Chapter 4

Arabidopsis Database and Stock Resources

Donghui Li, Kate Dreher, Emma Knee, Jelena Brkljacic, Erich Grotewold, Tanya Z. Berardini, Philippe Lamesch, Margarita Garcia-Hernandez, Leonore Reiser, and Eva Huala

Abstract

The volume of *Arabidopsis* information has increased enormously in recent years as a result of the sequencing of the reference genome and other large-scale functional genomics projects. Much of the data is 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 and stock centers.

Key words Data mining, Database, Genomics, Gene expression, Bioinformatics, Computational biology, Stocks, *Arabidopsis thaliana*

1 Introduction

Arabidopsis thaliana serves as the primary model system for many aspects of plant biology. It was the first plant to have its entire nuclear genome sequenced [1]. Following the completion of the *Arabidopsis* genome sequencing in 2000, the international *Arabidopsis* community set an ambitious goal to determine the function of every *Arabidopsis* gene by the year 2010 [2]. Numerous laboratories internationally have taken part in this project (Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project). Large amounts of data about gene function, expression, metabolism, and protein and gene interactions have been generated by these labs. To accomplish the task of organizing and managing the data, lab consortia and individual labs have created databases to store the information generated and make it available to the research community. Community resources such as genome-wide DNA clones and knockout mutant libraries (e.g., SALK T-DNA insertion lines) were also created [3]. There are now extensive tools and resources for storage, curation, and

Jose J. Sanchez-Serrano and Julio Salinas (eds.), *Arabidopsis Protocols*, Methods in Molecular Biology, vol. 1062,
DOI 10.1007/978-1-62703-580-4_4,

retrieval of *Arabidopsis* data and DNA and seed stocks. Scientists doing research in this "postgenomic" era are compelled to know how to make use of these resources to find the relevant information and stocks needed to further their research.

In this chapter, we describe how to use databases to find what is known about *Arabidopsis* and to make inferences and predictions that can later be tested experimentally. We include a summary of the rationale, a brief description of the database/tool(s), and the specific steps for querying, retrieving, and interpreting the data. Methods on how to search and order DNA or seed stocks are also provided. The methods, along with the corresponding databases and tools, are outlined in Table 1. This table of contents can be used to find specific methods of interest within the chapter.

Databases described here represent a small portion of the vast collection of databases and bioinformatics resources available for *Arabidopsis* researchers. In this chapter, we focus on well-developed resources that provide comprehensive *Arabidopsis* data (including stocks) such as TAIR (The Arabidopsis Information Resource) [4–8] and ABRC (Arabidopsis Biological Resource Center) [9]. There are many more databases that focus on specific types of *Arabidopsis* information such as subcellular localization (SUBA: SUB cellular location database for *Arabidopsis* proteins, http://suba.plantenergy.uwa.edu.au/) [10], whereas others focus on specific classes of genes or disseminate data from a functional genomics project, e.g., the Chloroplast 2010 database (http://www.plastid.msu.edu/) [11]. Many links to these external resources and US National Science Foundation 2010 *Arabidopsis* functional genomics project pages (http://www.arabidopsis.org/portals/masc/projects.jsp) are provided on the TAIR Portal pages (http://www.arabidopsis.org/portals/). There is also currently an ongoing effort aiming to integrate all *Arabidopsis* database resources by the proposed International Arabidopsis Informatics Consortium [12]. In addition to databases that are entirely devoted to *Arabidopsis* ("*Arabidopsis* specific"), there are also numerous multi-species databases containing *Arabidopsis* data along with information about other organisms, such as the National Center for Biotechnology Information's (NCBI) GenBank (http://www.ncbi.nlm.nih.gov/genbank/), the European Bioinformatics Institute's (EBI) InterPro (http://www.ebi.ac.uk/interpro/), UniProt (http://www.uniprot.org/), and PlantGDB (http://www.plantgdb.org/), to name a few. Some of these databases are listed in Table 1. This chapter does not intend to cover all these databases in depth; instead we hope it will serve as a good starting point for anyone who wishes to explore these valuable resources.

Arabidopsis seed and DNA stocks and other biological materials can be obtained from a number of different institutions around the world. These stock centers provide different kinds of materials and different levels of service. The Arabidopsis Biological Resource

Table 1
Selected *Arabidopsis* databases and stock resources

Database: tool	URL	Protocol/description
Finding comprehensive information about *Arabidopsis* genes		
TAIR: Gene Search	http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene	3.1.1. Finding genes in TAIR by name
NCBI: Gene Search	http://www.ncbi.nlm.nih.gov/gene/	Finding genes in NCBI's Reference Genome Collection. Search by locus identifiers, symbol, etc.
Ensembl Plants Genome Browser	http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index	Multi-species plant genome database providing access to *Arabidopsis* and other plant genomic data
SUBA (SUBcellular location database for Arabidopsis proteins)	http://suba.plantenergy.uwa.edu.au/	Finding protein subcellular location information
TAIR: NBrowse	http://www.arabidopsis.org/tools/nbrowse.jsp	Finding protein–protein interaction information
BioGRID (Biological General Repository for Interaction Datasets)	http://thebiogrid.org/	Finding protein–protein interaction information
MIND (Membrane protein interaction database)	http://www.associomics.org/Associomics/Home.html	Finding *Arabidopsis* membrane interactome data
Finding gene sequence and structure data		
TAIR: Sequence Bulk Download and Analysis	http://www.arabidopsis.org/tools/bulk/sequences/index.jsp	3.2.1. Retrieving DNA and protein sequence data
TAIR: GBrowse	http://gbrowse.tacc.utexas.edu/cgi-bin/gb2/gbrowse/arabidopsis/	3.2.2. GBrowse
TAIR: WU-BLAST	http://www.arabidopsis.org/wublast/index2.jsp	3.2.3. Finding related DNA or protein sequences
TAIR: Bulk Protein Download	http://www.arabidopsis.org/tools/bulk/protein/index.jsp	3.2.4. Finding protein structure and domain information
Phytozome	http://www.phytozome.net/	Comparative genomic database providing access to 25 green plant genomes which have been clustered into gene families
InterPro	http://www.ebi.ac.uk/interpro/	Finding predicted protein signature (domain) information
1001 Genomes	http://1001genomes.org/	*Arabidopsis thaliana* genetic variation database

(continued)

Table 1
(continued)

Database: tool	URL	Protocol/description
Finding Gene Ontology (GO) annotations		
TAIR: Gene Ontology Annotations Search	http://www.arabidopsis.org/tools/bulk/go/index.jsp	3.3.1. Finding GO annotations
TAIR: Keyword Search	http://www.arabidopsis.org/servlets/Search?action=new_search&type=keyword	3.3.2. Finding genes annotated to related functions or processes
AmiGO	http://amigo.geneontology.org/cgi-bin/amigo/go.cgi	Searching the Gene Ontology database
Finding information about gene expression		
TAIR: Plant Ontology Search	http://www.arabidopsis.org/tools/bulk/po/index.jsp	3.4.1. Finding Plant Ontology annotations
TAIR: Microarray Expression Search	http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression	3.4.2. Finding DNA microarray data
Plant Ontology Consortium Database	http://www.plantontology.org/	Searching or browsing PO and PO annotations
NASCArrays	http://affy.arabidopsis.info/narrays/experimentbrowse.pl	Finding microarray data from the European Arabidopsis Stock Center's microarray database
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/	European Bioinformatics Institute's microarray database
NCBI GEO (Gene Expression Omnibus)	http://www.ncbi.nlm.nih.gov/geo/	NCBI's gene expression data repository
Genevestigator	https://www.genevestigator.com	Analysis tool for mining microarray datasets
eFP Browser	http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi	Analysis tool for mining microarray datasets
Obtaining information about metabolism in *Arabidopsis*		
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg/	3.5. Obtaining information about metabolism in *Arabidopsis*
Kazusa Plant Pathway Viewer (KaPPA-View4)	http://kpv.kazusa.or.jp	3.5. Obtaining information about metabolism in *Arabidopsis*
KNApSAck	http://kanaya.naist.jp/KNApSAcK/	3.5. Obtaining information about metabolism in *Arabidopsis*
MetNet	http://metnetonline.org/	3.5. Obtaining information about metabolism in *Arabidopsis*
Arabidopsis Reactome	http://www.arabidopsisreactome.org/	3.5. Obtaining information about metabolism in *Arabidopsis*
MetaCrop	http://metacrop.ipk-gatersleben.de.	3.5. Obtaining information about metabolism in *Arabidopsis*
AraCyc/PlantCyc/PMN	http://www.plantcyc.org	3.5. Obtaining information about metabolism in *Arabidopsis*

Finding and ordering seed and other resources		
ABRC: Stock Catalog	http://www.arabidopsis.org/servlets/Order?state=catalog	3.6. Finding and ordering seed resources from ABRC
TAIR/ABRC: Seed Germplasm Search	http://www.arabidopsis.org/servlets/Search?action=new_search&type=germplasm	3.6. Finding and ordering seed resources from ABRC
TAIR/ABRC: DNA/Clones Search	http://www.arabidopsis.org/servlets/Search?action=new_search&type=dna	3.7. Finding and ordering other (non-seed) resources from ABRC
NASC (European Arabidopsis Stock Centre)	http://arabidopsis.info/	Finding and ordering seed and clone stocks from the European Arabidopsis Stock center
RIKEN Biological Resource Center Experimental Plant Division (Japan)	http://www.brc.riken.jp/lab/epd/Eng/catalog/seed.shtml	Providing *Arabidopsis* transposon-tagged lines and activation tagging lines
French National Institute for Agricultural research (INRA) Arabidopsis Resource Center for Genomics	http://www-ijpb.versailles.inra.fr/en/cra/cra_accueil.htm	Providing *Arabidopsis* T-DNA lines (FLAG lines)
Bielefeld University	http://www.gabi-kat.de/	Providing *Arabidopsis* T-DNA lines (GABI-Kat lines)
SIGnAL(Salk Institute Genomic Analysis Laboratory): T-DNA Express	http://signal.salk.edu/cgi-bin/tdnaexpress	Finding T-DNA insertion sites
Searching literature databases		
NCBI PubMed Database	http://www.ncbi.nlm.nih.gov/pubmed/	3.8.1. Finding articles in PubMed
TAIR: Publication Search	http://arabidopsis.org/servlets/Search?action=new_search&type=publication	3.8.2. Finding publications in TAIR
TAIR: Textpresso Full-Text Search	http://www.textpresso.org/arabidopsis/	3.8.3. Searching full-text literature
Submitting your data or DNA/seed stocks		
TAIR: Submit Data	http://www.arabidopsis.org/submit/index.jsp http://www.arabidopsis.org/doc/submit/functional_annotation/123	3.9.1. Submitting data to TAIR
PMN: Submit Data	http://www.plantcyc.org/feedback/data_submission.faces	3.9.2. Submitting data to PMN
ABRC: Stock Donation	https://abrc.osu.edu/donate-stocks	3.9.3. Donating seed and DNA stocks to ABRC

Center (ABRC), located in North America, and the European (Nottingham) Arabidopsis Stock Centre (NASC) represent the two largest stock centers and essentially mirror each other's seed collections. The collections of both centers will be discussed in more detail in the next section (Subheadings 2.2.2 and 2.2.3). The RIKEN BioResource Center (BRC) Experimental Plant Division in Japan has some unique resources, e.g., lines overexpressing *Arabidopsis* full-length cDNAs (FOX), and operates under the restriction of Material Transfer Agreements (MTAs) (Table 1). The French National Institute for Agricultural Research (INRA) in France and the Bielefeld University in Germany distribute locally developed collections of T-DNA lines (FLAG and GABI-Kat, respectively) [13, 14]. Although historically both institutions restricted the distribution by requiring an MTA, these restrictions have been lifted either completely or for the greater part of their collections.

As with any web-based informatics resource, database content and tools change over time. The protocols described here use tools and data available in databases and stock centers as of December 2011.

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 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. A high-speed network connection is desirable to ensure faster data access. Up-to-date web browser software, such as Internet Explorer, Firefox, or Safari, is also 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. The browser must have cookies enabled if users want to log in and place stock orders through TAIR. JavaScript must also be enabled to use TAIR since TAIR makes extensive use of this feature. See http://www.arabidopsis.org/help/index.jsp for information on properly configuring your browser. Note that for other databases mentioned in this chapter, there may be specific browser preferences.

2.2 Databases and Stock Centers

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. This section describes three commonly used *Arabidopsis* resources. Additional databases can be found in Table 1.

2.2.1 The Arabidopsis Information Resource

TAIR (http://www.arabidopsis.org) is a comprehensive web resource for the biology of *A. thaliana* [4–8]. It provides a centralized, curated gateway to *Arabidopsis* biology, research materials, and community members. Data available from TAIR includes the complete *Arabidopsis* genome sequence along with gene structure, gene product information, metabolism, expression data, genome maps, genetic and physical markers, publications, and information about the *Arabidopsis* research community. In addition, seed and DNA stock information and ordering from the Arabidopsis Biological Resource Center (ABRC) are fully integrated into TAIR. TAIR is a curated database; data are processed by biocurators with biology Ph.Ds who ensure their accuracy. TAIR data come from a variety of sources including in-house manual curation of published literature and sequence data, locally run computational pipelines for annotating gene structure and function, integration of data from other biological databases and resources (GenBank, ABRC, Gene Ontology Consortium, etc.), and submissions from the research community. TAIR also provides researchers with an extensive set of data visualization and analysis tools. A comprehensive guide on how to use TAIR is available [15].

2.2.2 The Arabidopsis Biological Resource Center

The ABRC collects, preserves, reproduces, and distributes diverse seed and other resources for *A. thaliana* and related species. The center is located at The Ohio State University in Columbus, Ohio, USA. The ABRC serves a dynamic community of plant researchers with a common goal to understand the basic processes of flowering plants, as well as to apply this understanding to further crop improvement. Seed stocks at the ABRC include classical mutants (*see* **Note 1**), natural accessions, T-DNA and transposon insertion collections, mapping populations, the TILLING collection, and seeds from related species (e.g., *Arabidopsis arenosa*, *Brassica rapa*, *Capsella rubella*). Other resources include cell suspension cultures, protein chips, full-length cDNA and ORF clones in recombination-ready and expression vectors, expressed sequence tagged (EST) and bacterial artificial chromosome (BAC) clones of *Arabidopsis* and related species, phage and plasmid libraries, and diverse vectors for cloning and expression. In addition, the ABRC has recently started distributing educational resources. Due to a large demand, this type of resource will be expanded further. This example illustrates how the resources provided by the ABRC closely track the emerging needs of the community. Seed resources are exchanged with the European Arabidopsis Stock Centre (NASC)

in Nottingham, UK. Researchers in the Americas are required to order seed stocks from ABRC, while researchers in Europe are required to order seeds from the NASC, but both can order DNA and other stocks from either center. Researchers outside of the Americas and Europe may order seed and other resources from either the ABRC or the NASC. The ABRC stock information and ordering system are hosted by TAIR (http://www.arabidopsis.org), and all functions can be accessed through the ABRC Stocks drop-down menu on the right side of the menu bar at the top of most TAIR pages.

2.2.3 The European Arabidopsis Stock Centre (NASC)

The European (Nottingham) Arabidopsis Stock Centre (NASC) provides *Arabidopsis* seed and information resources to the plant research community in coordination with the ABRC as described in the previous section. The NASC's stock collection includes seeds of *A. thaliana* and related species, tomato seed resources, DNA clones, and diverse cloning vectors. In addition, the NASC provides an International Affymetrix GeneChip hybridization service for a wide range of species including *Arabidopsis* and many other plants [16]. The data they collect through their hybridizations as well as other user-supplied *Arabidopsis* data are made publicly available through their NASCArrays database. The NASC stock information, ordering, and NASCArrays database are available at http://www.arabidopsis.info.

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 the following: What's the sequence and structure of my gene? What type of protein 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.

3.1 Finding Comprehensive Information About Arabidopsis Genes

After over 12 years of development, TAIR now serves as a central access point for *Arabidopsis* data. The TAIR home page (http://www.arabidopsis.org) is the main entry point to the database. The navigation toolbar provides easy access to the eight major functionalities: Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. When mousing over each item in the toolbar, a drop-down menu appears with clickable submenus that lead to a variety of dataset, tools, and external links. Log-in is not required for searching and viewing data but is required for ordering DNA or seed stocks from the ABRC and for submitting gene

functional data. Here, we describe how to use the TAIR Gene Search tool and locus detail page to find information about *Arabidopsis* genes.

3.1.1 Finding Genes in TAIR by Name

TAIR's locus detail page represents the most comprehensive starting point for a researcher to find out what is known about a gene. There are two commonly used ways to find genes and to get to the locus detail page: using the quick search and advanced Gene Search form.

1. To perform a quick search, go to the header on any TAIR page that has a quick search tool in the upper right corner. Enter the gene name (e.g., ABI3 or AT3G24650) in the text box and use the default "Gene" option on the drop-down menu. Click Search. A list of all matching records is displayed on a page titled TAIR Gene Search Results (*see* **Note 2**).
2. To perform a gene search using the advanced Gene Search form, on any TAIR page with a top navigation bar, select "Genes" from the Search drop-down menu (http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene).
3. Define the name search criteria. To search by name, choose "Gene name" as the option from the Search Name drop-down menu. This option is used to search by symbolic names (e.g., ABI3), full names (e.g., ABA INSENSITIVE 3), or AGI locus identifiers (e.g., AT3g24560). AGI (*Arabidopsis* Gene Identifier) locus identifiers are systematic names assigned based on chromosomal location.
4. 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 (e.g., ARF for Auxin Response Factor 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, switch from a "Gene name" search to a "description" search and choose the "contains" option.
5. Select the output format. The default values are 25 records, sorted by name. The position option can be used when finding genes by location.
6. Click "Submit Query" to start your search. All of the loci that match the query term will be displayed in a list of results (on a page titled TAIR Gene Search Results). Click on the locus name to view the locus detail page.

3.1.2 Using TAIR's Locus Detail Page to Find Information About a Gene

The locus detail page contains a wealth of information about a gene including its sequence, and function, and associated polymorphism, mutant phenotypes, and publication. This page also includes

links to a large number of external databases and tools. To see an example locus detail page, go to http://www.arabidopsis.org/servlets/TairObject?type=locus&name=AT3G24650. This section describes the typical data types displayed on the locus page.

1. Gene summary information: TAIR uses the AGI locus identifier (e.g., AT3G24650) as the primary gene name. Other names including both abbreviated gene symbols and the corresponding full names are displayed in the Other Names section. The Description field provides a short summary of the gene's function either manually composed by a curator or computationally generated (*see* **Note 3**).
2. Gene model information: A locus in TAIR refers to the physical location of an annotated gene on the chromosome. One locus can have several gene models or splice variants associated to it based on alternatively spliced mRNAs (e.g., At5g01810.1, At5g01810.2, At5g01810.3). The representative gene model for a protein coding gene is the gene with the longest coding sequence (CDS); for other gene types, the representative model is set as default to the .1 model. The Gene model page contains model-specific information such as exon–intron positions, protein domains, and gene model-specific function information. The Map Detail Image section displays the exon–intron structures of all gene models of a locus. Clicking on the image directs the user to GBrowse (*see* Subheading 3.2.2).
3. Gene function annotations: The Annotations section displays all of the Gene Ontology (GO) [17] and Plant Ontology (PO) [18] controlled vocabulary terms that describe the function and expression of the gene product. GO terms describe the molecular function, biological process, and subcellular localization of the gene product, while the PO consists of growth and development stages and plant structure terms capturing the temporal and spatial expression of the gene product. Detailed information including references and supporting evidence can be obtained by clicking on the "Annotation Detail" link located at the bottom of this section. How to find GO and PO annotations is described later (Subheadings 3.3.1 and 3.4.1).
4. Gene expression: Information about gene expression can be found in the Plant Ontology annotations section and in the RNA Data section. In the RNA Data section, array elements from pre-2005 one-channel and/or two-channel microarray experiments that map to the locus are listed. For elements whose expression has been analyzed across all experiments, the average log ratio of expression values, along with standard error, is provided. For these elements, links to the Expression Viewer (for finding similarly expressed genes) and Spot History (only available for microarray elements from arrays in the Stanford Microarray Database) are also available [15].

Please note that no new microarray expression datasets have been entered into TAIR since June 2005; instead TAIR provides links to high-quality gene expression resources in its External Links section on every locus page. The Associated Transcripts subsection within the RNA Data section lists full-length cDNAs and expressed sequence tags (ESTs) associated with the locus.

5. Nucleotide sequence: Links to the full-length CDS and full-length cDNA of the representative gene model plus the full-length genomic sequence are provided in this section.
6. Protein Data: This section displays the structural and physical characteristics of the protein encoded by the representative gene model, including length (amino acid), predicted molecular weight, isoelectric point, and domains. Click on the AGI name in the protein section to open a Protein detail page. Protein detail page provides more detailed information and the amino acid sequence for the representative gene model. To view nucleotide and protein data for other gene models, go to the specific gene model page.
7. Map Locations: This section displays the chromosome and coordinates of the locus for the maps on which it is found. TAIR provides three tools to view a gene in a whole-genome context: Map Viewer, Sequence Viewer, and GBrowse.
8. Polymorphism: This section contains all of the polymorphisms mapped to the locus. Both natural variations found in different ecotypes and induced mutations (e.g., T-DNA insertions) are shown. Note that by default this section only displays 15 entries, but a complete list can be obtained by clicking on the "See All" link right under this section's name.
9. Germplasm: This section provides information on all germplasms currently in the database associated with a locus and includes phenotype descriptions, mutant images, stock numbers, and ordering options when available.
10. External Link: TAIR links extensively to external sites that offer either alternative views of or different information about a locus, e.g., other *Arabidopsis* genome annotation databases, gene expression databases, functional genomics sites, and data analysis tools. Links to external sites and tools can also be found on the Portal pages (http://www.arabidopsis.org/portals/index.jsp).
11. Comments: This section contains statements contributed by registered TAIR users. Comments can be added to nearly all of the TAIR detail pages. This function can be used to report new data, errors, or omissions related to the displayed object.
12. Publications: Publications include published literature imported from PubMed, Agricola, and BIOSIS, along with abstracts from the International Conference on Arabidopsis

Research (ICAR). Only 15 entries are initially displayed on the locus page. At the bottom of the Publications section, click on "View Complete List" to see all records. Click on the title of the publication to view a detailed publication record that provides a link to the corresponding PubMed abstract or publication text when available.

3.2 Finding Gene Sequence and Structure Data

The primary source of *Arabidopsis* gene sequence and structure data at many biological databases is TAIR. In an ongoing effort to improve the annotation of the *Arabidopsis* genome, TAIR has released updated versions of the *Arabidopsis* gene set on a yearly basis since 2005 [7, 8]. TAIR's genome annotation is widely distributed to other major databases such as GenBank and UniProt. Therefore, these databases often have overlapping datasets. Here, it is described how to find sequence and structure data from TAIR.

TAIR provides gene sequence and structural data (i.e., the exon–intron architecture of a gene) in a variety of formats. DNA and protein data for an individual gene can be found on the locus and gene model pages (*see* Subheading 3.1.2). For those users interested in downloading complete sequence datasets, the TAIR ftp site provides sets of sequence files in FASTA format organized by TAIR release (e.g., TAIR10, TAIR9) and data types (e.g., coding sequence or CDS, cDNA, genomic DNA, and promoter regions) at ftp://ftp.arabidopsis.org/home/tair/Genes/. For users looking for a subset of gene sequences, the Sequence Bulk Download and Analysis tool generates sequence files based on a list of AGI locus identifiers (*see* Subheading 3.2.1).

In addition to sequence-based information, TAIR also provides structural information about each gene. The complete set of genome coordinates of each feature (such as exon, CDS, and 5′ untranslated region or 5′UTR) for every gene in the TAIR genome release is available in GFF3 format (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/). For a visual snapshot of a gene's exon–intron structure, users can refer to the TAIR locus page where a graphic displays the structural architecture of each splice variant annotated at that locus. TAIR also offers two different genome browsers: GBrowse and SeqViewer. While both browsers allow users to explore their genomic region of interest, the browsers are quite distinct and are used for different purposes. GBrowse is especially useful for analyzing a wide variety of data types that overlap with a chromosomal/gene region of interest. The tool contains a menu of datasets divided into sections such as expression data and sequence similarity data, which users can select to visualize in the main browser window. SeqViewer lends itself especially well for nucleotide-based analysis. Users can search the genome using either a name or a sequence, and thanks to the detailed "SeqViewer Nucleotide View," users can have a detailed look at the corresponding genome-based nucleotide sequence

decorated with annotated genes, T-DNA insertions from the SALK T-DNA insertion lines and other mutant collections, polymorphisms, and more. Detailed instructions on how to use SeqViewer are described elsewhere [15].

3.2.1 Retrieving DNA and Protein Sequence Data

TAIR's Sequence Bulk Download and Analysis tool allows the user to retrieve DNA and protein sequence data in bulk for a list of genes (or a single gene).

1. On any TAIR page with a top navigation bar, select "Bulk Data Retrieval" from the Tools drop-down menu. Then select "Sequences." Alternatively go directly to the URL http://www.arabidopsis.org/tools/bulk/sequences/index.jsp.
2. Enter individual or a set of AGI locus or gene model identifiers (e.g., At5g01810, AT1G01040.2) into the text box or upload a text file containing AGI locus or gene model identifiers. Select the desired data type from the Dataset drop-down menu (e.g., transcripts, coding sequence, and genomic locus sequences). This tool allows the user to retrieve sequences for the representative gene model, all gene models, or only the gene model matching the user query.
3. Select the FASTA or tab-delimited text output options. Click on "Get Sequences" to perform the search. More information on how to use the tool can be found by following the link to the Help document.
4. This tool can also be accessed whenever a user generates a "Gene Search Results" page by clicking on the "get all sequences" or "get checked sequences" button at the top of the page.

3.2.2 Searching for a Gene, Its Overlapping Transcripts, and Orthologous Genes in Other Organisms Using GBrowse

1. On any TAIR page with a top navigation bar, select GBrowse from the Tools drop-down menu (http://gbrowse.tacc.utexas.edu/cgi-bin/gb2/gbrowse/arabidopsis/). The GBrowse display is divided into five main sections: (1) Instructions, which provides directions and examples of GBrowse search queries; (2) Search, which allows the user to enter a query and select a data source; (3) Overview, which shows a graphical representation of the chromosome and region currently displayed; (4) Details, which provides a graphical representation of the genomic features in the selected region; and (5) Tracks, which allow the user to customize the display settings and select which features are displayed in the detail section [15].
2. To select a region of the genome to view, enter its name in the "Landmark or Region" search box (e.g., At1g01040). Select the desired dataset from the "Data Source" drop-down menu. The most recent TAIR genome release is the default data source. Clicking on "Search" will update the overview

and the detail display. Use the "Scroll/Zoom" feature to move along the chromosome or display a larger-/smaller-scale view of the genome.

3. Customize the GBrowse display. TAIR GBrowse has 11 track categories: Assembly, Community/Alternative Annotation, DNA, Expression, Gene, Genomic Features, Methylation and Phosphorylation, Orthologs and Gene Families, Sequence Similarity, Variation, and Analysis. New tracks may be added in the future. Each track category has multiple check boxes for different types of data. Mouse over a track name to display further information about the track. You can add or remove tracks from the detail display by checking or unchecking the required tracks and clicking the "Update Image" button. You can also upload your own annotation data in a special format to GBrowse using the "Add your own tracks" feature. For instructions on file format and uploading click on the "Help" link in this section.
4. To download the sequence in a particular region, go to the "Reports and Analysis" feature box and select "Download Decorated FASTA File" from the menu options. This file format allows you to highlight specific features of interest (e.g., coding regions in red) on the FASTA sequence file. Use the "Configure" option to select which features to highlight and the desired markup options such as font styles and background colors and then click "GO." The new web page will display the FASTA sequence for the region displayed in the detail view with the selected features highlighted.

3.2.3 Finding Related DNA or Protein Sequences in Arabidopsis

For sequenced genes with limited experimental data, one of the first steps toward understanding a gene's function is to search for evolutionarily related genes. The function of an unknown gene may be inferred from its similarity to a well-characterized homolog. Searching for similar DNA or protein sequences in *Arabidopsis* using local sequence alignment methods can be performed at TAIR and NCBI. These groups share some overlapping *Arabidopsis* datasets; but TAIR has some *Arabidopsis*-specific datasets not found at NCBI (http://www.arabidopsis.org/help/helppages/BLAST_help.jsp#datasets). These datasets are used by all of TAIR's sequence similarity programs (WU-BLAST, NCBI BLAST, FASTA, PatMatch) [15]. This section illustrates how to use TAIR's WU-BLAST tool to identify similar genes in *Arabidopsis*.

1. On any TAIR page with a top navigation bar, select WU-BLAST from the Tools drop-down menu (http://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.

3. Input your query. The tool accepts sequences or locus identifiers as inputs. 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 representative gene model; 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 non-redundant set of all known *Arabidopsis* proteins and includes all proteins generated through alternative splicing.
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 expected 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.

3.2.4 Finding Protein Structure and Domain Information

The function of an unknown gene may also be inferred from the presence of conserved domains. For example, proteins with an F-box domain (IPR001810, http://www.ebi.ac.uk/interpro/entry/IPR001810),) typically form part of an SCF E3 ubiquitin

ligase, whereas proteins with a kinesin motor domain (IPR001752, http://www.ebi.ac.uk/interpro/entry/IPR001752) may be involved in intracellular transport in association with the cytoskeleton. Additional sequence-based features, such as transmembrane domains or a KDEL endoplasmic reticulum retention signal, can be used to infer protein localization. Protein structural data including predicted domain can be found at various databases such as TAIR, NCBI, and InterPro. Here, we describe 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. On any TAIR page with a top navigation bar, select "Bulk Data Retrieval" from the Tools drop-down menu. Then select "Proteins" (http://www.arabidopsis.org/tools/bulk/protein/index.jsp).
2. Choose the output display. The output options include molecular weight, isoelectric point, intracellular locations, domains, number of transmembrane domains, UniProt ID, and SCOP's structural class. Selecting the HTML format option will display links to TAIR locus detail pages, protein sequences, SeqViewer graphical displays, and protein records in UniProt/Swiss-Prot, 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 1,000 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, check the box next to "Predicted Molecular Weight" and enter the lower and upper limits of the desired weight range in the adjacent text boxes.
4. Submit the query by clicking on the "Get Protein Data" button.

Protein domain annotations may not be consistent 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.

3.3 Finding Gene Ontology (GO) Annotations

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. The Gene Ontology vocabularies developed by the Gene Ontology Consortium (http://www.geneontology.org) have been widely adopted by many biological databases and are considered to be the standard for gene function annotation. GO describes three aspects

of a gene product: molecular function, biological process, and cellular component (subcellular localization) [17]. TAIR is the primary source of GO annotations for *Arabidopsis* genes. Additional sources of *Arabidopsis* GO annotations include TIGR (The Institute for Genomic Research) (*see* **Note 4**), UniProtKB-GOA (UniProt Knowledge Base Gene Ontology Annotation group) and the GO Consortium [19]. Members of the research community also contribute GO annotations through TAIR's journal collaboration program and through voluntary user submissions [20]. Annotations from all the above sources are displayed in TAIR. Users can also access these annotations from the central GO database using the AmiGO query tool for making cross species queries (http://amigo.geneontology.org/). This section describes how to find GO annotations at TAIR (*see* **Note 5** for information about how to correctly interpret them).

3.3.1 Finding GO Annotations

Users can view GO annotations for a single gene from its locus detail page and can also download TAIR's whole-genome GO annotation file from its ftp site (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/). This file is updated on a weekly basis. To retrieve GO annotations for a specific gene or set of genes, use TAIR's Gene Ontology Annotations Search tool.

1. On any TAIR page with a top navigation bar, select "Gene Ontology Annotations" from the Search drop-down menu (http://www.arabidopsis.org/tools/bulk/go/index.jsp).
2. Input the locus identifier(s) in the query box. Type, paste, or upload a file containing your list of locus identifiers.
3. Define output options. Select HTML to view hyperlinked results. Choose text for saving the results as a text file.
4. To obtain a list of annotations, click on the "Get all GO Annotations" button at the bottom of the page.
5. Alternatively, instead of getting a list of all annotations, the genes can be grouped into broader categories based on their annotations. After inputting the locus identifiers (**step 2** above), choose "HTML" output and click the Functional Categorization button. The functional categorization table data can be transformed into a pie chart by clicking on the "Draw Annotation Pie Chart" button at the top of the results page. Further details on functional categorization based on GO annotations are described in Chapter 5.

3.3.2 Finding Genes Annotated to Related Functions or Processes

By using structured controlled vocabularies, GO annotations allow researchers to quickly find what genes may act in a pathway (genes annotated to the same biological process term) or have similar function (genes annotated to the same molecular function term). For example, ERA1 (AT5G40280) encodes a protein

farnesyltransferase; mutants have low prenylation levels and defects in meristem organization and abscisic acid-mediated responses [8, 21–23]. A researcher may want to know the following: What other genes might be involved in prenylation, and do they act in the same or another pathway?

1. On any TAIR page with a top navigation bar, select "Keywords" from drop-down menu under Search (http://www.arabidopsis.org/servlets/Search?action=new_search&type=keyword).
2. Enter the term (keyword) "farnesyltransferase" in the text box and choose "contains" (an inexact search) from the drop-down menu to the left of the text box (*see* **Note 6**). Restrict the keyword category to "GO Molecular Function" and click the "Submit Query" button.
3. The Keyword Search Results page displays terms retrieved along with a count of data objects (loci, publications, annotations) annotated to that term and to its child terms. Click "loci" to display the genes annotated to "farnesyltranstransferase activity" and its child terms (e.g., farnesyl-diphosphate farnesyltransferase activity). Click on the "Download All" button on the result page to save the list of all loci.
4. On the Keyword Search Results page, click on "treeview" to view the term in a hierarchical tree view. Click on the plus sign next to a term to expand the node and display all of the child terms. To display genes annotated to each of the parent and child terms, select the "loci" radio button at the top of the tree view page and then click on the Display button. The display will show a count of the number of loci annotated to each term and the number of loci annotated to the children of each term. Click on the numbers to view more details.

The above example used a GO molecular function term "farnesyltranstransferase activity" to show how to find genes sharing similar function by searching for genes annotated to the same function term. Similarly searching for genes annotated to a process term (e.g., protein prenylation) will retrieve a list of genes involved in the related process.

3.4 Finding Information About Gene Expression

An important method of finding functional information comes from the analysis of gene expression data (*see* **Note 7**). There are many reasons to analyze these data, such as finding the expression pattern 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. A number of methods have been applied to study gene expression in *Arabidopsis* including low-throughput methods such as Northern blot, reverse transcription-polymerase chain reaction

(RT-PCR), in situ hybridization, and various reporter assays (e.g., GFP or green fluorescence protein, GUS or ß-glucuronidase reporters) and high-throughput methods such as DNA microarray analysis or RNA-Seq. Expression data obtained by the use of low-throughput methods can be found mainly in the literature. Some of these data in published literature have been captured in the form of Plant Ontology (PO) annotations through TAIR's literature curation effort. High-throughput DNA microarray data are for the most part stored in databases allowing for download and further analysis. Some of the DNA microarray expression data have also been converted into PO annotations and can be found in TAIR. For example, TAIR contains close to half million PO annotations based on the AtGenExpress microarray data (http://www.weigelworld.org/resources/microarray/AtGenExpress/) [24].

3.4.1 Finding an Expression Pattern by Searching for Plant Ontology Annotations

Following the model of Gene Ontology, the Plant Ontology Consortium (POC; http://www.plantontology.org/) has developed an ontology of controlled vocabulary terms for plant structure as well as growth and developmental stages [18]. Examples of plant structure terms are leaf, leaf stomatal complex, phyllome vascular system, etc. Examples of growth and developmental stages terms are seedling shoot emergence stage, late rosette growth, etc. In TAIR, PO are used to annotate gene expression data from low-throughput experiments such as Northern blot and reporter assays as well as high-throughput data from DNA microarray experiments and proteomics studies. PO annotations are displayed on the locus detail page along with GO annotations in the Annotations section. To retrieve PO annotations for a set of genes, use the Plant Ontology Annotations Search tool accessible from the Search drop-down menu (http://www.arabidopsis.org/tools/bulk/po/index.jsp). To find genes co-expressed in the same tissue or developmental stage, use the Keyword Search tool described previously (Subheading 3.3.2) by simply replacing a GO term with a PO term. TAIR's whole-genome PO annotation file is available for download from its ftp site (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Plant_Ontology/).

3.4.2 Finding DNA Microarray Data

DNA microarrays are one of the most powerful tools for investigating the expression pattern of thousands of genes in parallel, and microarray experiments are now commonly performed in many *Arabidopsis* laboratories. A vast amount of DNA microarray data has been generated, either through coordinated community effort such as the AtGenExpress project (http://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp) or as a result of individual research projects carried out by numerous laboratories. *Arabidopsis* microarray data can be found in several public repositories. TAIR provides access to experimental results from

both cDNA- and Affymetrix-based platforms of microarray data that TAIR received before June 2005. Newer and more comprehensive data can be found in NASCArrays (http://affy.arabidopsis.info/narrays/experimentbrowse.pl) [16], ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) [25], and GEO (http://www.ncbi.nlm.nih.gov/geo/) [26]. The emphasis of these public databases with microarray data is to provide long-term storage and access to publicly available data. There are many other academic and commercial groups that have focused on developing advanced analysis tools for mining microarray datasets. Notable examples include Genevestigator (https://www.genevestigator.com) [27] and eFP Browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) [28]. These tools will be covered in Chapter 5. This section shows how to use the TAIR microarray database to find expression profiles of a gene or genes in specific experiments.

1. Start at the TAIR Microarray Expression Search (http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression). This search can be used to find expression data for up to 100 genes using gene names, locus identifiers, microarray element names, or GenBank accession numbers.
2. Choose the default "Locus" from the Search by Name or GenBank Accession drop-down menu and enter At5g01810 in the query text box.
3. Choose Array Type/Design. This feature allows the search to be restricted to a specific type of arrays. Choose the default option (Affymetrix GeneChips, any design).
4. Limit Search by Expression Values (optional). This option allows one to adjust expression value parameters for either Affymetrix or cDNA arrays. Since this example involves Affymetrix data, use the parameter selections for this type of array. In the Detection section, choose Present, which will only include data from hybridizations where the transcript was detected.
5. Limit search by Experiment Parameters (optional). This is an advanced option to restrict a search to only certain experiments. If no limits are imposed, all the experiments in the database are searched.
6. Select the output options (optional).
7. Submit the query. The summarized results include array name (locus identifier), information about the experiment design (Experiment Name, Sample Variables), and specific data for each experiment. Click on the links to go to respective detail pages. The results can be downloaded in text format by clicking the check boxes for the records of interest and then clicking Download Checked.

3.5 Obtaining Information About Metabolism in Arabidopsis

There are a number of different databases that focus on providing information related to metabolism and metabolites in *Arabidopsis* including AraCyc and PlantCyc from the Plant Metabolic Network (PMN) [29], Arabidopsis Reactome [30], KEGG [31], KaPPA-View4 [32], MetNet [33], MetaCrop [34], and KNApSAck [35]. Although these resources may each offer specific benefits and their combined use might be ideal for optimal data analysis, based on the historical and ongoing connection between TAIR and Pathway Tools/PMN databases, this section will only describe how to access information from PMN databases with a focus on PlantCyc and AraCyc. PlantCyc can house biochemical data for all plant species, whereas AraCyc serves as a metabolic encyclopedia of *Arabidopsis* [29, 36–38]. Both databases provide information about genes, enzymes, compounds, reactions, and pathways that can have experimental and/or computational support. Semiannual releases, including the latest in July 2013, continue to improve upon the depth, breadth, accuracy, and coverage of these resources. In many cases, links are provided to connect these items found in the PMN to outside metabolism resources such as KEGG, BRENDA, ChEBI, and PubChem, as well as to more general databases such as TAIR, Phytozome, and UniProt.

3.5.1 Finding Information About Metabolic Pathways by Name

Although plant metabolism can only be completely described through an extremely dense and highly interconnected metabolic web, many scientists want to search for "pathways" that describe a comprehensible subset of connected reactions. These can be found in AraCyc from TAIR or in AraCyc or PlantCyc directly through the PMN.

1. From any TAIR page, enter the common name of a pathway (e.g., chlorophyll biosynthesis) or a prominent compound expected to be in the pathway (e.g., ascorbate) in the Quick Search tool in the header. Select "Metabolic Pathways" from the drop-down menu of search types and click the "Search" button. The search by default is a "contains" search, so, on the results page, all pathways, enzymes, reactions, and compounds associated with the input keyword will be retrieved. In the case of ascorbate, four different pathways are retrieved.
2. The same search can be performed from within the Plant Metabolic Network (www.plantcyc.org). From any PMN page, enter the search term (e.g., "ascorbate") in the Quick Search bar in the header, select the database to query, and click on "Quick Search" or "Search" (*see* **Note 8**). Again the default search will return all entries in the database that "contain" the term including pathways, enzymes, compounds, and/or reactions.
3. To learn more about a specific pathway, click on its name in the search results. This opens a page that provides a diagrammatic

representation of the pathway, evidence code(s), taxonomic information, a curator-written summary, literature references, and more. When a pathway page is initially opened, an overview diagram that lacks detailed information about enzyme identities, chemical structure, etc., may be shown. Click on the "More Detail" button one or more times to display the pathway with increasing amounts of information. When enzyme names appear (in gold), they are shown in bold if they are supported by experimental evidence or non-bold face type if they are supported by computational predictions. Each item on the pathway can be clicked on to open another page with more information, such as an "enzyme detail page."

3.5.2 Finding Information About Metabolic Pathways Based on Pathway Properties

To find a specific pathway or group of pathways that cannot be identified solely by name, at least four additional search strategies are available.

1. The Pathway Search page enables a user to select one or more pathways based on a variety of criteria. To access it from any page, expand the "Search" drop-down menu and choose "Pathways" (http://pmn.plantcyc.org/pwy-search.shtml). On the resulting page, nine different gray headers describe the type of filtering criteria available. To use one or more types of filter, click on the small box, e.g., to the left of the text that says "Search/Filter by number of reactions," and enter the desired restrictions. Multiple criteria can be combined before clicking on the "Submit Query" button.
2. The Advanced Search page gives users even more power to generate detailed requests. To access it from any page, expand the "Search" drop-down menu and choose "Advanced Query" (http://pmn.plantcyc.org/query.shtml). Several steps must be taken to construct a query in Section 1 of the page, beginning with choosing the appropriate database to search. Multiple "conditions" may be included in the search using the "add a condition" button and may be connected through Boolean operators. Once the request has been formulated, select the columns of data to output and choose a column to sort by in Section 2 of the page. Specify the output format (html or tab delimited) in Section 3 and then click on "Submit Query" to initiate the search. It should be noted that a familiarity with the underlying structure of the Pathway Tools database facilitates the use of this search tool.
3. Pathways can also be identified based on their membership in a particular class, such as "Amino Acids Biosynthesis" by using the Pathway Ontology Browser. To access it from any page, expand the "Search" drop-down menu and choose "Browse Ontologies" and then "Pathway Ontology." In the

resulting page, navigate through the ontology by clicking on any plus sign to expand a category and any minus sign to contract it.

4. Experimental data can also be used to highlight specific pathways that may be of interest to a user. Briefly, quantitative or qualitative results from transcriptomic, proteomic, and metabolomic experiments can be projected onto the entire *Arabidopsis* metabolic map using the "Metabolic Map/Omics Viewer" present under the "Tools" menu. A tutorial for this procedure is available at the PMN and has been described in previous publications [15].

3.6 Finding and Ordering Seed Resources from the Arabidopsis Biological Resource Center

The ABRC provides access to thousands of seed stocks which can be identified through a number of different search strategies at TAIR. Queries can be entered into the quick search bar in the header, or using the advanced Seed/Germplasm search, located on the ABRC Stocks drop-down menu on the TAIR navigation bar. The quick search allows searching by germplasm or polymorphism name or seed stock number. In addition to this, the advanced Seed/Germplasm search allows searching by germplasm/seed stock-associated information such as donor name, gene name, allele name, and phenotype. Searches can be limited by species, by germplasm type, and by a range of other attributes including genetic background, mutagen, and genotype. A specific search for ecotypes allows searching for natural variants of *A. thaliana* and related species by donor or germplasm attributes. The search can be limited by location and habitat. Search results pages for both germplasm and ecotype searches include check boxes for ordering and links to detail pages. Detail pages also contain links to other relevant information, for example, to clone detail pages for transgenic germplasm and to community detail pages for donors.

Stock-browsing functions are also supported by ABRC's catalog that can be accessed from the ABRC Stocks drop-down menu in the navigation bar available on most TAIR pages. Seed stocks in the catalog are divided into eight categories and include a range of different types of mutants, mapping lines, transgenic and RNAi lines, natural accessions, and seeds from other closely related species. Some sections link to detail pages with check boxes for ordering. Other sections link to summary pages describing available resources in that category with tips for finding them through advanced searches.

Arabidopsis seed stocks with associated sequence information, such as flank sequenced insertion lines, can be found by searching using the AGI locus identifiers through TAIR's GBrowse genome viewer and are fully integrated in the TAIR database. GBrowse is accessible from the navigation bar under "Tools." Locus-associated polymorphisms are displayed on the T-DNAs/Transposons and Polymorphisms tracks. Clicking on a polymorphism on the viewer links out to the polymorphism detail page where the corresponding

germplasm/stock can be found and ordered. Germplasm names/stock numbers are also displayed on locus detail pages with check boxes for ordering. Stock numbers and germplasm names link to germplasm detail pages where specific information and an ordering button are displayed.

Individuals can access their own order history and invoices from their personal home page when logged in to the TAIR web site. Other TAIR users cannot access an individual's complete order history, but the order history for a specific stock can be accessed through a link on the germplasm detail page for that stock.

In addition to TAIR's Seed/Germplasm Search, the T-DNA Express (http://signal.salk.edu/cgi-bin/tdnaexpress) developed by the Salk Institute Genomic Analysis Laboratory (SIGnAL) is another popular tool that helps users to find mutant resources associated with specific loci or chromosome locations [3]. T-DNA Express provides links to directly connect users to the ABRC, NASC, or other appropriate stock center to order them. In a reciprocal manner, TAIR provides direct links to this tool from the External Links section of its Locus Detail page (*see* Subheading 3.1.2).

3.7 Finding and Ordering Other (Non-seed) Resources from the Arabidopsis Biological Resource Center

Arabidopsis clone information is fully integrated into the TAIR database. For sequenced clones, links to clone detail pages can be accessed from TAIR's GBrowse genome viewer and from Locus detail pages. Clone detail pages contain a link to a stock detail page where information such as price, special handling, and other stock specific data can be found.

Clones and all other non-seed stocks can also be found through the TAIR quick search, but it is necessary to provide some name information, such as stock number or clone name. Advanced search options for these resources are provided by the TAIR DNA search (http://www.arabidopsis.org/servlets/Search?action=new_search&type=dna). Drop-down menus allow selection of the type of resource sought (e.g., vector, clone, or host strain), the species, and the type of information supplied (e.g., name, AGI, or stock number). A wide range of features to restrict the search are also available. Results pages from the search provide check boxes for ordering stocks, links to clone, vector and/or stock detail pages, as well as links out to NCBI for sequence information if available. Detail pages provide check boxes for ordering and links out to publications, images, and external web pages with information relevant to the stocks. The order history for a specific stock can be accessed through a link on the stock detail page.

DNA stocks can be found by browsing the ABRC catalog. They are divided into five categories, including libraries, clones, vectors, and host strains. The catalog provides links to detail pages with check boxes for ordering or to summary pages describing available resources in that category with tips for finding them through advanced searches. Access to other non-seed resources,

including protein chips, cell cultures, and educational resources, is also provided by the catalog. More details about educational resources developed by the ABRC can be obtained from the ABRC outreach portal at http://abrcoutreach.osu.edu.

3.8 Searching Literature Databases

Researchers have published a wealth of data about all aspects of *Arabidopsis* physiology, biochemistry, and development. Databases such as PubMed, Agricola, and 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 (http://www.ncbi.nlm.nih.gov/pubmed/) is the primary database for life-science literature. At the end of 2011 the number of *Arabidopsis* publications in PubMed totaled over 36,000. PubMed has a powerful search interface and links to the rest of the databases within the NCBI system, such as sequence and expression databases. 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/bsd/disted/pubmed.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 International Conference on *Arabidopsis* Research, defunct *Arabidopsis* electronic journals (The Arabidopsis Information Service and Weeds World), books, and dissertations. The following sections describe how to find *Arabidopsis* publications in PubMed and TAIR.

3.8.1 Finding Articles in the NCBI PubMed Database

1. Start at the PubMed search page (http://www.ncbi.nlm.nih.gov/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, use the advanced search page (http://www.ncbi.nlm.nih.gov/pubmed/advanced). The Search Builder allows users to build complex queries.
3. Finding the article text and saving relevant citations: The default display format is a summary of the citation. The complete citation, including available abstracts, can be viewed by clicking on the titles. 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.

3.8.2 Finding Arabidopsis Publications Using TAIR's Publication Search

1. On any TAIR page with a top navigation bar select "Publication" from drop-down menu under Search (http://arabidopsis.org/servlets/Search?action=new_search&type=publication).
2. To search with a specific author's name or phrase, enter the desired terms in the text query boxes and choose the field to search from the drop-down menu (abstract, author, journal/book title, title, title/abstract, URL for electronic publications, journal, or PubMed ID). 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. The 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 keywords and genes. For articles with a PubMed ID, a link to the PubMed database is also provided.

3.8.3 Searching Full-Text Arabidopsis Literature Using Textpresso

Textpresso is an information extracting and processing package for biological literature [39]. Textpresso for *Arabidopsis* allows users to search over 40,000 abstracts and 27,000 full-text publications in TAIR as of August 2011.

1. To use this tool, go to http://www.arabidopsis.org/ and select "Textpresso Full Text" from Tools drop-down list. Alternatively go to http://www.textpresso.org/arabidopsis/.
2. Enter the search term in the Keywords query box. Textpresso is extremely useful for tracking down specific information like the mutation sites in certain alleles. For example, enter SALK_099519, and click "Search." Sentences that contain the matching keyword are displayed together with bibliographic information so that users can quickly confirm the usefulness of a particular paper and link directly to the full text, if they have an appropriate subscription to the journal in question. At the Textpresso site, searches can be narrowed by searching in specific keyword categories (mouse over "List >") including

Arabidopsis gene names, Gene Ontology and Plant Ontology (terms), or a combination of keywords. Advanced search options are described in the User Guide accessible from the top navigation bar.

3.9 Submitting Your Data or DNA/Seed Stocks

Funding agencies such as the National Science Foundation (NSF) have invested heavily in the development of community resources such as biological databases and stock centers. These resources play a crucial role in driving research forward by providing access to data and research materials. The long-term sustainability of such resources depends upon contributions by the research community. In an age when data influx has outstripped the organizational ability of the staff of any one database, it is essential to involve the research community in the data collection and curation process. It is important that researchers share their findings not only through publication but also by contributing their data directly to scientific databases. This section describes how to submit your data and/or DNA/seed stocks to various databases.

3.9.1 Submitting Data to TAIR

TAIR accepts a wide range of data types including gene function, structure, interaction partners, expression patterns, markers, phenotypes, and several others. Instructions for data submission are available on the Submit Overview page (http://www.arabidopsis.org/submit/index.jsp), accessible from the Submit drop-down menu in the top navigation bar.

TAIR provides several ways for researchers to submit their data. For gene function data submission, the use of the online submission tool (http://www.arabidopsis.org/doc/submit/functional_annotation/123) is encouraged. This tool requires the submitting user to log into the TAIR system with a registered user ID, which provides an automatic provenance for the submitted annotations. Reference information (PubMed ID or DOI identifier) is also required. The use of DOIs allows a user to submit annotations before public release of the manuscript; however, the annotations are only released from TAIR upon publication of the corresponding article.

Users can also prepare various types of data for submission formatted according to the guidelines listed on the Submission Overview page or download and use the preformatted Excel spreadsheets available there [15]. Data can then be submitted to TAIR by e-mail to curator@arabidopsis.org. In addition, each data detail page contains a Comments section; registered TAIR users can submit comments by clicking on the "Add My Comment" button. Comments submitted are immediately displayed in the Comments section of the detail page.

For corrections to existing data, users may contact TAIR by e-mail to curator@arabidopsis.org.

3.9.2 Submitting Data to the PMN

The Plant Metabolic Network is eager to receive data submissions of published findings related to pathways, enzymes, reactions, or compounds found in plants. To help researchers submit these data types, three Excel forms and simple instructions are provided on the Data Submission page (http://www.plantcyc.org/feedback/data_submission.faces). This can be accessed from the "Submit Data" heading on the menu bar. Submitters are encouraged to enter the data on the forms, save them locally, and then send them to the PMN. The forms may be e-mailed or may be uploaded and submitted via the Feedback Form (http://www.plantcyc.org/feedback/feedback_form.faces) that can also be found on the "Submit Data" menu. Although thoroughness on the forms is appreciated, incomplete forms are always accepted. In addition, supporting materials, such as .gif files that depict pathway layouts or .mol files that provide compound structures, can also be submitted. The PMN also welcomes experts to volunteer to help review particular domains of metabolism to check for completeness and accuracy.

Feedback and corrections concerning data found in the PMN can be submitted using the Feedback Form or through a direct e-mail to curator@plantcyc.org.

3.9.3 Donating Seed and DNA Stocks to the ABRC

The ABRC accepts all *Arabidopsis* seed resources and is particularly interested in receiving confirmed insertion mutants, characterized mutants, transgenic lines, and cDNA/ORF clones. For other types of resources, it is necessary to contact the stock center in advance to ensure that the resource can be accommodated. All seed resources are shared with NASC after propagation at the ABRC or immediately if enough seed is supplied. Other resources may also be shared with NASC if requested by NASC customers. The ABRC has developed stock donation forms to collect data associated with stock donations. This data is curated by ABRC staff and uploaded to TAIR within a month of receiving the material. Donated stocks are being made available for distribution either at the time related data is uploaded or upon amplification. Although it is preferable that donors fill out ABRC donation forms, a simple donation form is available for published resources and data in other formats is accepted, particularly for large collections of stocks. Links for downloading ABRC donation forms are available from the ABRC Stocks drop-down menu. A donation form for a contribution of educational materials for high school and undergraduate-level classes has recently been developed and is available upon request.

4 Notes

1. Classical mutants are mostly characterized and published mutants derived from forward genetic screens utilizing populations generated with various mutagens (X-rays, fast

neutrons, ethyl methanesulfonate or EMS, agrobacterium transformation, etc.).

2. The quick search performs a name search for most of the objects in the TAIR database (e.g., Genes, Clones, ESTs or BAC ends, People/Labs, Polymorphisms/Alleles, Germplasms, Ecotypes, Keywords, Genetic Markers, Proteins, Seed and DNA Stocks, and Vectors). By default, this is a "contains" search (a search for aba1 retrieves both ABA1 and ATRABA1A). This search is not limited to the name field. For example, when performing a quick search for "Gene," the gene description and keywords fields will be searched as well as the name. This is to avoid missing any potentially relevant results, but sometimes too many results are returned. To perform an exact name search, choose the "exact name search" option from the drop-down menu to the right of the search box. This option will only search the name field for all the data types listed in the drop-down menu [15].
3. The computational description contains the gene's full name, Gene Ontology and Plant Ontology terms, best BLAST-identified *A. thaliana* protein match, and the number of protein BLAST hits in other species (NCBI BLink) [15]. A computational description is only shown if the locus has not yet been curated manually. Users are especially welcome to submit suggested gene descriptions for loci that only have a computational description.
4. TIGR, now the J. Craig Venter Institute (http://www.jcvi.org/), no longer actively produces GO annotations for *Arabidopsis* genes, but past TIGR annotations are still stored in TAIR.
5. GO annotations can be divided into two broad categories: (1) annotations based on experimental data including results from low- and high-throughput experiments (e.g., DNA microarray and proteomics studies) and (2) computationally predicted annotations. Computational annotations are based on an in silico analysis of the gene product sequence and/or other data as described in the cited reference and may or may not be individually reviewed by a curator. For example, TAIR uses a combination of InterProScan and InterPro2GO mapping file to create GO annotations for proteins based on the presence of domains with mapped GO terms [8]. Such annotations are not reviewed on an individual basis by a curator. Alternatively, annotations can be made by a curator on an individual basis by examining relevant computational analyses (e.g., sequence alignment, protein family information). Computational annotations provide the basis to form testable hypothesis particularly for genes with little known experimental data. For example, AT3G24560 (RASPBERRY 3) is annotated to

the GO term "ligase activity, forming carbon–nitrogen bonds" based on an InterPro domain scan. A researcher can then design an experiment to test whether indeed this protein has ligase activity. The GO Consortium has developed a set of evidence codes to indicate how an annotation to a particular term is supported. In order to correctly interpret a GO annotation, it is essential to review the evidence code together with the GO term. For a complete list of evidence codes currently in use, go to http://www.geneontology.org/GO.evidence.shtml. In TAIR, annotations also include an evidence description. For example, an annotation with the evidence code "inferred from mutant phenotype" (IMP) may be further specified by including an evidence description "RNAi experiments." Since more than one gene may be affected by RNA interference, the GO annotation should be viewed with the understanding that the phenotype may be due to the loss of function of more than one homologous locus. An in-depth discussion on how to avoid the common misuse of GO is available [40].

6. Many of the GO terms exist as complex phrases. TAIR searches treat the entire entered term or phrase as a complete phrase rather than a set of words. Consequently, an "exact match" search will often not retrieve any entries. Therefore, using the "contains" option for keyword searches is recommended [15].
7. Gene expression data historically and most properly refers to the expression of gene transcripts; however, the expression of protein constructs and/or the analysis of proteomic experiments is also often grouped into this category.
8. The PMN offers a collection of PMN-generated pages (www.plantcyc.org/…) and Pathway Tools-generated pages (pmn.plantcyc.org/…) which have some differences, particularly in the header. Most notably, a simple drop-down menu is used to select a database to query via the Quick Search bar on PMN-generated pages, whereas the "change organism database" link can be used to select a new database to query on all Pathway Tools-generated pages.

Acknowledgements

This project was supported by the National Science Foundation (grant number DBI-0850219, DBI-0640769, IOS-1026003), the National Institute of Health National Human Genome Research Institute (NIH-NHGRI) (grant number 5P41HG002273-09), and the TAIR sponsorship program (http://www.arabidopsis.org/doc/about/tair_sponsors/413).

References

1. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815
2. The Multinational *Arabidopsis* Steering Committee (2011) The multinational coordinated *Arabidopsis thaliana* functional genomics project annual report 2011. http://www.arabidopsis.org/portals/masc/2011_MASC_Report.pdf
3. Alonso JM, Stepanova AN, Leisse TJ et al (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301:653–657
4. Garcia-Hernandez M, Berardini TZ, Chen G et al (2002) TAIR: a resource for integrated *Arabidopsis* data. Funct Integr Genomics 2:239–253
5. Huala E, Dickerman AW, 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 SY, Beavis W, Berardini TZ 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. Swarbreck D, Wilks C, Lamesch P et al (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36:D1009–D1014
8. Lamesch P, Berardini TZ, Li D et al (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. doi:10.1093/nar/gkr1090
9. Meinke D, Scholl R (2003) The preservation of plant genetic resources: experiences with *Arabidopsis*. Plant Physiol 133:1046–1050
10. Heazlewood JL, Verboom RE, Tonti-Filippini J et al (2007) SUBA: the *Arabidopsis* subcellular database. Nucleic Acids Res 35:D213–D218
11. Lu Y, Savage LJ, Larson M et al (2011) Chloroplast 2010: a database for large-scale phenotypic screening of *Arabidopsis* mutants. Plant Physiol 155:1589–1900
12. International Arabidopsis Informatics Consortium (2010) An international bioinformatics infrastructure to underpin the *Arabidopsis* community. Plant Cell 22:2530–2536
13. Samson F, Brunaud V, Balzergue S et al (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. Nucleic Acids Res 30:94–97
14. Kleinboelting N, Huep G, Kloetgen A et al (2011) GABI-Kat Simple Search: new features of the *Arabidopsis thaliana* T-DNA mutant database. Nucleic Acids Res. doi:10.1093/nar/gkr1047
15. Lamesch P, Dreher K, Swarbreck D, et al. (2010) Using the Arabidopsis Information Resource (TAIR) to find information about *Arabidopsis* genes. Curr Protoc Bioinformatics. Chapter 1:Unit1.11
16. Craigon DJ, James N, Okyere J, Higgins J et al (2004) A repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res 32:D575–D577
17. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
18. Jaiswal P, Avraham S, Ilic K et al (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp Funct Genomics 6:388–397
19. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome project: a unified framework for functional annotation across species. PLoS Comput Biol 5:e1000431
20. Ort DR, Grennan AK (2008) Plant physiology and TAIR partnership. Plant Physiol 146: 1022–1023
21. Cutler S, Ghassemian M, Bonetta D et al (1996) A protein farnesyl transferase involved in abscisic acid signal transduction in *Arabidopsis*. Science 273:1239–1241
22. Yalovsky S, Kulukian A, Rodriguez-Concepcion M et al (2000) Functional requirement of plant farnesyltransferase during development in *Arabidopsis*. Plant Cell 12:1267–1278
23. Ziegelhoffer EC, Medrano LJ, Meyerowitz EM (2000) Cloning of the *Arabidopsis* WIGGUM gene identifies a role for farnesylation in meristem development. Proc Natl Acad Sci USA 97:7633–7638
24. Schmid M, Davison TS, Henz SR et al (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet 37:501–506
25. Parkinson H, Sarkans U, Kolesnikov N et al (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 39:D1002–1004
26. Barrett T, Troup DB, Wilhite SE et al (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 39:D1005–1010

27. Hruz T, Laule O, Szabo G et al (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008:420747
28. Winter D, Vinegar B, Nahal H et al (2007) An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One 2:e718. doi:10.1371/journal.pone.0000718
29. Zhang P, Dreher K, Karthikeyan A et al (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol 153:1479–1491
30. Tsesmetzis N, Couchman M, Higgins J et al (2008) *Arabidopsis* reactome: a foundation knowledgebase for plant systems biology. Plant Cell 20:1426–1436
31. Masoudi-Nejad A, Goto S, Endo TR et al (2007) KEGG bioinformatics resource for plant genomics research. Methods Mol Biol 406:437–458
32. Sakurai N, Ara T, Ogata Y et al (2011) KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic Acids Res 39:D677–684
33. Wurtele ES, Li L, Berleant D et al (2007) MetNet: systems biology software for *Arabidopsis*. In: Nikolau BJ, Wurtele ES (eds) Concepts in plant metabolomics. Springer, Berlin, pp 145–158
34. Grafahrend-Belau E, Weise S, Koschützki D et al (2008) MetaCrop: a detailed database of crop plant metabolism. Nucleic Acids Res 36:D954–958
35. Shinbo Y, Nakamura Y, Altaf-Ul-Amin M et al (2006) KNApSAcK: A comprehensive species-metabolite relationship database. In: Saito K, Dixon RA, Willmitzer L (ed) Plant metabolomics. Berlin, Springer, pp 165–181. doi: 10.1007/3-540-29782-0_13
36. Karp P, Paley S, Romero P (2002) The pathway tools software. Bioinformatics 18: S225–S232
37. Mueller LA, Zhang P, Rhee SY (2003) AraCyc. A biochemical pathway database for Arabidopsis. Plant Physiol 132:453–460
38. Zhang P, Foerster H, Tissier C et al (2005) MetaCyc and AraCyc: metabolic pathway databases for plant research. Plant Physiol 138:27–37
39. Müller HM, Kenny EE, Sternberg PW (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2:e309
40. Rhee SY, Wood V, Dolinski K et al (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9:509–515