Funct Integr Genomics (2002) 2:239–253
DOI 10.1007/s10142-002-0077-z

REVIEW

**Margarita Garcia-Hernandez · Tanya Z Berardini
Guanghong Chen · Debbie Crist · Aisling Doyle
Eva Huala · Emma Knee · Mark Lambrecht
Neil Miller · Lukas A. Mueller · Suparna Mundodi
Leonore Reiser · Seung Y. Rhee · Randy Scholl
Julie Tacklind · Dan C. Weems · Yihe Wu · Iris Xu
Daniel Yoo · Jungwon Yoon · Peifen Zhang**

TAIR: a resource for integrated *Arabidopsis* data

Received: 11 June 2002 / Accepted: 12 August 2002 / Published online: 3 October 2002
© Springer-Verlag 2002

Abstract The *Arabidopsis* Information Resource (TAIR; http://arabidopsis.org) provides an integrated view of genomic data for *Arabidopsis thaliana*. The information is obtained from a battery of sources, including the *Arabidopsis* user community, the literature, and the major genome centers. Currently TAIR provides information about genes, markers, polymorphisms, maps, sequences, clones, DNA and seed stocks, gene families and proteins. In addition, users can find *Arabidopsis* publications and information about *Arabidopsis* researchers. Our emphasis is now on incorporating functional annotations of genes and gene products, genome-wide expression, and biochemical pathway data. Among the tools developed at TAIR, the most notable is the Sequence Viewer, which displays gene annotation, clones, transcripts, markers and polymorphisms on the *Arabidopsis* genome, and allows zooming in to the nucleotide level. A tool recently released is AraCyc, which is designed for visualization of biochemical pathways. We are also developing tools to extract information from the literature in a systematic way, and building controlled vocabularies to describe biological concepts in collaboration with other database groups. A significant new feature is the integration of the ABRC database functions and stock ordering system, which allows users to place orders for seed and DNA stocks directly from the TAIR site.

Keywords *Arabidopsis* · Genome · Database · Information management

M. Garcia-Hernandez (✉) · T.Z. Berardini · A. Doyle · E. Huala
M. Lambrecht · L.A. Mueller · S. Mundodi · L. Reiser
S.Y. Rhee · J. Tacklind · I. Xu · D. Yoo · J. Yoon · P. Zhang
TAIR, Carnegie Institution of Washington, 260 Panama Street,
Stanford, CA 94305, USA
e-mail: garcia@acoma.stanford.edu
Fax: +1-650-3256857

G. Chen · N. Miller · D.C. Weems · Y. Wu
National Center for Genome Resources,
2939 Rodeo Park Drive East, Santa Fe, NM 87505, USA

D. Crist · E. Knee · R. Scholl
Arabidopsis Biological Resource Center,
The Ohio State University, 309 Botany and Zoology Bldg.,
1735 Neil Avenue, Columbus, OH 43210, USA

Introduction

Arabidopsis thaliana is a small flowering plant, closely related to many crop plants such as cauliflower, cabbage, and broccoli. In the last 20 years, *Arabidopsis* has progressed from being an obscure weed to being a respected member of the "Security Council of Model Genetic Organisms" (Meinke et al. 1998). Its short life cycle, easy genetic manipulation, facile transformation, and more recently, the decoding of its complete genome, has made *Arabidopsis* a model organism for many academic and commercial laboratories all over the world. About 11,000 researchers representing nearly 4,000 organizations worldwide are registered as TAIR users and use *Arabidopsis* in their research. Many biotechnology companies are counting on *Arabidopsis* research to help solve practical problems related to agriculture, energy, and the environment (Meinke et al. 1998).

The volume of data resulting from *Arabidopsis* research has grown exponentially over the years, largely due to whole genome analyses. For example, there are currently about 400,000 nucleotide sequences, 200,000 mutant lines, 4,000 genetic markers, 90,000 polymorphisms, more than 600 microarray experiments, and over 14,000 publications available to the researchers.

The extensive body of data derived from experimentation using *Arabidopsis* has been, for the most part, scattered in different places or not accessible to the research community. The need for a comprehensive and centralized resource for *Arabidopsis* information has been long recognized by the *Arabidopsis* community (http://arabidopsis.org/info/carnegieworkshop.html). The

Table 1 Selected TAIR URLs described in this paper

Name	Description	URL
TAIR search	Text-based database search	http://arabidopsis.org/servlets/Search?type=general&action=new_search
SeqViewer	View genomic sequence	http://arabidopsis.org/servlets/sv
MapViewer	View all maps	http://arabidopsis.org/servlets/mapper
BLAST	Sequence homology search	http://arabidopsis.org/Blast/; http://arabidopsis.org/wublast/
PatMatch	Sequence pattern search	http://arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl
AraCyc	Visualization of biochemical pathways	http://arabidopsis.org/tools/aracyc/
Bulk sequence download	Download locus sequences in bulk	http://arabidopsis.org/tools/bulk/
GO annotation download	Download gene ontology annotation of loci in bulk	http://arabidopsis.org/tools/bulk/go/
FTP	Download of complete datasets	ftp://ftp.arabidopsis.org/home/tair/
Seed stock search	Search ABRC seed stocks	http://arabidopsis.org/servlets/SeedSearcher
DNA stock search	Search ABRC DNA stocks	http://arabidopsis.org/servlets/StockSearcher
Stock order history	Search stock order history	http://arabidopsis.org/servlets/Order?state=search&mode=stock_number
Nomenclature and protein classification	Nomenclature resources	http://www.arabidopsis.org/links/nomenclature.html
Gene families	Gene family data	http://arabidopsis.org/info/genefamily/genefamily.html
2010 Projects	Information on functional genomic projects	http://arabidopsis.org/info/2010_projects/

Arabidopsis Information Resource (TAIR; http://arabidopsis.org) is an attempt to fulfill this requirement.

TAIR is a collaborative effort funded by the NSF to collect, curate and distribute information about *Arabidopsis* (Rhee 2000; Huala et al. 2001). It is the successor of the former *Arabidopsis thaliana* DataBase (AtDB; Flanders et al. 1998), which was created during the genome sequencing project to coordinate the sequencing effort and to integrate the *Arabidopsis* genetic and physical maps. TAIR has expanded to include new data types and analysis tools, and is committed to the curation of the *Arabidopsis* literature. Thus, TAIR is in transition from a genome sequence-centric perspective to one that focuses on all aspects of *Arabidopsis* biology. For detailed information about the TAIR project, including the full proposal, data sources, and projects under development, please visit http://arabidopsis.org/about/.

Since TAIR's inception, its usage has increased steadily from 20,000 web page visits per month in November 1999 (3 months after the project started) to about 500,000 per month in July 2002 (http://arabidopsis.org/usage/monthly/2002/05/01/index.html). The majority of users belong to academic USA and European domains (30% each), followed by Asia and corporate domains (10% each).

This paper describes TAIR's database contents and user interfaces, as well as the tools available for the analysis, visualization and retrieval of data offered by the resource. A summary of the relevant TAIR URLs is shown in Table 1.

Arabidopsis data available at TAIR

The current major data types stored in TAIR database are: genes, markers, polymorphisms, assignments (map positions), sequences, clones, stocks, proteins, expression, researchers and references. These data are integrated with each other and associated to references and attributions in the database. Table 2 summarizes the current number of entries of each major data type and their sources. The data stored in the database are available from TAIR's text-based search and visualization tools as described in the Tools section.

There are also some datasets that are not yet integrated in the database, which are available only from the static web pages (e.g. gene families from the user community and single nucleotide polymorphism and Landsberg *erecta* sequence from Cereon genomics) or from the FTP site. The data files available from the FTP site are listed in Table 3.

New data types are continuously added into the resource as we progressively expand the database structure to accommodate them. In addition, data updates and improved ways for accessing existing data types are incorporated on a regular basis. These additions are announced in the "Breaking News" section on the web site, and in some cases, to the *Arabidopsis* Newsgroup (http://www.arabidopsis.org/news/newsgroup.html).

The *Arabidopsis* genome

Arabidopsis thaliana was the first higher plant to have its genome fully sequenced (Arabidopsis Genome Initiative 2000), and the first example in which a multinational collaboration of public and private laboratories (Arabidopsis Genome Initiative –AGI) contributed to the sequencing effort.

TAIR contains the essentially complete genome of *Arabidopsis*, which now stands at 125 Mb interrupted by the centromeric regions, 3 5SrDNA regions and 11 other gaps and potential gaps. For an updated summary of the genome completion status see http://arabidopsis.org/

Table 2 Data available at TAIR's database and their sources

Data class	Data object type	Number of entries	Sources
Gene	Loci	28,595	TIGR, GenBank, AtDB, Meinke Lab, Literature, user submissions
	Gene models	37,363	
Nucleotide sequences	Genomic	226,193	GenBank, AtDB, TIGR, user submissions, Kazusa, Salk
	mRNAs	194,745	
	ESTs	174,122	
	Full-length CDS	8,271	
Clones	BACs	23,651	AtDB, GenBank, ABRC
	YACs	7,978	
	Lambda	97	
	P1	9,690	
	TAC	10,593	
	Plasmids	127,445	
	Cosmids	243	
Polymorphism	Digest pattern	2,906	AtDB, NASC, Stanford Sequencing Center, user submissions
	PCR product length	2,979	
	Visible	462	
	INDEL	392	
	Insertion	58	
	Substitution	40,247	
	SNP	3,434	
Genetic markers	Visible	463	AtDB, NASC, user submissions
	RFLP	643	
	AFLP	1,254	
	CAPS	290	
	RAPD	1	
	SNP	176	
	SSLP	184	
Protein data	Protein domains	9,073	TIGR, InterPro, SCOP, TargetP analysis, TAIR
	Subcellular locations, predicted physical properties (pI, molecular weight)	26,082	
	Sequences	26,082	
DNA stocks	BAC clones	50,000	ABRC, NASC
	ESTs	38,000	
	Libraries	25	
	Full-length cDNA clones	8,000	
	Mapped YACs	408	
	RFLP phage mapping lines	77	
	Cosmids	209	
	Pools of genomic DNA from T-DNA mutant populations	692	
Seed stocks	Characterized mutants	1,950	ABRC, NASC
	RI mapping lines	840	
	Tetrad lines	300	
	Sequence indexed T-DNA lines	140,000	
	Transposon and T-DNA stocks	35,825	
	Lines expressing transgenes	100	
	Natural accessions and other species	1,500	
References	Publications	14,430	AtDB, PubMed, Agricola, Biosis, NAL
	Database references	1,602	
	Analysis references	6	
	Communications	27,727	
Community	Persons	10,992	ABRC, AtDB, NASC, user submissions
	Organizations	4,081	
Physical map data	Mapped clones	10,871	User submissions, AtDB websites, Literature
	Mapped probes	3,539	
RI map data	Mapped markers	1,309	NASC

Table 3 Datasets available for downloading from TAIR's FTP site (ftp://ftp.arabidopsis.org/home/tair/) and their sources

Main data types	Data files	Sources
BLAST datasets	Whole *Arabidopsis* genome Genes (DNA)	TIGR
	Whole *Arabidopsis* genome CDS (DNA)	TIGR
	Whole *Arabidopsis* genome proteins (Protein)	TIGR
	AGI BACS	TIGR (Annotation units)
	AGI BACS	GenBank
	All *Arabidopsis* DNA, including ESTs and BAC ends	GenBank
	All *Arabidopsis* DNA without ESTs and BAC ends	GenBank
	All *Arabidopsis* proteins	GenPept, PIR, and SwissPROT
	All *Arabidopsis* BAC end sequences	GenBank and Kazusa
	All ESTs	GenBank
	All higher plant sequences (DNA)	GenBank
	Insertion flank sequences	
	1,000 bp upstream *Arabidopsis* genome CDS	TAIR
	3,000 bp upstream *Arabidopsis* genome CDS	TAIR
	1,000 bp downstream *Arabidopsis* genome CDS	TAIR
	3,000 bp downstream *Arabidopsis* genome CDS	TAIR
	Intergenic regions	TAIR
	Introns	TAIR
Other sequences	Chloroplast sequence	Kazusa, TIGR
	Whole chromosomes	TIGR
	Mitochondrial sequence	TIGR, Brennicke lab
	Clones	AGI
	Markers	AtDB
Genes	2010 projects	User submission
	Gene ontology	TAIR, GO, TIGR
	Gene families	User submission
	Gene name aliases	TAIR, literature, Meinke lab
	EST mapping	TAIR
	Genome annotation in XML	TIGR
Maps	SeqMap data	AtDB
	MapViewer data	TAIR
	SeqViewer data	TAIR
Microarrays	Affymetrix probes	Affymetrix, TAIR
	AFGC probes	AFGC, TAIR
Proteins	Protein domains, Id conversions and properties	InterPro, SCOP, TargetP analysis, TAIR
Ontologies	Anatomy and developmental controlled vocabularies	TAIR

info/agicomplete.html. The AGI sequencing effort predicted approximately 26,000 genes (Arabidopsis Genome Initiative 2000), of which more than 14,000 have been structurally verified by full-length cDNA sequencing (http://signal.salk.edu/SSP/index.html).

The sequenced genome was initially annotated by the AGI sequencers (Arabidopsis Genome Initiative 2000) and is now undergoing a major re-annotation of both structure and function by TIGR, MIPS and TAIR. The genomic sequence was assembled by TIGR (http://www.tigr.org/tdb/e2k1/ath1/) and MIPS (http://mips.gsf.de/proj/thal/db/index.html).

During the sequencing phase, some overlapping clone sequences were trimmed before submission to GenBank. Subsequently, during the initial assembly phase, some of these clones were increased in size at the overlapping regions by adding sequences from adjacent clones to facilitate annotation of genes lying near the ends of clones. The resulting units of sequence have been designated "annotation units" in TAIR to distinguish them from the original clone sequences in GenBank. These annotation units constitute the basis for the genome annotation coordination between TIGR and TAIR and for the assembled chromosome sequences used in the sequence map (see SeqViewer section).

Immediately after the genome assembly was complete, each genomic sequence location corresponding to an experimentally verified or predicted transcription unit was assigned a chromosome-based locus name (see Gene data section). As a consequence of the re-annotation effort, which has included merging and splitting of the original loci, some locus names have been made obsolete and new locus names have been created. The history of the locus names and the current usage can be searched and downloaded from TAIR (http://arabidopsis.org/tools/bulk/locushistory/).

TIGR's major re-annotation efforts have focused on incorporating the existing full-length cDNA sequences to verify gene structures and using protein sequence similarity information to annotate functional categories

(http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml). TAIR's major annotation efforts have been made at the functional level by using various sources of data such as the literature, sequence analysis, other database annotations, and community input (http://arabidopsis.org/info/tair-annotation.html). Both TAIR and TIGR are using controlled vocabularies for functional annotations developed by the Gene Ontology Consortium (http://geneontology.org), which can be searched and downloaded in bulk at TAIR (http://arabidopsis.org/tools/bulk/go/). The structural annotation information is searchable, browsable, and downloadable from GenBank (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search?chr=arabid.inf), TIGR (http://www.tigr.org/tdb/e2k1/ath1/), MIPS (http://mips.gsf.de/proj/thal/db/index.html), and TAIR (http://arabidopsis.org/info/agilinks.html).

Gene data

The gene information currently in TAIR amounts to 28,595 loci (including 28,038 physical loci and around 560 genetic loci) and 37,363 gene models (Table 2). A TAIR physical locus is defined as a region of the chromosome that corresponds to a single transcription unit in the *Arabidopsis* genome, and it is the minimum physical representation of a gene. A gene model is defined by a single CDS (coding sequence) or a transcribed sequence (for non-coding genes). Genes that correspond to the same genomic region but have different structural annotations are stored as separate gene models, but are all associated with the same physical locus.

The physical loci (genes resulting from the genome sequencing effort) are named according to their chromosome location. Chromosome-based names are a combination of organism name (AT), chromosomal location (1–5), G (for gene), and a unique accession number (e.g. AT2G34400). This nomenclature has been generally accepted as the standard nomenclature for *Arabidopsis* loci and replaces the previous one used during the sequencing phase, which was based on BAC names (e.g. F23H14.2).

The 560 genetic loci for which no sequence is yet available have been included in the Locus table temporarily until the sequence is resolved and can be associated to a physical locus. These genetic loci have been identified from mutant screens and can also be associated to one or many gene models that have been experimentally determined to be allelic. Genes identified from mutant phenotype characterization have symbolic names, the majority of them made of three-letters followed by a number (http://arabidopsis.org/info/guidelines.html).

The Gene data includes aliases, sequences, gene product descriptions, physical and/or genetic map locations, and associations to markers, attributions and publications. A significant amount of effort has been made in associating genes to the respective loci and sorting the gene names and aliases. There are now approximately 4,800 gene symbols in TAIR and about 66,000 aliases, including ORF names, full names, and symbols. In addition, about 10,000 *Arabidopsis* genes have an experimentally verified full-length cDNA sequence. About 900 genes have been identified by mutant phenotype and are included as "visible" markers in the Marker data class. Only a small portion of genes (about 3,500) have been studied or mentioned in publications.

The ongoing gene annotation effort carried out by TAIR, TIGR, and others will result in an expansion of the Gene data as the information is extracted from the literature and other sources. Currently 10,302 unique gene models have been functionally annotated with terms from the Gene Ontology Consortium's controlled vocabularies. The functional annotation data includes an updated set of literature-based annotations and electronic annotations based upon matches to INTERPRO domains. These data can be downloaded in bulk from the FTP site at: ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_Ontology/ (Table 3), and be searched by locus name at: http://arabidopsis.org/tools/bulk/go/. The structural annotation data can be downloaded from the FTP site at: ftp://ftp.arabidopsis.org/home/tair/Maps/seqviewer_data/.

Genes can be searched by name (symbolic, full name, ORF or gene product name), GenBank accession or description. The search can be restricted using different criteria, such as chromosome or map location, or whether the gene structure has been experimentally verified (http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene). The gene result page presents all the hits by locus name and gene model. The locus name is linked to a detail page where the associated gene models and EST sequences are displayed (Fig. 1a). In the future, all polymorphisms including insertions and deletions will be displayed on Locus detail pages. The gene model name is hyperlinked to a detail page (Fig. 1b) that displays all the associated information for that gene model, which includes description, aliases, sequences, map positions, markers, gene products, attributions, and publications related to that gene and links to other associated data, such as markers (Fig. 1c) and maps (Fig. 1d).

Protein and gene family data

TAIR has recently incorporated protein and gene family data to the resource. The protein data includes amino acid sequence, predicted physical-chemical properties such as isoelectric point and molecular mass, predicted subcellular localization (TargetP analysis; Emanuelsson et al. 2000), protein domains from INTERPRO (Apweiler et al. 2000) and superfamilies from the Structural Classification of Proteins (SCOP) database (Julian Gough, personal communication). Currently, all 26,082 predicted proteins from the genome sequencing project are stored in the database (Table 2). We plan to curate other protein information such as post-translational modification and protein-complex data from the literature and obtain pre-

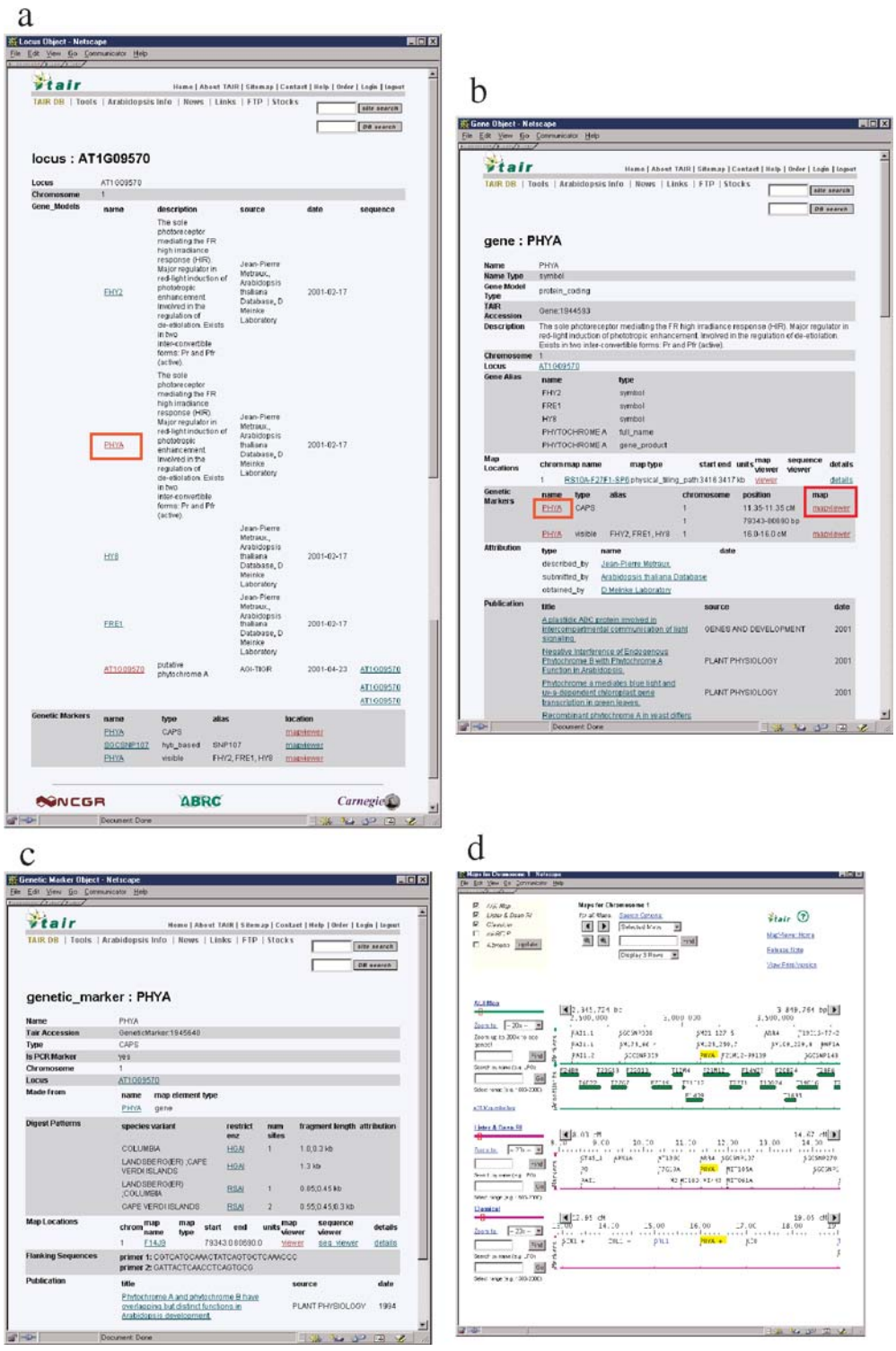

Fig. 1a–d Gene search interfaces.**a** Locus detail page displaying all the associated gene models and other related data resulting from searching for the gene PHYA. **b** Detail page of the gene model resulting from clicking on the gene model *enclosed in a square* on page in **a**. **c** Marker detail page linked from the gene model detail page from **b**. **d** Visualization of the marker PHYA on the MapViewer tool linked from the gene model detail page shown in **b**

dicted tryptic digest patterns using sequence analysis programs. The protein data can be searched at: (http://www.arabidopsis.org/tools/bulk/protein/index.html).

The gene family information has been collected from individual researchers and the literature, and for the moment is available only as a set of static html files at: http://www.arabidopsis.org/info/genefamily/genefamily.html. Currently, the dataset contains 3,630 genes classified into 450 gene families, and includes nucleotide or amino acid sequence information, as well as references and attributions to the source of the data. The gene families are generally defined according to overall protein similarity, but the specific software, parameters, and cri-

teria used to define each family vary for each gene family and the researcher who provided the data. We are currently developing the infrastructure and visualization tools to display these "expertly" curated gene family data along with computationally derived sequence similarity groupings from the database. For those interested in sharing their gene family data, TAIR provides a guideline for data submission that describes the information required and the appropriate format (http://www.arabidopsis.org/info/genefamily_submission.html).

Markers and polymorphisms

Genetic markers are defined in TAIR as polymorphisms associated with a method of detection. CAPS are Cleaved Amplified Polymorphic Sequences detected by PCR amplification, followed by digestion and gel electrophoresis of digestion products. SSLPs are Simple Sequence Length Polymorphisms, detected by PCR amplification and gel electrophoresis of PCR products. AFLP (Amplified Fragment Length Polymorphisms) are sequence polymorphisms, amplified by PCR and detected by gel electrophoresis. Other genetic marker types/methods, include Single Nucleotide Polymorphisms (SNPs) detected via hybridization (e.g. SNP chips), or allele specific PCR amplification (e.g. SNAP markers), Random Amplified Polymorphic Digested sequences (RAPDs), and Restriction Fragment Length Polymorphisms (RFLPs). Sources of genetic marker data include user submissions, the literature and data for the Recombinant inbred map from the Nottingham Arabidopsis Resource Center (NASC; http://nasc.nott.ac.uk/). When available, the associated primer sequences and special conditions for detection are shown on the marker detail pages. Genetic markers can be located on genetic maps (e.g. miRFLP, Lister and Dean RI map, Classical Genetic Map) and can be visualized with the MapViewer (see below) and if the sequence is known, they are placed on a sequence map, and visualized with the MapViewer or the Sequence Viewer (see Tools section). Links to relevant literature include papers that describe marker detection methods and map locations. An example of a detail page of a marker is shown in Fig. 1c. The number of each marker type available at TAIR is shown in Table 2.

Polymorphisms in TAIR can be sequence-based (e.g. SNPS), digest patterns, PCR product lengths, INDELS (insertion/deletions), insertions (e.g. T-DNA, or transgene insertions), and substitutions (Table 2). Polymorphisms that represent altered forms of a gene (i.e. alleles) may be included as visible polymorphisms, and associated with the sequence polymorphism that is the basis of the allele. Polymorphisms can be associated to genes, loci, genetic markers, and germplasm. When available, polymorphisms are associated to relevant literature.

Aside from searching by allele name and polymorphism name, users are able to retrieve data based on genetic marker type, polymorphism type, allele type, allele inheritance, transgene construct type, polymorphism site, insertion type, and mutagen (http://arabidopsis.org/servlets/Search?action=new_search&type=marker).

Maps

Currently the most comprehensive map in TAIR is the sequence-based one from the AGI, which contains about 31,000 genes, 2,450 genetic markers, 108,000 ESTs, and 1,600 annotation units mapped on them on the basis of sequence matching. In addition to the sequence-based map, there are three genetic maps and over ten physical maps. Available genetic maps include the classical genetic map (based on scoring of visible phenotypes) maintained by David Meinke (Meinke et al. 1998; http://mutant.lse.okstate.edu/), the Lister and Dean RI map (Lister and Dean 1993) maintained by NASC (http://nasc.nott.ac.uk/new_ri_map.html) and the mi-RFLP genetic map (Liu et al. 1996). Physical maps were developed by a number of *Arabidopsis* researchers using clone hybridization and fingerprint experiments, and were collected at AtDB (Rhee et al. 1999) and references therein (http://arabidopsis.org/mapViewer/help/mapkey3.htm#913144). These maps are displayed using two main map tools, MapViewer and SeqViewer, described in detail in the Tools section. The SeqViewer displays all the entities on the sequence-based AGI map. The map comparison tool, MapViewer, can display and align the AGI, genetic, and physical maps at once.

Clones

Over 175,000 different clones are included in TAIR database, including BAC, YAC, TAC, P1, cosmid, plasmid and lambda clones (Table 2). The TAIR clone search page allows clones to be searched based on name, GenBank accession, vector type, insert type, availability as a stock, genomic location of the cloned DNA, and other features such as existence of sequence information or whether the clone has been used as a marker. Information that can be accessed from the clone detail page includes aliases, insert type, GenBank accession if available, any EST or BAC end sequences, library and vector name, cloning enzyme, map locations, name of the person or organization that registered, submitted or sequenced the clone, and the stock name. Clones that are Arabidopsis Biological Resource Center (ABRC) stocks can be ordered directly from the search summary page or the clone detail page.

Detail pages of clones that were sequenced as part of the genome initiative also carry a link to the annotation unit derived from that clone. Annotation units are based on the original clone sequences but frequently include sequence extensions derived from neighboring clones or PCR products and were created by MIPS to facilitate annotation of genes located at the clone ends (Heiko Schoof, personal communication). Since annotation units

are the basis of the genome annotation and the TAIR AGI map, information about which genes exist on a sequence are viewable from the annotation unit detail page rather than the clone page. Other information available from the annotation unit page includes adjacent annotation units and amount of overlap, sequence status, length and a list of ESTs matching the annotation unit sequence.

Sequences

TAIR database includes gene, protein and genomic clone sequences from the Arabidopsis Genome Initiative as well as mRNA, BAC end and EST sequences from GenBank. Sequences can be retrieved through their association to genes (ESTs and mRNAs) or clones (clone and BAC end sequences). In addition, a simple sequence retrieval interface is available which can be used to download sequences by locus name (http://www.arabidopsis.org/tools/bulk/sequences/index.html). Sequences can be downloaded in FASTA or tab-delimited formats. Also, these and other sequence files in FASTA format are updated regularly for the FTP site (Table 3) and TAIR sequence analysis programs online (see Tools section). In the future, a more advanced database search for sequences is planned, which will permit searching by name, sequenced object type, location on the genome, and other parameters that will allow users to build customized sequence sets for further analysis.

DNA and seed stocks

TAIR has recently incorporated the ABRC database functions and ordering system (http://arabidopsis.org/abrc/), allowing users to obtain information about the DNA and seed stocks available at the stock center, to place orders on line and to view the order history of ABRC users.

The DNA resources available from ABRC include BAC clones, ESTs, libraries, full-length cDNAs, YACs, cosmids, RFLP phage lines, and pools of genomic DNA from T-DNA mutant population (Table 2). The cDNA libraries have been isolated from different sources including whole plants, seedlings, and flowers. Some libraries have been size-selected to facilitate isolation of larger clones. Genomic libraries include phage, cosmid, yeast artificial chromosomes, bacterial artificial chromosomes (BACs), and plant-transformable BACs (transformable BACs and binary BACs). Available BACs include those utilized by the Arabidopsis Genome Initiative public genome sequencing projects. The available BACs and ESTs, in conjunction with the mapping resources of the stock centers, provide a powerful set of resources for positional gene cloning, complementation, expression, and protein characterization (Scholl et al. 2000).

Pooled genomic DNA obtained from pooled transgene insertion lines is also available for PCR screening to identify insertions in a gene of interest, complementing the "knockout" services of the Arabidopsis Knockout Facility (AKF; http://www.biotech.wisc.edu/Arabidopsis/). DNA isolated from 12,000 T-DNA lines is currently available from ABRC and will be expanded to a total of 40,000 lines.

ABRC maintains and distributes seeds from public and private sources. Mapping resources include mapped mutant lines, multiple marker lines, four populations of recombinant inbred lines, trisomic lines, and a population organized by tetrads (Table 2). Transposon and T-DNA populations with random insertions throughout the genome have been produced which enable a line to be identified by phenotype and utilized to identify and characterize the associated insertion point. The current holdings of T-DNA stocks exceed 140,000. Currently there are 50,256 insertion flank sequences available, which are included as a dataset for BLAST and the other sequence analysis programs offered at TAIR (see Tools section). Large numbers of characterized lines, having transpositions to random locations, are also available, together with lines transformed with specific transgenes and molecular tags, transposon parental stocks and natural variants collected in the wild from around the world (Table 2).

The current Seed stock search allows the user to specify gene or allele name, background ecotype, donor, phenotype, and stock number to narrow down their choices http://arabidopsis.org/servlets/SeedSearcher). More sophisticated search capabilities are in development that will allow searching based on mutagen, insertion type, transgene construct type, allele type, polymorphism inheritance, and chromosomal constitution in addition to the existing criteria.

Metabolic pathways

Biochemical pathways constitute complex data that include information on molecules, reactions, co-factors, enzymes, locations, genes, etc. TAIR has relied on the Pathway Tools software (Karp et al. 2002) to build an *Arabidopsis*-specific pathway database called AraCyc (see Tools section). It currently features over 1,800 enzymes and 1,001 enzymatic reactions in 167 pathways containing 719 different compounds. The AraCyc pathways have been generated computationally using Pathway Tools software (Karp et al. 2002) with TIGR's gene annotations and the MetaCyc database as a reference database (Karp et al. 2002). The resulting database was then edited manually to correct mistakes, remove non-plant pathways and add missing plant pathways. To date, more than a dozen plant-specific pathways, including carotenoid, brassinosteroid, and gibberellin biosyntheses have been added from the literature. A list of more than 40 plant pathways will be added in the coming months. The aim is to reflect current knowledge of *Arabidopsis* metabolism as precisely and extensively as possible, and to allow users to make educated guesses

about parts of the metabolism where little experimental evidence is available.

References

TAIR stores four types of references in the database: publications, database references, analysis references, and personal communications. Publications include peer-reviewed research articles, reviews, theses, books, abstracts from conferences, and book chapters. Publications are downloaded monthly from PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), and quarterly from Agricola (http://www.nal.usda.gov/ag98/) and Biosis (http://www.biosis.org/) databases, (kindly provided by Barbara Buchanan at USDA). In addition, abstracts of the past international conferences on *Arabidopsis* research are included in the collection. Currently the number of publications in TAIR amounts to more than 14,000 (Table 2). Publications can be searched by author, keyword, source, and year (http://arabidopsis.org/servlets/Search?action=new_search&type=publication), and the result page displays the full citation, including the abstract and hyperlinks to full-text and to community detail pages, if available.

Information about the software and parameters used on computational analysis of data performed at TAIR, such as restriction analysis, domain matches, or BLAST analysis, is stored in the Analysis reference class. The Database reference class stores information on data obtained from external databases, such as TIGR or GenBank. The Personal communications table stores the references submitted to TAIR by individual users, which are not yet published or will not be published in a journal.

Community

Arabidopsis researchers and organizations that use *Arabidopsis* in their research comprise most of TAIR's community. The purpose of storing community data is multiple: it facilitates the exchange of information among members of the community, it allows us to make attributions to the data in TAIR, and it is necessary to maintain the stock ordering system and order history recording of ABRC. Currently we have contact and research interest information on more than 10,500 persons and about 4,000 organizations (Table 2). Organizations include labs, companies, research institutes, consortiums, and committees. The Community information can be searched by name, e-mail, city, country and keywords, and can be restricted to only persons or organizations (http://arabidopsis.org/servlets/Search?action=new_search&type=community). New users can register in TAIR through the Web (http://www.arabidopsis.org/servlets/Community?action=edit&new=true&type=person). Once they have obtained a password, users are able to use web forms to update their own information. In the future, we plan to add a function that would allow users to update their bibliography record from references stored in the data base.

Microarray expression

TAIR is in the process of incorporating microarray expression data from users and external databases into the resource. In the future, expression information in TAIR will expand to include in situ hybridization, northern blots, RT-PCR, reporter gene fusions, and other expression data available from the literature.

Besides the "numeric" data, the Expression data class includes detailed information about the experimental design and the source of the RNA used for hybridization, as well as the data analysis and protocols employed.

The first set of data that are being incorporated represents the bulk of public experiments performed by the Arabidopsis Functional Genomics Consortium (AFGC; http://afgc.stanford.edu/afgc_html/site2.htm), which currently consists of about 400 hybridizations. In addition, data files that contain information on the probes used in AFGC and Affymetrix arrays, including annotation and their corresponding locus assignment generated by TAIR, are currently available for download at: ftp://ftp.arabidopsis.org/home/tair/Microarrays/ (Table 3).

Functional genomic projects

As a follow-up to the *Arabidopsis* genome sequencing efforts by the AGI, a new initiative known as the "The 2010 Project", was proposed to the *Arabidopsis* community with the goal of understanding the function of all plant genes by the year 2010 (Chory et al. 2000). As a result of this initiative, functional genomic projects aimed to identify the molecular function of all *Arabidopsis* genes have been initiated all over the world. The role of TAIR is to provide the bioinformatics infrastructure to facilitate the dissemination of the information generated from these projects. Information about the functional genomic projects currently funded can be found at: http://www.arabidopsis.org/info/2010_projects/. The information provided here allows any user to search for the genes included in each research proposal, as well as learn about the individual projects and researchers. This information is also available in the database from the Gene and community pages.

Cereon polymorphism collection and Landsberg *erecta* sequences

TAIR provides access to the collection of 56,670 predicted *Arabidopsis* single-nucleotide polymorphisms (SNP) and small insertions/deletions (INDELs) between the publicly available Columbia (Col) sequence and Landsberg *erecta* (L*er*) sequence generated by Cereon Geno-

mics (http://arabidopsis.org/Cereon/index.html). In addition to the polymorphisms, Cereon has recently released approximately 95 Mb of sequence from the L*er* strain. The L*er* sequence is comprised of 81,306 sequence entries from a single-pass shotgun sequencing.

These data are accessible at TAIR only to non-profit institutions, universities, and colleges. In order to access the data, users must first fill out a one-time registration form for either Polymorphism or the L*er* sequence data to get access to each data set. Registered users must then login to access the data, which can be viewed as HTML pages or downloaded in their entirety as tab-delimited text files.

FTP datasets

Aside from the data stored in the database, TAIR provides a variety of data files in tab-delimited format on the FTP site (ftp://ftp.arabidopsis.org/home/tair/). The contents of the FTP site are summarized in Table 3. The collection includes all the sequence datasets used by the sequence analysis programs (see below), the complete *Arabidopsis* genomes (nuclear, chloroplast and mitochondrial), mapping data, a variety of gene and protein data files, and microarray and ontology information.

Tools

TAIR provides a series of sequence analysis, visualization and data retrieval tools. They are all accessible from the Tools section of TAIR's web page at: (http://arabidopsis.org). In addition, a tool that allows placing stock orders to the ABRC stock center directly from TAIR's web pages is now available. A brief description of each tool is presented below.

Visualization tools

SeqViewer

TAIR's SeqViewer is a tool for viewing the *Arabidopsis* genome and its associated annotation starting with a whole genome view and zooming down to the nucleotide level (http://arabidopsis.org/servlets/sv). The SeqViewer displays gene annotation, clones, markers, transcripts, and polymorphisms on the *Arabidopsis* genome sequence. Users can search with clone, marker or gene names or a short nucleotide sequence and visualize the locations of one or many search hits on the whole genome, in a close-up view or in a nucleotide window.

The SeqViewer entry page, shown in Fig. 2a, displays a whole genome view of the five *Arabidopsis* chromosomes and provides the following functions: (1) Search by name to find the location of a gene, marker or clone; (2) search by sequence to find the location of up to four short sequences (15–150 nt each): and (3) open a close-

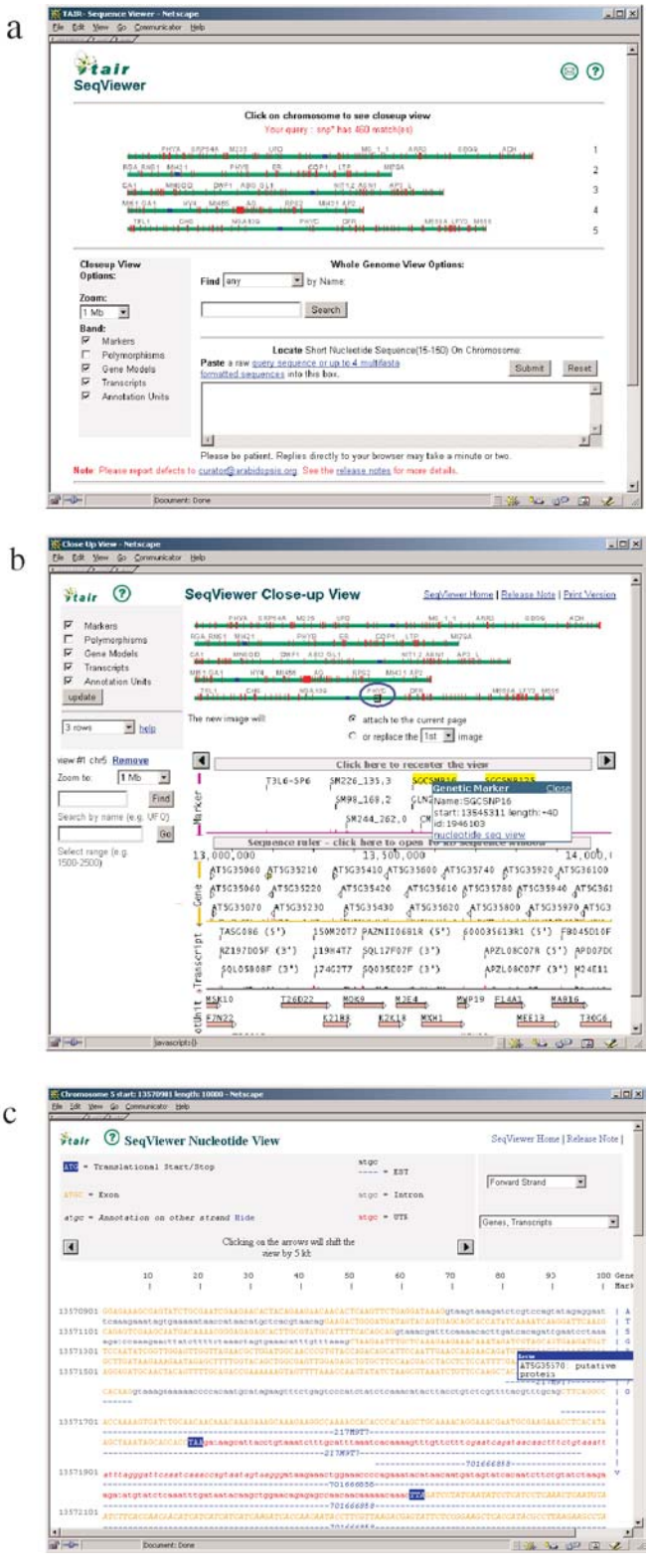

Fig. 2a–c Sequence Viewer tool. **a** Display of the result of searching for "SNP*" as *red tick marks* on the chromosome bars. **b** Close-up window. This page is the result of clicking on one of the tick marks shown in the previous window (marked as *1*). It displays the markers, gene models, transcripts and annotation units that lie on the selected position on the chromosome as separate transversal bands. **c** Nucleotide window showing a region that was selected from the close-up window in **b** by clicking on one of the objects on the Gene band

up view by clicking on a point on one of the five chromosomes.

The Close-up view page displays a whole genome view of the five *Arabidopsis* chromosomes at the top and one or more close-up views below containing marker, gene model, transcript and clone map bands (Fig. 2b). Additional information for each mapped object is available as a pop-up window that appears on mouseover. To signal when the zoom level does not allow all objects in a band to be displayed we have added tick marks below each map band representing all the objects in the band. The tick marks are color-coded: black marks represent objects that don't appear at the current zoom level because of space constraints and red marks represent objects currently being displayed. Functions available from the close-up page include: scroll, change zoom level, enter a range or re-center to customize the view; search for a marker, gene or clone name; click on names to link to a detailed page describing the object; open a second close-up view; and open a nucleotide window to view 10 kb of nucleotide sequence with genes and other features highlighted.

The Nucleotide window (Fig. 2c) can be scrolled 5 kb at a time, and allows users to view genes on one or both strands with UTRs, start and stop codons, exons and introns highlighted in different colors. EST matches, markers and clone ends can also be precisely located in the nucleotide window. Text from this page can be copied and pasted into other applications with the gene annotation preserved as uppercase (coding regions) and lowercase (introns, UTRs and intergenic regions).

MapViewer

The MapViewer is a tool that integrates the visualization and analysis of different maps, such as genetic, physical, and sequence-based, on each chromosome (http://arabidopsis.org/servlets/mapper). It allows searching, browsing, and aligning of maps and map elements on each chromosome and displays the information graphically. A screen shot is shown in Fig. 1d. This tool was developed to facilitate forward and reverse genetics, where researchers can start with a mutant phenotype and get to the gene of interest, or start with a gene or gene family of interest and find out what the roles of these genes might be by looking for mutations in the gene. Each entity on the maps is hyperlinked to a page with detailed information from the database. In addition, there is extensive help documentation on how to use the map viewer, from interpretation of the data to navigation of the tool. The help document is hyperlinked from every page on the MapViewer (http://arabidopsis.org/mapViewer/help/tairmapa.htm).

AraCyc

AraCyc is a tool for querying and visualizing *Arabidopsis* biochemical pathways (http://arabidopsis.org/tools/aracyc). Some of the interfaces are shown in Fig. 3. The main query page (arabidopsis.org:1555/server.html) provides links to descriptions of the dataset, an overview diagram/expression viewer, and the results of the initial computational build of the database. The tool allows users to perform queries with text strings to find pathways, reactions, proteins, genes or molecules, as well as to browse a classification hierarchy of pathways, reactions, compounds or genes. The pathways are displayed graphically and the tool allows zooming in using the "More detail" button to reveal more details about the reactions down to the molecular structures of all compounds and co-factors. All the reactions, compounds, genes, proteins and cofactors displayed in the pathways are clickable and display detail pages on the clicked object. An overview diagram gives a bird's-eye view of all pathways in the database. In the diagram, the pathways are represented schematically with compounds being little empty squares (sugars), triangles (amino acids), etc., where filled symbols denote phosphorylated compounds and lines represent reactions. When the cursor is moved on the expression viewer, information on the pathway and compounds under the cursor are given in the browser status line. Reactions are represented as lines between the compounds. The expression viewer allows expression data to be overlaid on the overview diagram by color-coding the lines representing the reactions according to the expression values.

Sequence analysis tools and datasets

BLAST and FASTA

TAIR provides web access to NCBI BLAST2.0 (http://arabidopsis.org/Blast/), WU-BLAST2.0 (http://arabidopsis.org/wublast/index2.html) and FASTA (http://arabidopsis.org/cgi-bin/fasta/nph-TAIRfasta.pl) sequence similarity searches. Users can choose to search against several nucleotide and protein sequence datasets (Table 3). Nucleotide sequences include CDS and protein sequences from the AGI, BAC sequences from TIGR, BAC ends from Kasuza and subsets of sequences from GenBank (i.e. all *Arabidopsis*, only *Arabidopsis* BACs, only *Arabidopsis* ESTs, all *Arabidopsis* minus ESTs and BACs, and all Viridiplantae sequences), and T-DNA/transposon insertion flank sequences. Additional genomic sequences include 1,000- and 3,000-base-pair upstream or downstream sequences flanking loci as well as intergenic and intron sequences generated by TAIR. In addition to nucleotide sequences, a non-redundant *Arabidopsis* protein sequence dataset created by merging all entries with identical sequences to form a unique set from GenPept, SWISS-PROT, and PIR is also searchable.

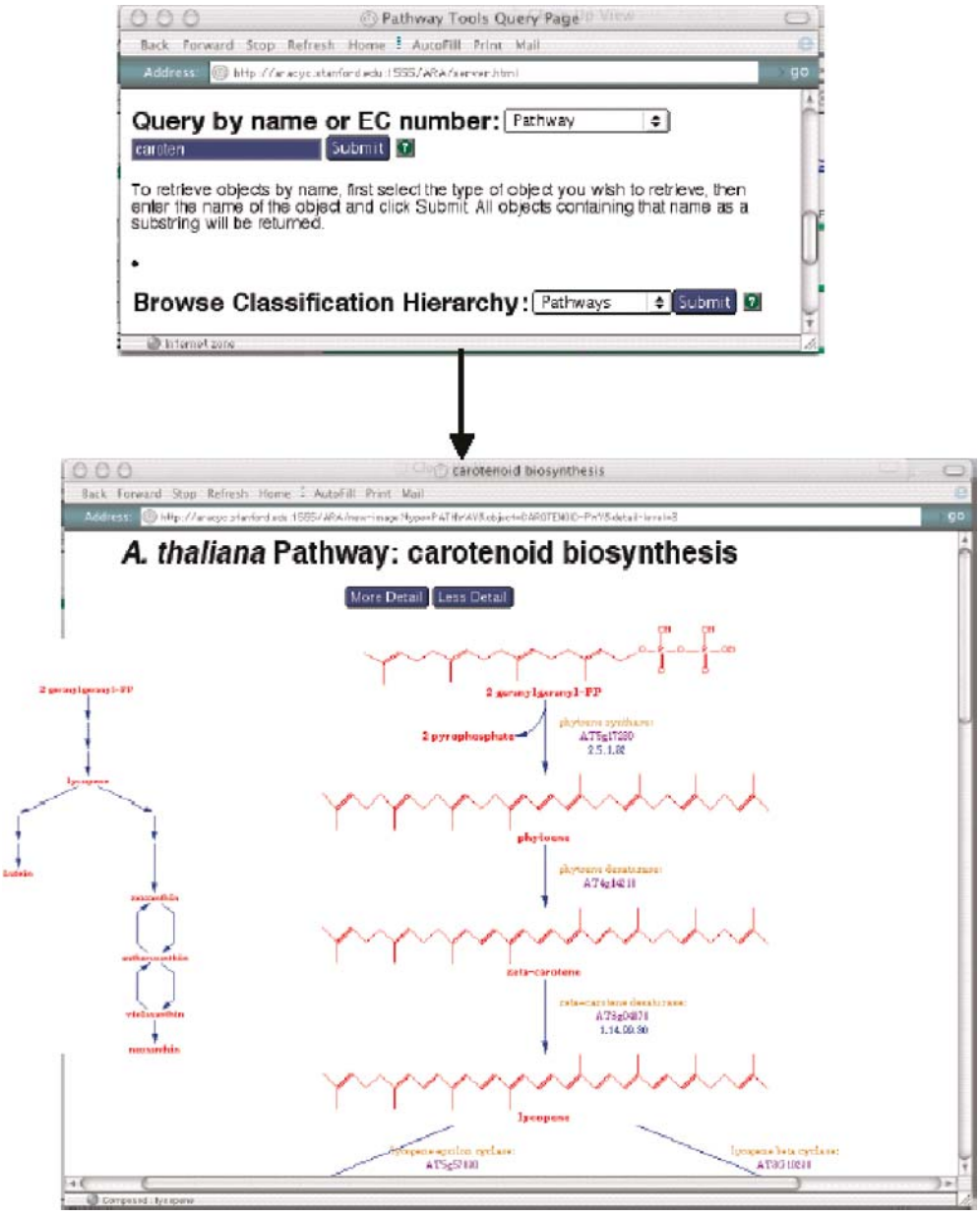

Fig. 3 AraCyc, TAIR's biochemical pathways querying and visualization tool. The query page is shown *at the top* of the figure and a sample query response shown *in the lower part*. The screen shot shows a "zoomed-in" version of the pathway that was initially found by the search (*box*)

The BLAST result page displays a one-line description of all matching sequences, followed by the actual alignment of the query sequence with the database sequences, and the last section lists the parameters used and the statistics generated during the search. The one-line descriptions give information about the dataset sequences that form a high scoring segment pair (HSP) with the query sequence. Each line is hyperlinked to the TAIR locus detail page for the matching sequence, as well as to MIPS and TIGR databases. BLAST datasets are available from FTP (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/).

PatMatch

PatMatch is a program for identifying short (less than 30) nucleotide and amino acid motifs (http://arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl). It is a tool specially suitable for searching short stretches of degenerate or conserved motifs such as *cis* regulatory elements and protein modification sites that are not easily searched by local alignment programs like BLAST and FASTA. It is based on the use of simple terms or regular expressions, which allow for up to three mismatches, insertions or deletions, as well as perfect matches. It searches the same datasets used by BLAST and FASTA. For DNA searches, either strand or both strands together can be searched. An example of a regular expression

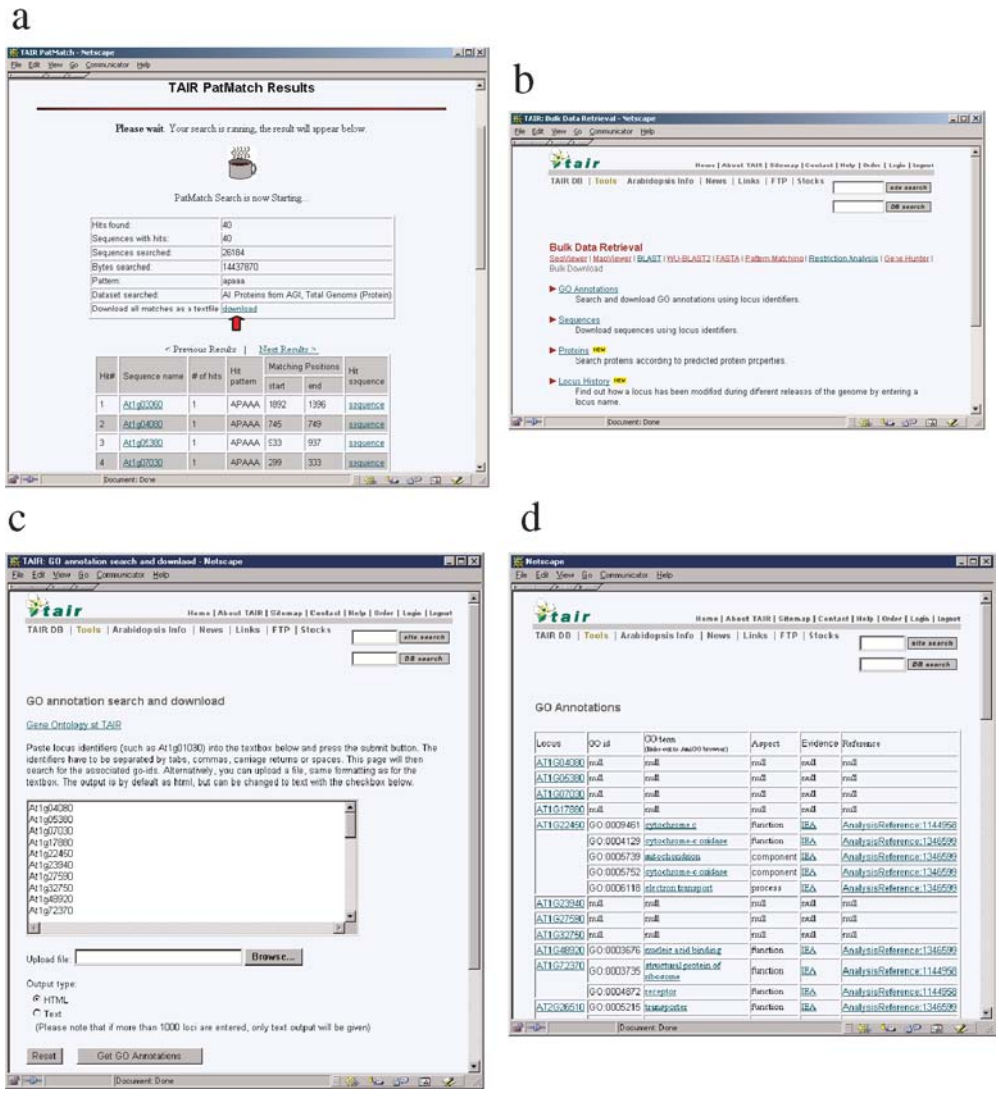

Fig. 4a–d PatMatch and Bulk Download tools. **a** The result of querying PatMatch for the pattern APAAA using the AGI total genome dataset. **b** The Bulk Data Retrieval tool entry page, displaying the several data retrieval options available (GO, sequence and protein data). **c** The page for retrieval of GO Annotation. In this example, the loci chosen were taken from the result obtained using PatMatch in **a**, using the download function marked with an *arrow*

search is <[AVLI]{3}X{4}KP{2,5}. This will find proteins that have, at the N terminus (<), three aliphatic amino acids ([AVLI]{3}) followed by any four (X{4}) amino acids, followed by lysine (K), followed by two to five proline residues (P{2,5}). An example of a PatMatch result page is shown in Fig. 4a. The results can be downloaded and used to retrieve other associated data, such as sequences or GO annotations using the bulk data retrieval tool described below.

Data retrieval tools

Text-based query tools

We provide several types of querying methods, including general and specific text searches, as well as the graphical searches already described.

The general text-based query tool allows users to type in any word and the output returns all the different hits found for each object type (e.g. markers, genes, community). This search can be accessed from a box identified as “DB search” located at the top of every web page, and from the entry page to the database at: http://arabidopsis.org/servlets/Search?type=general&action=new_search. Currently this functionality only works with object names, but will soon be extended to keywords.

More complex searches are facilitated by web forms that allow users to choose the type of object to be searched (e.g. genes, clones), and then specify multiple criteria for locating their data of interest. Currently, the types of objects that are searchable are genes, markers, proteins, clones, sequences, stocks, maps, community and publications (http://arabidopsis.org/servlets/ Search?type=general&action=new_search; http://arabidopsis.org/abrc/).

Search results are summarized on an intermediate page, where all the matches are displayed as a one-line description. Users can access a detail page for each entry by clicking on the name of the objects. The detail page

displays all the data associated to the chosen object, including the metadata. A long-term goal is to have more customizable result pages to allow the browsing and downloading of data of interest.

Bulk data retrieval tool

To facilitate the retrieval of large numbers of data including sequences, annotations of genes and proteins, TAIR has developed a bulk data retrieval tool (http://arabidopsis.org/tools/bulk/), shown in Fig. 4b, c, and d. Users can either upload a file with locus names, or paste the list directly on the box. For the retrieval of sequences, users can customize the type of sequence to be retrieved (coding, protein, transcript, or promoter), and the format (FASTA or tab-delimited text).

Arabidopsis *GeneHunter*

The *Arabidopsis* GeneHunter searches information about *Arabidopsis* genes from a selectable list of *Arabidopsis*-related web sites and databases, including TAIR, TIGR, AGR, Meinke DB, GenBank, PubMed and Swiss-Prot (http://arabidopsis.org/cgi-bin/geneform/geneform.pl). The result is a concatenation of clickable web pages from the sites searched. It is a useful tool to retrieve information about a given gene from different sources.

Stock ordering

The ABRC stock order processing has now been incorporated into TAIR. In collaboration with ABRC, we have developed tools for stock searching and ordering to be conducted from TAIR. The current ordering system allows the DNA sequence, genetic and physical maps, and comprehensive genomics data and capabilities of TAIR to be utilized to efficiently locate clones for ordering. Searchable seed stock data are available, but these data are not linked to the genetic and sequence sections of TAIR yet. The present ordering process in the TAIR system involves user login, searching for stocks and placement of orders through the TAIR web site. In addition, the clone searches are linked to the ordering pages, and in the future the seed search will also be linked to the ordering process. Also, a recent tool allows access to viewing order history information by user or stock number (http://arabidopsis.org/servlets/Order?state=search&mode=stock_number).

Database and software design and implementation

TAIR DB uses an object-oriented approach to data representation, implemented in a relational database (Sybase). The data is organized according to a hierarchical structure. Each parent table links to tables with generic information for all the members in the same branch, and all the children inherit the data fields of the parent tables. At the top of the hierarchy is the Tair Object class. Each TAIR object is in turn linked to a set of data that provide information about its source and history (Attribution and Reference classes), and its description (Annotation class). The Attribution, Reference and Annotation classes constitute the metadata of TAIR objects. This design has the advantage of allowing easy expansion of new data types as needed. More detailed information about the database schemes and documentation can be found on http://www.arabidopsis.org/search/schemas.html.

Controlled vocabulary design and implementation

The development and implementation of controlled vocabularies play a pivotal role in advancing genomic research. There is a movement to standardize controlled vocabularies for function, process, and sub-cellular location of gene products among several organism databases. This initiative has been developed by members of the Gene Ontology Consortium (http://www.geneontology.org/), which includes representatives of several model organism genome databases. The information stored by the GO project includes ontologies, definitions of terms and gene associations. Each ontology is composed of well-defined terms (based on published data), and their relationships with other terms. A given term can have multiple parents and children, their association defining a non-cyclic structure, and can be linked by different types of relationships, e.g. “an instance of”, or “a part of”.

TAIR joined the consortium in June 2000 and is contributing to the development of controlled vocabularies for describing the function, process, and localization of gene products that apply to higher plants. We have also developed controlled vocabularies to describe *Arabidopsis* anatomy and developmental stages (ftp://ftp.arabidopsis.org/home/tair/Ontologies/; Table 3) and are working to extend these ontologies to describe the anatomy and development of other plant species in conjunction with several plant genome databases: Gramene (http://www.gramene.org/), MaizeDB (http://www.agron.missouri.edu/), International Rice Research Institute (http://www.irri.org/), under the auspices of the Plant Ontology Consortium.

GO annotation of *Arabidopsis* genes and the anatomy and developmental ontologies can be found at the FTP site (Table 3).

Future prospects

Because the value of *Arabidopsis* derives from its utility in understanding other plants, our goal is to build a structure for TAIR that permits facile high resolution linking of specific information about *Arabidopsis* to similar information in all other plants (and vice versa).

Ultimately, TAIR's goal is to provide the common vocabulary, visualization tools, and information retrieval mechanisms that permit integration of all knowledge about an organism into a seamless whole that can be queried from any perspective. Of equal importance for plant biologists, the ideal TAIR will permit a user to use information about one organism to develop hypotheses about less well-studied organisms. Thus, TAIR's goal is to develop user-friendly tools that permit an individual working outside this model species to formulate a query based on their organism of interest, have that query directed to the relevant knowledge for the plant models, and present the information about the models in a way that can be understood by the plant biology community at large.

Acknowledgements We would like to thank the TAIR users and the *Arabidopsis* research community for their continuing support, feedback, and particularly for sharing their data and expertise. This project was supported in part by the National Science Foundation (grant numbers DBI-9978564, DBI-0091471, and DBI-9813586) and by the National Institute of Health (grant number HG-02273). This is the Carnegie Institution of Washington Department of Plant Biology Publication 1546.

References

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM (2000) InterPro–an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16(12):1145–1150

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM, Cook D, Dangl J, Grant S, Guerinot ML, Henikoff S, Martienssen R, Okada K, Raikhel NV, Somerville CR, Weigel D (2000) A blueprint for understanding how plants are built and how to improve them. National Science Foundation-sponsored workshop report: the 2010 project functional genomics and the virtual plant. Plant Physiol 123:423–426

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300(4):1005–1016

Flanders DJ, Weng S, Petel FX, Cherry JM (1998) AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the Arabidopsis Genome Initiative. Nucleic Acids Res 1(26):80–84

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based formation retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29(1):102–105

Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. Nucleic Acids Res 30(1):59–61

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4:745–750

Liu YG, Mitsukawa N, Lister C, Dean C, Whittier RF (1996) Isolation and mapping of a new set of 129 RFLP markers in *Arabidopsis thaliana* using recombinant inbred lines. Plant J 10:733–736

Meinke DW, Cherry JM, Dean M, Rounsley SD, Koornneef M (1998) *Arabidopsis thaliana*: a model plant for genome analysis. Science 3(282):662–682

Rhee SY (2000) Bioinformatic resources, challenges, and opportunities using *Arabidopsis* as a model organism in a post-genomic era. Plant Physiol 124:1460–1464

Rhee SY, Weng S, Bongard-Pierce DK, Garcia-Hernandez M, Malekian A, Flanders DJ, Cherry JM (1999) Unified display of *Arabidopsis thaliana* physical maps from AtDB, the *A. thaliana* database. Nucleic Acids Res 27(1):79–84

Scholl RL, May ST, Ware DH (2000) Seed and molecular resources for *Arabidopsis*. Plant Physiol 124:1477–1480