiers or array element names in the bigger query text box.
3. Select the output options (optional). This can be used to select the format in which the results will be presented. The default options are 25 records per page, sorting results by fold change, and green/red color scheme to display expression values. Once the results are returned to the browser it is possible to re-sort by other criteria.
4. Limit search by Expression Parameters (optional). Click on the plus sign alongside this heading to display the available parameters. This is an advanced option to restrict the search to only certain expression values. "Absolute expression" allows selection of qualitative absolute expression values, such as "expressed," "not expressed," or "absent." The "expressed" and "not expressed" option retrieves all the signal intensity values higher or lower, respectively, than an arbitrary expression value (350 in both channels for dual-channel arrays). The "absent" option retrieves the unreliable values only. "Relative expression" allows selection of array elements/genes that are differentially expressed, either upregulated or downregulated. The default is any relative value. "Std error" allows selection of the confidence level of the measurement. The lower the value, the most reliable is the expression signal. "Analysis level" allows you to retrieve only the expression measurements of replicate arrays ("replicate sets"), or all the arrays ("slide/chip").
5. Limit search by Experiment Parameters (optional). This is an advanced option to restrict a search to only certain experiments. Click on the plus sign alongside this heading to display the available parameters. Experiments can be searched by name, author, description and key word. Key word searches include experimental variables, plant tissue, experiment category (e.g., biotic treatment), and experiment goal. If no limits are imposed, all the experiments in the database are searched.
6. Limit search by Array Design (optional). This option is used to restrict the search to only certain array designs (e.g., Affymetrix 25K). The default option includes all array designs in the database.
7. Submit the query.

The results display the expression measurements (average fold change and standard errors) for each array element in each slide and/or replicate set. The values are color-coded according to expression level and whether the change in expression was positive or negative. Each record

also includes links to pages with detailed information about the experiments and array elements. The results can be downloaded in text format by clicking the check boxes for the records of interest, or selecting all records by pointing to the "Check All" button, and then clicking "Download."

When interpreting the results obtained from microarray analysis, several issues should be taken into consideration. The first one is related to the complexity of the microarray technology. All the numerous steps involved in microarray analysis, from array elements preparation to array manufacture, RNA preparation, hybridization, image processing, and data normalization, are potential sources of variability that can affect the accuracy of the final observation ***(16)***. Many systematic errors are removed by data transformation and normalization ***(17)***, but issues related to sensitivity and background noise, among others, are inherent to the technology and cannot be resolved. An additional error factor is the potential for cross-hybridization. cDNA-based microarrays have higher potential for cross-hybridization, and thus the intensity signal associated with a spot may be a composite of signals from related genes ***(14)***. Also, because hybridization properties vary from sequence to sequence, averaging results from different sequences (array elements) to obtain expression summaries for one gene can be misleading. Additionally, not all the clones spotted on the arrays are fully sequenced, which together with potential mislabeling of clones, results in the measurements associated to a spot not corresponding to the real gene ***(14)***. For example, 1.5% of EST clones used in the AFGC arrays were found to be mislabeled ***(14)***. Moreover, gene-specific dye effects can also add variability to the dual-channel arrays ***(18)***.

3.3.1.2. Finding Expression History of a Gene Across All Microarray Experiments

One way to estimate the variability of a gene's expression profile across all experiments is by plotting the distribution of expression values in those experiments. Both NASCArrays (http://affymetrix.arabidopsis.info/narrays/spothistory.pl) and SMD (http://genome-www5.stanford.edu/) offer tools designed for this purpose. This protocol demonstrates how to use the NASCArrays Spot History tool to find the expression history of a gene across all experiments in the database.

1. Start at the NASC Spot History page (http://affymetrix.arabidopsis.info/narrays/spothistory.pl).
2. Input the gene of interest in the query box and select the type of name from the drop-down menu (AGI code, symbolic name, or Affymetrix probe set name). If desired, check the corresponding box to plot the *x* axis in log scale.
3. Submit your query by clicking on the "Plot!" button.

This search will return a histogram of the frequency of each signal value bin.

To find detailed information about the experiments that have this expression level, click on a bar in the histogram.

3.3.2. Searching Massively Parallel Signature Sequencing Data

Massively parallel signature sequencing (MPSS) is a powerful technique to quantitatively measure gene expression ***(19)***. The relative abundance of the signatures in a given library represents a quantitative estimate of expression of that gene. The advantages of MPSS is the high precision and sensitivity of the expression levels, allowing the detection of very low expression levels not detectable with microarray nor EST methods. However, the cost is very high, so it is expensive to analyze many treatments. In *Arabidopsis*, only one public project developed in the Meyers lab has used this technique so far, and the data have been made public on the Web (http://mpss.udel.edu/at/). The *Arabidopsis* MPSS data set includes the signatures obtained from sequencing 14 different libraries from different tissue types (e.g., leaf, root, silique) and treatments (e.g., salicylic acid). The data can be queried by locus identifier, key word, and signature sequence, as well as by chromosomal position. Advanced query tools allow restricting the

search to one or more libraries, and a visualization tool is also provided to see pairwise comparisons of five of the libraries. This protocol shows how to search MPSS data with locus identifiers.

1. Start at the MPSS database Simple Query section (http://mpss.udel.edu/at/).
2. Type either the AGI code name (e.g., At2g34560) or paste the sequence in the appropriate box.
3. Submit the query by clicking on the "Get Data" button.

The results show a graphical representation of the gene structure overlaid with all the matching signatures. The signatures are classified according to their position on the DNA strands and the direction and gene features they fall in, and are color-coded accordingly to the class. A table showing all the signatures that matched the gene is also shown. The columns in the table include the signature sequence, class, number of hits, DNA strand, and abundance (transcripts per million [TPM] values) in each of the 12 libraries assayed. Gray TPM values indicate a lack of significant expression. A weakly expressed gene value is in the range of 1–10 TPM, while a very strongly expressed gene may be more than 1000 TPM.

When interpreting MPSS data, it is important to note the signature's class. Class 1 signatures fall anywhere within a transcript or the 3' untranslated region on the genes defined according to the available genome version. If a signature matching an intron (class 5) shows evidence of expression, the best explanation is that there is an alternative splice site, which means that is most likely part of an unannotated exon. Antisense signatures (class 3) may indicate the presence of an antisense transcript, although they could also result from mispriming of the oligo-dT during first-strand cDNA synthesis. Signatures outside a gene region may correspond to an unannotated gene. Moreover, not all signatures are unique in the genome. Around 10% occur two or more times (http://mpss.udel.edu/at/java.html?). Signatures that are not unique may result from duplications of genomic regions or of individual genes.

3.3.3. Searching Expressed Sequence Tags

Expressed sequence tags (ESTs) provide researchers with a quick route for discovering new genes, for obtaining data on gene expression and regulation, and for structural annotation. For many genes, the existence of EST matches is the only evidence for their expression. Information about the source of the cDNA library provides clues on expression localization and regulation of the gene. For example, a cDNA cloned from a leaf library indicates that the corresponding gene is expressed in leaves. In some cases, the number of EST clones in a library can be used to infer transcript levels, but EST data are not generally considered to be quantitative.

UniGene clusters (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) are built using pairwise similarity searches with nonredundant EST sequences from the dbEST division of GenBank. Each cluster contains sequences that represent a unique locus, as well as library information and map location. For a sequence to be included in a UniGene cluster, the insert sequence must have at least 100 bp of high quality sequence, not be repetitive, and have a 50-bp overlap in the 3' UTR with 100% identity. The UniGene Web site allows the user to view UniGene information on a per-cluster, per-sequence, or per-library basis ***(20)***.

TIGR has assembled *Arabidopsis* ESTs into tentative consensus (TC) sequences and provides the results as a service to the community (www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=arab). Assembly was made by clustering the EST sequences and then assembling the clusters into TCs. EST clusters and transcripts are clustered together only if they meet a certain criteria (minimum of 40 bp match, greater than 94% identity in the overlap region, and maximum unmatched overhang of 30 bp). AtGI allows searching by sequence using WU-BLAST, or using text-based searches, including GB identifiers, tissue and library names, and GO functional categories.

3.3.3.1. Using NCBI's UniGene to Find ESTs for a Sequenced Gene

1. Start at the NCBI's UniGene Search page (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene).

2. Type in the gene name or locus identifier in the text box. To restrict the search terms to specific fields, click "Preview/index." For example, to search for EST data on the Arabidopsis COP9 gene type in "COP9." Click the AND button and choose "organism" in the fields tab menu, and type in "Arabidopsis thaliana." The resulting query will read "COP9 AND organism [Arabidopsis thaliana]."
3. Submit the query.

The results will display a list of gene clusters that match your criteria. Clicking on the clusters brings a page with information on the libraries where the sequences were derived, as well as protein similarities with other organisms, mRNA and EST sequences, and links for downloading sequences. The EST library information is not curated, and as such may not be correct or may be described inconsistently. Users should keep in mind that not all the *Arabidopsis* genes are represented by a UniGene cluster.

3.3.3.2. Using ESTs to Find Genes That Are Differentially Expressed

This protocol describes how to use TIGR's AtGI database to find genes that are differentially expressed in two tissues or libraries. The program identifies TC sequences having statistically significant differential expression in two libraries.

1. Start at the AtGI Library Expression Search (www.tigr.org/tigr-scripts/tgi/libtc.pl?db=atest).
2. To compare tissues select the desired tissue by clicking on its name from the "A" tissue menu box. Press the Ctrl key while clicking to select more than one tissue, and then select the second tissue from the "B" menu box. To compare libraries, use the drop-down menus on the right, selecting a name from each box. To find information about the libraries click on the "Library Description" link.
3. Restrict the search to libraries with a certain number of ESTs (optional). Some libraries contain a small number of EST sequences, and the comparison against bigger libraries may be biased.
4. Press the "Get Expression" button.

The results display a list of TCs matching your query, along with the number of ESTs in the TC and the number of ESTs in each library or tissue compared. Significant differential expression is identified using the "R statistic." A large R indicates that is a significant bias toward one or more libraries in that TC. TCs with R-values in the top 5% are indicated with an asterisk and highlighted in red.

3.4. Arabidopsis *Metabolic Pathways*

Of the 26,207 protein coding genes included in the latest release of the genome sequence (TIGR version 5.0, January 2004) about 8170 unique loci can be classified as enzymes based on their GO annotations. To date, approx 1200 have been associated to metabolic pathways in the AraCyc Pathways Database (www.arabidopsis.org/tools/aracyc/) developed at TAIR in collaboration with SRI International ***(21)***. AraCyc contains information about *Arabidopsis* enzymes, biochemical pathways, reactions, substrates, and product compounds. The database was initially built using computational methods to generate pathways based on the correlation of predicted enzymes from the *Arabidopsis* genome with components of known pathways in other organisms ***(21)***. AraCyc pathways are then manually curated to correct, modify, and otherwise update existing pathways, and to add new pathways. When examining metabolic pathways in AraCyc it is important to bear in mind that computationally predicted pathways may be inaccurate and should be considered only as a starting point for experimentation. In addition, some pathways may also be incomplete, reflecting the current state of knowledge. There are several ways to search and browse AraCyc. The first protocol (**Subheading 3.4.1.**) illustrates how to find information about enzymes, compounds, and reactions for a specific pathway. The second protocol (**Subheading 3.4.2.**) describes how AraCyc can be used to analyze changes in the expression of genes involved in metabolism.

3.4.1. Finding Metabolic Pathways, Reactions, Enzymes, and Compounds

1. To search for information about a specific pathway, start at the main AraCyc query page at www.arabidopsis.org:1555/index.html.
2. Finding pathways: To search for pathways, choose "Pathway (by name)" and type the name of the pathway of interest into the input box. For example, type in "Lignin biosynthesis" to view this pathway. The results are displayed in a graphical format. Compounds are shown in red and reactions are indicated by blue arrows. The default view shows the broadest outline of the pathway with the least detail. A brief description of the pathway is shown in the comments section. The pathway evidence glyph summarizes the evidence supporting the reactions in the pathway. If an enzyme is present in Arabidopsis and the reaction is unique to the pathway, the reaction is well supported for that pathway. Conversely, a reaction in the pathway may not be well supported because the enzymatic function has not been identified in Arabidopsis or the reaction is present in multiple pathways.
3. Viewing the pathway in more detail: Pathways can be displayed in greater detail by clicking on the "more detail" button at the top of the diagram. At the highest level of resolution the display includes E.C. number, enzymes, locus, individual reactions, and compounds.
4. Displaying individual reactions: Clicking on the blue arrows will display a new page showing details about the specific reaction, including catalytic enzymes, a reaction diagram showing the structure of the compounds, a list of all pathways that include this reaction, and a gene-reaction schematic showing all enzymes/loci that catalyze the reaction.
5. Displaying information about specific enzymes: Clicking on the name of the enzyme will display a new page with a summary of the reactions catalyzed by the enzyme and all pathways in which the enzyme is known (or is predicted) to participate. To view the TAIR locus detail page for the enzyme, click on the symbolic gene name or locus identifier.
6. Displaying individual compounds: Clicking on the compound will display the compound formula, SMILES string, molecular weight, and a list of reactions for which the compound is either a reactant or a product.

3.4.2. Detecting Changes in Expression of Genes Involved in Metabolism

The Pathway Omics Viewer allows displaying expression data on the pathway diagrams, and thus can be used as a tool to examine how enzymes may be regulated in response to experimental treatments in a pathway context. Expression values are displayed on the pathway overview map as color-coded reaction lines indicating the expression level of the enzyme that catalyzes the reaction. For time-course experiments, data can be displayed as an animation, and the changes in gene expression are easily distinguished by following the color changes in a particular pathway for each time point. The animation can be stopped at any point to facilitate scanning and querying of the pathways.

1. Start at the AraCyc Omics Viewer at www.arabidopsis.org:1555/expression.html.
2. Click on "Browse" to find the expression data file in your local computer. The file should be a tab-delimited text file. Word or other word processing software files cannot be used. The first column must be the AGI locus identifier. The remaining columns should contain the expression values (absolute or relative) from each treatment. To display data from a time series, the expression values corresponding to consecutive time point should be included in the same order in consecutive columns (e.g., values for time point 1 in column 1, values for time point 2 in column 2).
3. Select the types of expression values present in your file (absolute or relative) and the number of data columns to be analyzed. If the data contain absolute values in a single column, choose "absolute" and "single" column display. If the values in the columns are already relative values, select only a single column to display and select relative expression.
4. Specify whether the data file has log values. Choose this option to display log values. If not, negative values will be discarded.
5. To display a single data column or ratio of two data columns, select the number of columns to display. By default, the first column of locus identifiers is 0 and the first data column is 1. Columns that are not selected will not be displayed on the Overview diagram. For example, to

display only the first data column, type in 1 or to show three time points in a file containing absolute values, type in 1, 2, and 3 in the single column data input box.

6. Specify a value for the maximum cutoff. By default the cutoff values are determined by the data values. To compare expression from different experiments, select the same maximum cutoff value for each display.
7. Click "Submit." The results will be displayed in the Pathway Overview Omics Viewer, which shows the metabolic pathway diagram colorized with expression data according to the relative or absolute expression level of the gene that codes for the enzyme that catalyzes that reaction step. The display includes a histogram showing the distribution of values in the dataset and a key to the diagram. Statistics, including minimum and maximum values and missing information, are shown on the bottom of the display, including the loci in the expression file that were not found in AraCyc.

3.5. Finding Related Protein Sequences in Arabidopsis

For sequenced genes for which limited experimental data are available, one of the first steps toward understanding a gene's function is to search for evolutionarily related proteins and conserved motifs. The function of an unknown gene may be inferred from its similarity to a well-characterized homolog or from the presence of conserved domains. Other motifs, such as transmembrane domains or signal peptides, can be used to infer protein localization. Structural classification based on protein folding is another way of finding more distantly related proteins. The assumption is that protein topology is functionally important and proteins with similar structures may have similar molecular activities.

A comprehensive analysis of protein families and domains typically requires queries of multiple databases and algorithms and can be very time-consuming. To facilitate gene family and domain analysis, TIGR, TAIR, and MIPS databases routinely analyze the entire *Arabidopsis* proteome as part of their annotation "pipelines" ***(2–4)***. Such a catalog of structural features facilitates the classification of domains that are over- or underrepresented in a genome. It also makes possible the grouping of members of gene families according to domain composition and order.

3.5.1. Finding Similar Protein Sequences

Searching for similar protein sequences in *Arabidopsis* using local sequence alignment methods can be performed at TAIR, TIGR, MIPS, and NCBI (**Table 2**). These groups have overlapping *Arabidopsis* datasets; TAIR has some other *Arabidopsis*-specific datasets not found at TIGR, MIPS, or NCBI (www.arabidopsis.org/help/helppages/BLAST_help.jsp#datasets). These datasets are used by all of TAIR's sequence similarity programs (FASTA, WU-BLAST, and NCBI BLAST and PatMatch).

This protocol illustrates how use TAIR's WU-BLAST tool to identify similar genes in *Arabidopsis*.

1. Start at the WU-BLAST search page. Point your browser to the URL www.arabidopsis.org/wublast/index2.jsp.
2. Select the appropriate BLAST program. Five different algorithms are available to match amino acid or nucleotide sequences. The choice of the program depends on the type of sequence to be queried and the query database. For example, when comparing a protein sequence to other protein sequences choose the BLASTP program (www.arabidopsis.org/help/helppages/BLAST_help.jsp#methods).
3. Input your query sequence. The tool accepts sequences or locus identifiers as input. To use a sequence as input, paste in the sequence as raw text or in FASTA format, or upload it from a file. Sequences pasted directly from GenBank records can also be used. To use a locus identifier as input, choose the locus name option under the input header, and type in the name of the locus, or upload it from a file. When using locus identifiers as input the program retrieves the coding sequence (CDS) for the locus; therefore it cannot be used with the BLASTP or TBLASTN options. To perform a search using more than one query sequence, submit multiple

sequences as a list of locus identifiers or as a set of FASTA formatted sequences, each sequence having its own FASTA header.
4. Define the dataset to search against. For example, to find homologous proteins in Arabidopsis choose the AGI protein dataset. This dataset is a nonredundant set of all known Arabidopsis proteins.
5. Customize the BLAST search parameters. The default parameters are filtering on an expect threshold (cutoff) of 10. The default S value is calculated based on the E value and represents the single high-scoring pair (HSP) score that satisfies the expect threshold.
6. Submit the query. Click on the "RUN BLAST" button. If you have chosen an inappropriate combination of query sequence and database, an error will be returned to your browser.

Results from the WU-BLAST search are presented in a graphical format that can be used to rapidly assess the significance of the results. The graph displays the query sequence in red and the HSP matches below. The length of the bar corresponds to the length of the HSP and the color of the bar indicates the range of expected values (the probability of finding the sequence match by random chance). The direction of the bar indicates whether the match is on the forward or reverse strand. Pointing the mouse over the HSP markers will display the description line of the matched sequence. Clicking on the HSP will display the selected sequence alignment. For AGI genes and loci, the name in the alignment is hyperlinked to the TAIR locus detail page (*see* **Subheading 3.2.1.**). Other similarity groupings can be found using the External links from locus detail pages. Both TIGR and MIPS databases include precomputed matches to similar proteins in their locus records. NCBI Blink shows similar proteins in GenBank. The Blink results show BLASTP searches against all proteins in GenBank in a graphical format along with the similarity score. To find the best matches for each taxonomic group, click on the button labeled "Best Hits." To access the GenBank record for each sequence, click on the sequence accession number.

3.5.2. Finding Similar Structures and Domains in Proteins

Comprehensive protein structural data can be found at TIGR, MIPS, NCBI, and TAIR (**Table 2**). The protocol described here demonstrates how to use TAIR's Bulk Protein Download tool to obtain a list of structural, physical, and chemical properties for a set of proteins.

1. Start at the Bulk Protein Download. Point the browser to the URL www.arabidopsis.org/tools/bulk/protein/index.jsp.
2. Choose the output display. The output options include molecular weight, isoelectric point, transmembrane domains, SCOP structural class, domains, and hyperlinked SwissProt IDs. Selecting the HTML option will display links to TAIR locus detail pages, protein sequence, SeqViewer graphical display, the protein record in SwissProt, and INTERPRO. The last two links are shown only if domains and Swiss Prot IDs are included in the output. Choose "text" output if you wish to download the data into your computer. Queries that return more than 1000 results will be returned as text-only format.
3. Limit the search by protein properties. For example, to obtain a list of proteins with a given range of molecular weights.
4. Submit the query by clicking on the "Get Protein Data" button.

Protein structural annotations may not be constant from database to database because different analysis methods or sequences are used. Domain databases are also updated frequently as new domain structures are identified. Frequent checks of genome databases should be done to determine whether new domains have been identified.

3.6. From Genotype to Phenotype—Finding Mutants for Analysis of Gene Function

By analyzing the effects of mutations it is possible to infer the function of a gene and, through epistasis analysis, define networks and interactions among genes. Different classes of mutations provide diverse tools for genetic analysis of gene function. Typically, when trying to ascertain a gene's function, knockouts or null mutations are preferred to determine the effect of

a complete loss of gene function on the phenotype of an organism. In some cases, a knockout does not give a detectable phenotype. There are many reasons why this may occur. If the gene is a member of a closely related gene family, the effect of the mutation may be masked by the activity of homologous genes. In this case double, triple, or even quadruple mutants may be needed before a phenotype can be detected. Loss-of-function mutations may also result in embryonic lethality or sterility, thus complicating the analysis. Another possibility is that the phenotype may not be observable under the growth conditions used. Correlation with functional annotations and gene expression data may be useful in determining when, where, and under what conditions a phenotype can be observed. Dominant gain-of-function or knock-on mutations can be useful in revealing the functions of genes that do not have a phenotype when knocked out, by causing over or ectopic gene expression ***(22)***. Analyzing the phenotypes of point mutations can provide important information about specific motifs ***(23,24)*** and can be a preferred source material for dissecting genetic pathways through enhancer/suppressor screens. This section describes how to find knockouts and other types of mutations in a gene or genes of interest.

3.6.1. Finding Knockout Mutations in a Gene

There are several large scale functional genomics projects that utilize modified T-DNA or transposons aimed to generated knockout mutations in every *Arabidopsis* gene ***(25)***. TAIR, AtIDB, SIGNAL T-DNA Express, and NASC have identical or at least largely overlapping datasets of the major T-DNA and transposon stocks (**Table 2**). See the appropriate database release notes to determine which sets of insertions are stored. The following method describes how to use BLASTN at the SALK T-DNA Express Gene Finding Tool to locate insertions in a gene of interest. The gene sequence is used to query a database of genomic sequences flanking the site of an insertion. T-DNA and transposon insertion sites are estimated, based on matching the flanking sequences to the genome. This calculated position is not always correct due to sequence errors and incomplete matches to the genome (http://signal.salk.edu/tdna_FAQs.html).

1. Start at the T-DNA Express search page (http://signal.salk.edu/cgi-bin/tdnaexpress).
2. Input the query sequence and select the search parameters. Paste your sequence into the input box and select the number of results to display and a cutoff similarity score (E-value). Choose a smaller E-value to display the best matches and a larger one to find lower-quality matches. For best results use the cDNA or coding sequence and BLASTN and the default cutoff value. The best matches may still be inexact, as genomic sequences flanking insertions are generally of moderate quality, and may be derived from a different genetic background than the Col-0 ecotype used for the genomic sequence. Submit the query.
3. Find the insertions. At the top of T-DNA Express BLAST report will be a summary box showing the locations of the best match to your sequence. The alignments are shown below the summary. To display a graphical view in T-DNA express click "To TDNA Express" hyperlink in the summary section. This will open a genome browser centered on the gene, which matched your query sequence. Genes are located on the top of the chromosome bar and insertion flank sequences are annotated below. The starting point of the arrow indicates the position of insertion and the arrowhead indicates the direction of the sequence match. Zoom controls can be used to obtain close-up views in order to better see where the insertion flanks are located within the gene. In the gene annotation band, exons are identified as green bands and introns are clear/white bands. To view detailed information about any insertion, click on the name of the insertion as it appears in the lower band. From the detail page you can see the exact coordinates and determine whether the insertion is in an exon, intron, intergenic region, or splice junction. To order stocks, use the appropriate links to NASC, ABRC, or FLAGDB.

TAIR, AtIDB, and NASC all have similar functionalities, but there are a few important differences to note. NASC BLAST differs in returning only alignments to the insertion flanking sequences, which are not linked to a graphical genome browser. NASC also features the Insert Watch utility, which allows researchers to submit a sequence of interest and receive notification when a new insertion flank sequence has been found with a significant match to the query sequence of interest. Researchers are automatically notified via email of the existence of

a new insertion that is potentially located in their gene of interest. If using TAIR's BLAST, clicking on the T-DNA flank sequence name from the BLAST results will open a page displaying the stock details for the T-DNA insertion line. To see a graphical display of the insertion site in the genome, click on the locus link, scroll down to the Map Links, and click on the Sequence Viewer. Registered users can order stocks from ABRC or NASC directly from this page. TAIR BLAST also allows for batch BLAST queries to identify insertions in more than one gene (e.g., multiple members of a gene family).

3.6.2. Finding Other Types of Mutations and Mutants With Similar Phenotypes

Mutants with similar phenotypes might be new alleles of known genes or function in the same genetic pathways. Quantitative trait analysis of natural variants can be used to find genes that may not be identified using standard mutational analysis. Comparing the phenotypic descriptions of related mutants can reveal other associations that may not be obvious from the literature. For example, a mutation identified in a screen for hormone resistance may have also been found in a screen for floral mutants. Finding correlations from phenotypes of different alleles may suggest new avenues of experimentation.

TAIR's Germplasm Search can be used to find natural variants and different types of mutants including T-DNA insertions, TILLED substitutions, and other types of induced mutations. The following protocol describes a method for searching for any germplasm having a mapped polymorphism at a locus of interest, or having a specific phenotype.

1. Start at TAIR's Germplasm Search (www.arabidopsis.org/servlets/Search?action=new_search&type=germplasm).
2. To search for mutants in a specific gene, define the gene of interest. In the Search by Name, Description, or Stock Number section choose "Locus Name" from the drop-down menu. Type in the AGI locus identifier for your gene of interest.
3. To find mutants with similar phenotypes, choose "description" from the drop-down menu found under the Search by Name section at the top of the page. For the broadest search choose the "contains" option and type in the terms to search for in the descriptions. Entering different terms in each input box will return those germplasms that contain all the specified terms within their descriptions (e.g., LEAF and SERRATED). Additional options can be selected to further refine your search, such as selecting the background of the mutation, which can be useful when searching for mutations in the same genetic background.
4. Submit the query. The results will list the germplasm name(s), polymorphisms, locus, background strain, a description of the germplasm (e.g., phenotype), donor name, and stock number. If images of the germplasm are available, a camera icon will be displayed.
5. Find allele(s) of interest. Click on the name of the polymorphism to view associated phenotypes, physical location of the lesion, mutagen that caused the mutation, and other germplasms that contain the allele. For some genes, it may be possible to acquire a range of mutations from knockouts (e.g., T-DNA insertion) to substitutions, or overexpressers (e.g., activation tag lines). A graphical representation of the gene structure and all the associated polymorphisms can be obtained by clicking on the Sequence Viewer link in the Map Links section on any polymorphism detail page. This will open a view of the SeqViewer genome browser with the polymorphism highlighted. The Zoom to drop-down menu can be used to display a closeup view of the gene, and the global controls (top left box) can be set to display specific annotations such as genes, transcripts, polymorphisms, and T-DNA/Transposons (www.arabidopsis.org/seqViewer/help/sv_intro.jsp). A more precise view of the sites of the polymorphisms can be obtained by mousing over the polymorphism of interest and selecting the nucleotide sequence view from the pop-up window. The nucleotide sequence view displays 10 kb of sequence. In the upper right corner of the nucleotide view window choose the appropriate objects to display. For example, choose Genes, Markers, and Polymorphisms to show the exact location of the polymorphisms within the gene. To get to a 10-kb centered view of a locus from a close up view anywhere on the chromosome, first zoom to a 10-kb view, then type in the name of the locus in the upper left text box and click on "Find."

6. Finding out about the germplasm: In the results page, the name of the germplasm is linked to a detailed record containing a full description of the line; information about the origins of the strain, such as pedigree and genetic background, images, and known polymorphisms; and, for germplasms available from ABRC or NASC, a description of the seed stock.
7. To order the seed stock from the ABRC, check the box in the "order" column. For each page of results, check the stocks you wish to order. When you are done selecting stocks, go to the top of the results page and click on the button to order the checked stock. If you are not logged in you will be prompted to do so. If you are not registered, you will need to register and be affiliated with a lab (for billing purposes) before you can order the stocks.

3.7. From Phenotype to Genotype—Databases and Tools For Map-Based Cloning

A primary method for gene discovery is the analysis of mutants and natural variations and subsequent cloning of the corresponding locus controlling the trait(s) of interest. Map-based cloning is a method for isolating genes for quantitative trait loci (QTLs) and mutations that are not tagged with foreign DNA. Fine mapping is done to reduce the interval to a small region (about 0.16 cM or approx 40 kb) containing only a few genes and then identifying candidate genes within the mapped region. Typically this involves generating a mapping population between two ecotypes and analyzing a large number of recombinants (approx 3000 chromosomes). Confirming the identity of the gene requires rescuing the mutant phenotype with a transgenic construct containing the wild-type locus. The combination of a sequenced genome, a rich resource of genetic markers, and publicly available polymorphism data has substantially reduced the amount of time required for positional cloning ***(26,27)***. For QTLs, the process is more complex and usually requires creating a population of isogenic lines (NILs) in order to fine-map the QTL ***(28,29)***. The protocols in this section describe how to use information at TAIR to assist in genetically mapping a mutation, finding candidate genes, and obtaining biological materials for complementation.

3.7.1. Finding and Downloading Sets of Genetic Markers for Mapping

1. Start TAIR's Genetic Marker Search page at www.arabidopsis.org/servlets/Search?action=new_search&type=marker).
2. Choose the type(s) of markers you want to use. PCR-based markers such as CAPS and SSLPs are by far the most commonly used marker types. In the section marked "Restrict by Features" check the boxes marked SSLP and CAPS.
3. Choose the mapping (parental) ecotypes. You can select markers that are known to reveal polymorphisms between two ecotypes. In the "Restrict by Features" section use the drop-down menus to select the two parental ecotypes. For the broadest search, or if your parental lines are not shown, keep the default selection (any).
4. Select the region of the chromosome of interest. If you have already determined linkage to a genetic marker or gene you can use the Range option to further restrict your search. First choose the map you are using. As most PCR-based markers are located on the AGI sequence map, click on the button next to this option in the "Map" section. Then select the region of the chromosome. To find markers 100 kb distal and 100 kb proximal to an existing marker, choose the "Around" option and type in the name of the linked marker. Note that the second option is disabled.
5. Choose the output option. You can choose to have the results displayed by name, position, or type. For example, if you want to see SSLP and CAPS markers separated, choose the "type" option.
6. Submit the query.
7. Viewing the results: The results of the query are returned to your browser window. You can view each genetic marker record by clicking on the marker name link. Alternatively, you can download the entire result set as a tab-delimited file and save it on your computer.
8. Downloading a list of markers: To save one or more of the markers to a list, use the check boxes to select the markers you want to save. If more than one page of results is returned, go through each page and select the markers you want. Once you have selected the markers, scroll to the top of the page and click on the "Download" button. The text file will have many fields, including primer sequences (if known), polymorphisms, and corresponding ecotypes.

3.7.2. Finding Polymorphisms Between Two Ecotypes for Generating New Markers

To generate a high-resolution map, it is often necessary to narrow down the interval by generating new markers in the region of interest. Single-nucleotide polymorphisms (SNPs) and small insertion/deletion (In Del) polymorphisms can be used to generate additional markers. SNPs can be used to make CAPS markers if the substitution causes a restriction site polymorphism. If no site is present, degenerate (dCAPs) markers can be made by introducing a mismatch in one primer that creates a restriction site polymorphism ***(30)***. In Dels can be used as a basis for generating SSLP markers. SNPs and In Dels can be used with high-throughput SNP detection methods that allow for faster mapping using smaller populations. Several large SNP identification projects have made their data available to the public. Monsanto has made its database of more than 57,000 polymorphisms between the Landsberg *erecta* and Columbia strains available to the research community (www.arabidopsis.org/Cereon/index.jsp; ***26***). SNP data from the Stanford Genome Sequencing Center (SGC; ***31***) and the Max Planck Arabidopsis SNP Consortium (MASC) database ***(32,33)*** are searchable at TAIR and are displayed on the genome browser. The SGC SNPs represent polymorphisms between L*er* and Columbia; MASC SNPs are tested in 12 different ecotypes.

TAIR's Polymorphism Search can be used to find many types of polymorphisms using a variety of criteria such as gene or locus name, polymorphism type, mutation site, and chromosomal location. This method describes how to find SNP or In Del polymorphisms between two specific ecotypes and shows how this information can be used to design PCR primers for genetic markers.

1. Start at TAIR's Polymorphism/Allele Search (www.arabidopsis.org/servlets/Search?action=new_search&type=polyallele).
2. Select the type(s) of polymorphisms to retrieve. For designing PCR markers, choose substitution, SNP, or In Del. To select more than one type, hold down the CTRL key (PC) or Apple key (Mac) and select each type with a mouse click.
3. Select the parental ecotypes. Use the drop-down menus to choose the ecotypes in which the polymorphism should be present. The selection is based on the current information in the database. If only one ecotype is listed, choose this as the first and "any" as the second option. If neither ecotype is represented, leave the default (any) option selected for both.
4. Define the chromosome of interest. Assuming a rough map position has been established, you can choose to limit the search to polymorphisms on a specific chromosome and further refine this by selecting the physical boundaries. Because sequenced polymorphisms are mapped on the sequence (AGI) map, choose this as the map type. Having selected this option, you can restrict to a range of kilobases or use any gene, marker, or clone in TAIR to define the boundaries.
5. Define the output format. Use the drop-down menus at the top of the search page to select the number of records to display and the sorting order of the display. Sorting by position will list all polymorphisms first by chromosome (if none was selected) and then by position from the top to bottom of the chromosome. Other options include sorting by name or type. The results include some or all of the following: name of the polymorphism linked to a detailed record, alternate names (aliases), polymorphism type, chromosome, starting position of the polymorphism, links to the SeqViewer close-up view centered on the polymorphism, links to the locus detail page if the polymorphism is within a locus, and a description.
6. Finding information about the polymorphic sequences: Click on the name of the polymorphism to display the detailed information, including the sequences that differ between ecotypes, along with their length and all ecotypes that contain this variant, and approx 20 bp of sequences flanking the polymorphism. For many polymorphisms this information can be used as a starting point to design markers.
7. To generate CAPS or dCAPs markers, go to the dCAPs finder program (http://helix.wustl.edu/dcaps/dcaps.html) by clicking on the link to the dCAPs finder from the "External Links" section. This program can be used to find existing restriction site polymorphisms or to design primers that will introduce a restriction site in one of the two alleles. You will need to use the flanking sequence data and the polymorphic sequence data from the polymorphism detail page to create the dCAPs primer.

3.7.3. Finding Candidate Genes in a Genetically Defined Interval

After the locus has been fine-mapped, the next step is to determine which genes located in the interval may be responsible for the phenotype. TAIR's genome browser, called SeqViewer, can be used to view the *Arabidopsis* genome from a whole-genome view down to the nucleotide level. SeqViewer displays genes, annotation units (BAC clones used to generate and assemble the genome sequence), transcripts (including ESTs and full-length cDNAs), polymorphisms, T-DNA/transposon insertions, and markers. The tool allows searching with up to 250 names or four short (<150 nucleotide) nucleotide sequences. The locations of one or many search hits on the whole genome can be shown in a close-up view (zoomable from 50 megabases to 10 kilobases) or in a 10-kb-nucleotide window. This protocol describes how to use SeqViewer to find candidate genes within a region containing a genetically defined interval using flanking marker data.

1. Start at the TAIR SeqViewer (www.arabidopsis.org/servlets/sv).
2. To view a region containing two or more markers, in the search box below the whole-chromosome view, type in the names of the flanking markers or polymorphisms and click "Submit." The matched hits will be displayed as red tick marks on the whole-chromosome view. Point your mouse to the region between the red tick marks and click. This will open a new view of the chromosome in that region. Adjust the zoom levels until both highlighted markers are visible.
3. Create a view centered on the region of interest between the two flanking markers. Type in the leftmost coordinate of the left marker and rightmost coordinate of the right flanking marker in the "Select Range" input box. Click "Go" to display a custom view of the chromosome. If the markers you have used for mapping are not in TAIR's database, you can use the primer sequences to locate the region of interest. From the SeqViewer home page, choose "search by sequence" and paste in the primer sequences in the input box and submit the query. Exact matches will be displayed as red tick marks on the whole-genome view. Click on the red marks to display a closeup view. Use the controller to align the view to display the region of interest.
4. Obtain a list of genes located between the marker endpoints and identify appropriate candidates. To obtain a list of genes use the Close Up view controller; click on the button to "List Genes in Range." This will open a new browser window containing a list of all loci in the close-up view. Click on the locus name to view the detailed record for this locus in TAIR or follow the links to view the corresponding locus details in the TIGR or MIPS *Arabidopsis* databases. To save the list, go to the upper right corner of the list of genes and click on "Download as text file." This list of locus identifiers can be used as input for a variety of tools. For example, it can be used to obtain a list of GO annotations (**Subheading 3.2.**) that may suggest a function or role consistent with the mutant phenotype or to obtain expression data for these genes (**Subheading 3.3.**). The expression data may show correlations with the phenotype in question (expressed in the same tissues that exhibit an abnormal phenotype). Go to the TAIR locus page to find cDNA clones to use as probes for Northern and *in situ* hybridization, genomic clones for generating complementation constructs, and additional alleles to confirm the phenotypes.

Acknowledgments

We would like to thank Eva Huala, Peifen Zhang, Tanya Berardini, Suparna Mundodi, and Sue Rhee for reviewing the manuscript. This project was supported in part by the National Science Foundation (grant numbers DBI-9978564, DBI-0091471, and DBI-9813586) and by the National Institutes of Health (grant number HG-02273). This is the Carnegie Institution of Washington Department of Plant Biology Publication 1679.

References

1. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.
2. Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H. W., and Mayer, K. F. (2004) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.* **32,** Database issue, D373–D376.

3. Wortman, J. R., Haas, B. J., Hannick, L. I., et al. (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol.* **132,** 461–468.
4. Garcia-Hernandez, M., Berardini, T. Z., Chen, G., et al. (2002) TAIR: a resource for integrated *Arabidopsis* data. *Funct. Integr. Genomics* **2,** 239–253.
5. Huala, E., Dickerman, A. W., Garcia-Hernandez, M., et al. (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.
6. Rhee, S. Y., Beavis, W., Berardini, T. Z., et al. (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.
7. Berardini, T. Z., Mundodi, S., Reiser, L., et al. (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* **135,** 745–755.
8. Alonso, J. M., Stepanova, A. N., Leisse, T. J., et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301,** 653–657.
9. Anderle, P., Duval, M., Draghici, S., et al. (2003) Gene expression databases and data mining. *Biotechniques* Suppl, 36–44.
10. Haas, B. J., Volfovsky, N., Town, C. D., et al. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3,** RESEARCH0029.
11. Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., and Mewes, H. W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics* **17,** 44–57.
12. Mewes, H. W., Amid, C., Arnold, R., et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32** Database issue, D41–D44.
13. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29.
14. Wisman, E. and Ohlrogge, J. (2000) *Arabidopsis* microarray service facilities. *Plant Physiol.* **124,** 1468–1471.
15. 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** Database issue, D575–D577.
16. Wang, X., Hessner, M. J., Wu, Y., Pati, N., and Ghosh, S. (2003) Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics* **19,** 1341–1347.
17. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* **2,** 418–427.
18. Kerr, K. M., Churchill, G. A. (2001) Statistical design and the analysis of gene expression. *Genet. Res.* **77,** 123–128.
19. Brenner, S., Johnson, M., Bridgham, J., et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18,** 630–634.
20. Wheeler, D. L., Church, D. M., Edgar, R., et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucl. Acids. Res.* **32,** D35–D40.
21. Mueller, L. A., Zhang, P., and Rhee, S. Y. (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.* **132,** 453–460.
22. Ichikawa, T., Nakazawa, M., Kawashima, M., et al. (2003) Sequence database of 1172 T-DNA insertion sites in *Arabidopsis* activation-tagging lines that showed phenotypes in T1 generation. *Plant J.* **36,** 421–429.
23. Colbert, T., Till, B. J., Tompa, R., et al. (2001) High-throughput screening for induced point mutations. *Plant Physiol.* **126,** 480–484.
24. McCallum, C. M., Comai, L., Greene, E. A., and Henikoff, S. (2000) Targeted screening for induced mutations. *Nat. Biotechnol.* **18,** 455–457.
25. May, S. T., Clements, D., and Bennett, M. J. (2002) Finding your knockout: reverse genetics techniques for plants. *Mol. Biotechnol.* **20,** 209–221.
26. Jander, G., Norris, S. R., Rounsley, S. D., Bush, D. F., Levin, I. M., and Last, R. L. (2002) *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiol.* **129,** 440–450.
27. Lukowitz, W., Gillmor, C. S., and Scheible WR. (2000) Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. *Plant Physiol.* **123,** 795–805.
28. Yano, M. (2001) Genetic and molecular dissection of naturally occurring variation. *Curr. Opin. Plant Biol.* **4,** 130–135.
29. Maloof, J. N. (2003) Genomic approaches to analyzing natural variation in *Arabidopsis thaliana*. *Curr. Opin. Genet. Dev.* **13,** 576–582.

30. Neff, M. M., Turk, E., and Kalishman, M. (2002) Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.* **18,** 613–615.
31. Cho, R. J., Mindrinos, M., Richards, D. R., et al. (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* **23,** 203–207.
32. Schmid, K. J., Sorensen, T. R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T., and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13,** 1250–1257.
33. Torjek, O., Berger, D., Meyer, R. C., et al. (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* **36,** 122–140.