AtDB, the *Arabidopsis thaliana* Database, and graphical-web-display of progress by the *Arabidopsis* Genome Initiative

David J. Flanders, Shuai Weng, Fabien X. Petel+ and J. Michael Cherry*

Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA

Received September 15, 1997; Revised and Accepted October 15, 1997

ABSTRACT

AtDB, the *Arabidopsis thaliana* Database, has a primary role to provide public access to the collected genomic information for *A.thaliana* via the World Wide Web (URL: http://genome-www.stanford.edu/). AtDB presents interactive physical and genetics maps that are hyperlinked with detailed information about the clones and markers placed on these maps. A large literature collection on *Arabidopsis*, contact information on researchers worldwide, laboratory method manuals and other information useful to plant molecular biologists are also provided. This paper discusses the database-driven clickable displays that provide easy navigation within a variety of genomic maps, including those summarizing progress of the international *Arabidopsis* genomic sequencing effort, AGI (the *Arabidopsis* Genome Initiative). The interface uses client-side hyperlinked GIF-images that direct the user to detailed database-information. A new BLAST service is also described. This gives users access to the thousands of *Arabidopsis* BAC clone end-sequences and includes hyperlinked images summarizing the search results. The linking of genetic and physically mapped regions and their sequence into information for loci within that region is an ongoing goal for this project.

INTRODUCTION

The small flowering plant, *Arabidopsis thaliana*, a member of the mustard family, has become the model system for plant biology. Its unique features allow it to be applied to a wide variety of research problems, including plant physiology, the molecular genetic study of organ development and plant–pathogen interactions. *Arabidopsis* has established itself in this role for several reasons: its small size; short life cycle; small genome size; low amount of repetitive DNA sequences; and the relative ease with which it is manipulated by molecular and genetic methods. The importance of *Arabidopsis* to plant biology is increasing with the international development of germplasm and DNA stock resource centers, genetic and physical maps and their associated techniques, expression sequence tags (EST, single pass cDNA sequences), and, of particular relevance to this paper, genomic sequencing projects.

Arabidopsis provides a model for higher plants much in the way that *Saccharomyces* has provided a model for cell biology in general, the value of the latter is enhanced as a result of having the complete genomic sequence. Molecular genetic techniques for *Saccharomyces* utilize this and allow explicit targetting of genes being studied. The genomic sequence has thus fostered the creation of many systematic projects to study gene expression and function, protein–protein interactions, protein localization and molecular evolutionary changes of genomes. These projects rely strongly upon bioinformatics, including up-to-date databases. The *Arabidopsis* community intends to learn from the efforts on *Saccharomyces* and *Caenorhabditis*, which is nearing completion, about the computational and database requirements for the public use of the large volume of information produced by systematic sequencing projects. In this regard, the AtDB project is well placed being in the same group as the *Saccharomyces* Genome Database (1).

The *Arabidopsis* genomic sequence will be important for all of plant biology. It is anticipated that molecular genetic tools for *Arabidopsis* will continue to allow targeted mutagenesis techniques to be discovered. Furthermore, because flowering plants evolved relatively recently, characterizing a gene in *Arabidopsis* often simplifies the isolation of a corresponding gene in another flowering plant. In addition, it is likely that an *Arabidopsis* gene will be functionally homologous within many other plant species, and vice versa (for a review of the biology and advantages of *Arabidopsis* see ref. 2).

In 1996, a major internationally-orchestrated effort thus began with the goal to sequence the entire *A.thaliana* genome by the year 2004. Termed the *Arabidopsis* Genome Initiative (AGI), it comprises six main groups (Table 2: The Institute of Genomic Research (TIGR); the SPP Consortium (Stanford University, Plant Gene Expression Center, and University of Pennsylvania); the CSHL Consortium (Cold Spring Harbor Laboratory, Washington University at St Louis, and Applied Biosystems, Inc.); the European Union's European Scientist Sequencing *Arabidopsis* (ESSA) network of laboratories; The Centre National de Sequençage (CNS), France; and the Kazusa DNA Research Institute, Japan. The

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Fax: +1 650 723 7016; Email: cherry@genome.stanford.edu

+Present address: Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990, USA completely sequenced and annotated BAC and P1 clones began to flow into the international DNA sequence databases (GenBank/ EMBL/DDBJ) in the fall of 1996. At the time of writing, we estimate that ~10% of the genome has been completed.

Most of the AGI participants have web sites giving full details of that group's efforts. AtDB's role is not to duplicate these sites. Rather its mission within the AGI project is to maintain a searchable, integrated summary of all AGI efforts including the name, genomic location and progress of all clones being sequenced. AtDB thus provides a single site for users to get answers to such questions as; 'who's doing what, where; and how are they progressing?' AtDB provides links to the individual AGI sites from its 'Information about AGI' page (Table 1) for users to glean more detailed information. There are two related aspects to the AGI data within AtDB: data acquisition and its display.

AGI DATA ACQUISITION: ON LINE TRANSACTION PROCESSING

Members of AGI use Web forms produced from AtDB's Illustra database to register clones that they propose to sequence and their subsequent progress in sequencing these clones. (Illustra was purchased by Informix Software, Inc. in 1995. We are currently using the Illustra product as released by Informix, and not the new Informix product termed Informix Universal Server.) These forms are password protected to help prevent spurious entry of data. Note that the data submitted are freely and instantly available from the database. There are no 'hidden' data on AtDB. All information submitted to AtDB is made available to the users as soon as possible.

AGI submitters then supply details of their name (their group is known through the password entry-process), the clone, and its sequencing status—the progression of their work is recorded for sequencing of the clone from intention to sequence, through clone library preparation, sequencing in progress, then completed sequence and, finally, to fully-annotated released-sequence. If known or appropriate, also supplied are the clone's insert size, GenBank/EMBL/DDBJ accession number, hybridizing marker or clone, the chromosome on which it's located and the nearest genetic marker from the Lister and Dean recombinant inbred (RI) map (3,4).

A complete physical map of BAC clones is not currently available. The AGI is using several methods to choose BAC clones dispersed throughout the genome, including hybridization to either mapped YAC clones or RFLP probes. A summary of the AGI effort has been described by Bevan *et al*. (5). As a complete physical map is not yet available, the AGI sequencing summary-graphic must use the genetic map as its base-coordinate system. The inclusion of the nearest RI marker to the BAC being sequenced is thus important as it allows the clone to be placed on the graphical display described below (Fig. 1).

Upon submission of the form, various checks are automatically conducted before the database is updated. These include, but are not limited to, ensuring that the clone name is standard and that the clone's sequencing status has progressed. Once these checks have been satisfactorily completed, a summary of the submission is displayed to the submitter and the updated data are incorporated immediately into the database and become instantly available to database users.

If a mistake has been made upon entry of the clone's details, a de-registration web-form is available. AGI members can also use this form to de-register a clone upon which they have ceased to work. This can happen for a variety of reasons, such as: either the clone becoming redundant as others are identified; or technical problems, e.g., the clone could not be grown.

In a recent enhancement, AtDB's weekly list of new or changed *Arabidopsis* sequences in GenBank (see below) is scanned by a Perl program for AGI-registered clones. Any AGI-database entry found to have a missing GenBank accession number, an updated sequence size, or an out-of-date sequencing status is automatically updated to reflect the clone's status in GenBank. This has the advantage both of reducing the effort required by the participating groups to keep their AGI records in AtDB current; and also of helping to keep the database as up-to-date as possible. This automatic updating of the AGI entries in the database, based on GenBank entries, will gain in importance as the number of clones being sequenced at any one time continues to increase as the AGI groups become ever-more efficient.

ON-THE-FLY GRAPHICAL DISPLAY OF AGI DATA

Users gain access to AGI data simply by clicking on the 'AGI' button in the Table of Contents of AtDB's database (see Table 1 for URL). The left-hand frame is used to access the AGI data in either of two easy-to-use ways.

(i) Selecting the 'Map-based search' button displays a schematic representation (genome cartoon) of the genome in the main frame (Fig. 1). Clicking on any region within the bars representing the five chromosomes of *Arabidopsis* gives a magnified view of a 10 cM region of the genome showing the AGI clones together with nearest RI- and hybridizing-markers (Fig. 2). Clicking on the clones presents the details of their sequencing status and anything else available about them within the database. In both the genome cartoon and the enlarged views, colour is used to indicate the status of each clone. In the latter, the top of the window shows the chromosome containing the 10 cM magnified region and users can move to any other part of the same chromosome simply by clicking in the appropriate region. Left and right arrows move the magnified view to the adjacent 10 cM regions.

Table 1. AtDB URLs

AtDB home page	http://genome-www.stanford.edu/Arabidopsis/
AtDB Database Table of Contents	http://genome-www3.stanford.edu/atdb_welcome.html
BLAST Arabidopsis-only sequences	http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-blast2atdb
FASTA Arabidopsis-only sequences	http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-fastaatdb
AGI information and links	http://genome-www3.stanford.edu/cgi-bin/Webdriver?MIval=atdb_registry_info.html
AGI data-release policy	http://genome-www.stanford.edu/Arabidopsis/AGI/AGI_data_release.html

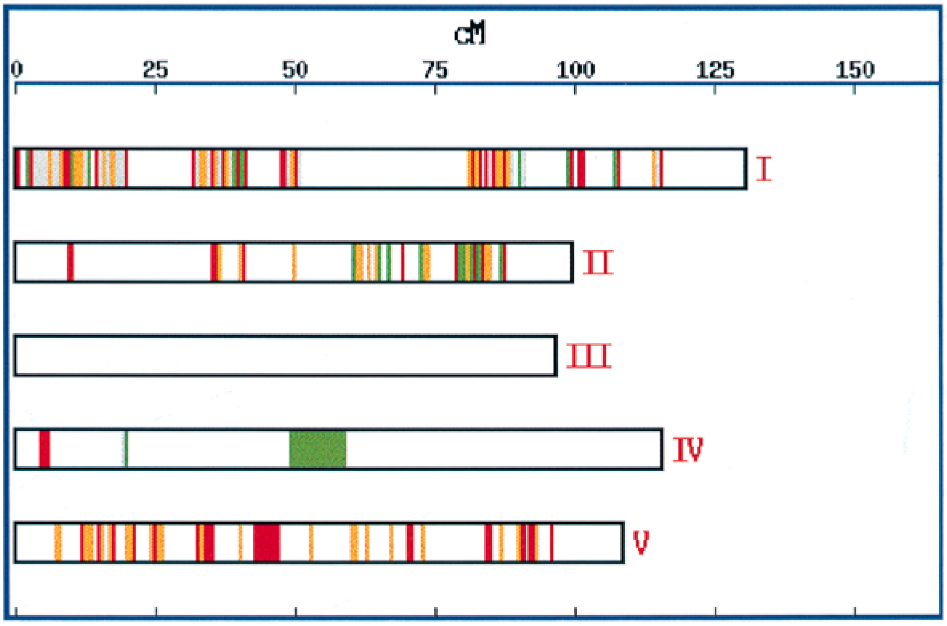

Figure 1. *Arabidopsis* Sequencing View. This shows the position and sequencing status (small, coloured, vertical bars) of clones in production by AGI, the international sequencing effort, across the five chromosomes of *Arabidopsis*. Clicking on a region returns a detailed view of that area (see Fig. 2).

As mentioned above, a precise physical-map for the genome is not yet available. However, defined maps of local regions are known. In addition, as neighbouring BACs are sequenced the precise amount of overlap is determined. With either of these cases, the orientation of the overlapping BAC cluster within the genome is typically not known. This will be resolved as the BAC cluster grows to include BACs known to contain dispersed, mapped genetic markers. To signify the uncertainty of the BAC cluster's orientation, a small box is drawn, in the enlarged view, above the BAC group, which says 'Flip Group' (Fig. 2). Clicking on this box results in a new image being drawn with the BAC cluster flipped to the opposite orientation. This allows the user to visualize the alternative ways in which clones may be positioned in the genomic region of interest.

(ii) Selecting single or multiple search options using one or more of the following criteria: clone, (AGI) group, (clone) library, chromosome, sequencing status, nearest recombinant inbred marker, and date. After making his or her selections, and pressing the 'Submit' button, the user is presented with the list of clones that match the search criteria. Clone names are Web links and take the user to the same magnified, clickable, graphical view (Fig. 2) obtained by clicking on the schematic representation of the whole genome (Fig. 1), as described above. Also displayed in the results of search are the status of the sequencing of each clone, the date of database entry, and links to AtDB 'colleague' entry for the person and AGI group of the supplier of the data. In addition, where the search criteria include a group that is a member of a consortium, clones matching the search criteria, but from other members of the same consortium, are also shown. This is done as the same clone may have different stages of its sequencing performed by more than one member of a given consortium. An additional hyperlink provides a tabular summary of sequences produced by AGI.

GRAPHICAL DISPLAYS OF GENETIC- AND PHYSICAL-MAP INFORMATION

The graphical displays used to represent AGI data are at the heart of recent developments from the AtDB project to provide an easy and effective view of the information contained within AtDB. All of these images are clickable client-side ISMAPs. The images are created by cgi programs written in C that directly, on-the-fly, query the database and then produce a GIF image plus the appropriate HTML code to add hyperlinks to the images. This is important because it means that when the data in the database are changed, this is instantly reflected in the graphical display presented to the user. This is not the case with conventionally-produced images, which need to be updated 'by hand'. This is a time-consuming task, which often results in many genomic and other images that are available over the web being out-of-date. Java provides another approach to providing on-the-fly graphics over the Web, but our experience with the use of C-programs that query the database—or, indeed, any table inside or outside a database, e.g., see Cherry *et al.* (1)—is that they are much quicker than Java and provide an appropriately simple, intuitive interface on any computer platform.

In addition to AGI data, these on-the-fly graphical displays are also used to show both genetic- and physical-map information. Users gain access to genetic-map displays simply by clicking on the 'Gen. Maps' button in the Table of Contents of AtDB's database. They may then select a genetic map, from one of three of greatest community interest: the Lister and Dean (3,4) recombinant inbred map; the 'classical' map of visible markers (6); or the mi-RFLP map (7). Other genetic maps will be displayed as they become available.

This initial graphic shows all five chromosomes with only a few landmark markers displayed. The user can click on any region and

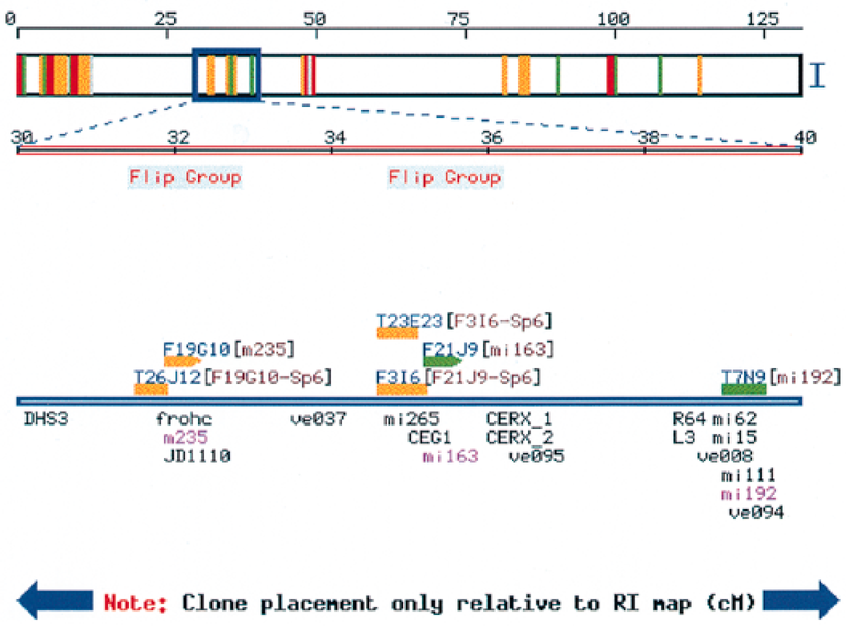

Figure 2. AGI Sequencing Graphic. Obtained by clicking on Figure 1. This shows in more detail the position and sequencing status (coded by colour) of the AGI-registered clones, which are in blue and clickable. The clone's hybridizing marker is given in [brackets]. The RI (genetic) markers are given below the narrow, blue box, with the nearest marker to a given clone shown in magenta and aligned left with the clone. The box titled Flip Group indicates that the orientation of the group of BACs below the box is not known relative the chromosome.

is then shown a 'zoomed-in' view (Fig. 3), which shows all the markers within a 10 cM region. The markers themselves are clickable and take the user to the relevant full-text entry in the database. The user can also go directly to the magnified view by selecting a marker name from the list, displayed in the left-hand frame, of all markers for a particular map. Manœuvring to adjacent regions is performed in the same way as for the AGI display (above).

A similar protocol is followed in order to obtain physical-map views. Like the other views, a summary image is available first and then a zoomed-in detailed view is presented where contig information is available in a given genomic region. There are, however, two types of detailed view. One is a tiling-path display, which is presented when simple probe-clone hybridization data are available for that contig. The other is a detailed physical-map display, where, in addition to the hybridization data, the size and physical relationships of the probes and clones are also known. Both the tiling-path and the detailed physical map displays are fully clickable and take the user to the appropriate entry in AtDB's database. The AtDB project is in the process of interconnecting the AGI and physical map views so the user can move from one display to the other.

Table 2. URLs of AGI (Arabidopsis Genome Initiative) participants' web sites

Centre National de Sequençage, France (CNS)	http://www.infobiogen.fr/CNS/Arabidopsis.html
CSH-WU-ABI Consortium:	
Cold Spring Harbor Laboratory, USA	http://www.cshl.org/arabweb
Washington University, St Louis, USA	http://genome.wustl.edu/gsc/
Applied Biosciences Incorporated	no URL known
John Innes Centre, UK [ESSA project]	http://www.uea.ac.uk/nrp/jic/
Kazusa DNA Research Institute, Japan	http://www.kazusa.or.jp/arabi/
SPP Consortium:	
Stanford DNA Sequencing & Technology Center, Stanford Univ.	http://sequence-www.stanford.edu/ara/ArabidopsisSeqStanford.html
Arabidopsis thaliana Genome Center, Univ. of Pennsylvania (ATGC)	http://cbil.humgen.upenn.edu/~atgc/ATGCUP.html
Plant Gene Expression Center, USDA	http://pgec-genome.pw.usda.gov
The Institute for Genomic Research (TIGR)	http://www.tigr.org/tdb/at/atgenome/atgenome.html

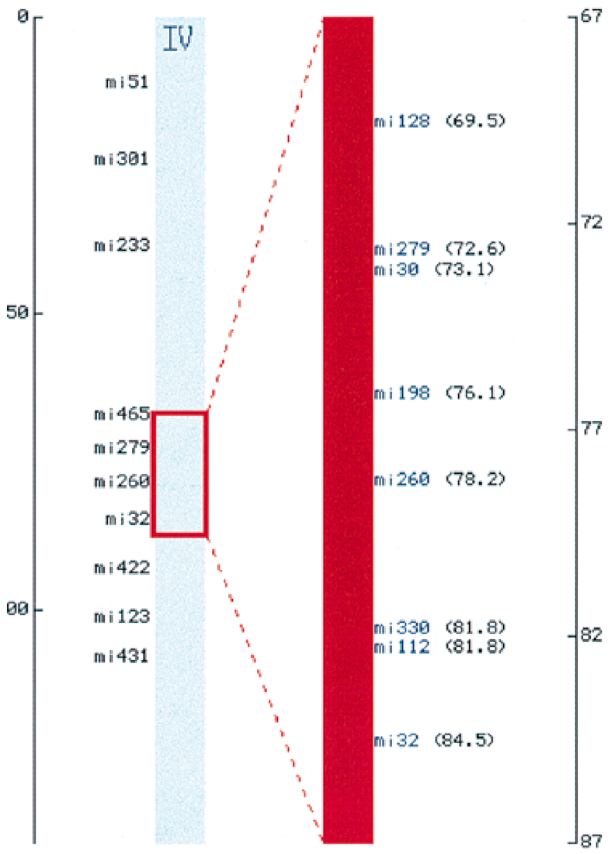

Figure 3. mi-RFLP Markers Graphic. Obtained by clicking on a cartoon of the genome, similar to that shown in Figure 1. This graphic shows all the markers within a 10 cM region. The markers themselves are clickable and, as with the 'hot' clones in Figure 2, take the user to the relevant full-entry in the database. Similar clickable genetic-map displays are available for the Lister and Dean (3,4) recombinant inbred map and the 'classical' map of visible markers (6).

BLAST AND FASTA

For some time, AtDB has been providing a web BLAST (8) and FASTA (9) service for matches against only *Arabidopsis* sequences. The two datasets used for these sequence similarity searches are:

(i) All sequences from GenBank that include the word 'Arabidopsis' in the 'Organism' line. This dataset is updated three times a week. (In a related service, a weekly summary is automatically generated and posted to the *Arabidopsis* electronic newsgroup [bionet.genome.arabidopsis]. This comprises *Arabidopsis* accessions that are either new or existing, but that have changed in their sequence or annotation.)

(ii) Non-redundant *Arabidopsis* protein sequences, these are created by merging all entries with identical sequences to form a unique set of sequences from GenPept, SwissPROT and PIR. The GenPept entries are updated three times a week, SwissPROT entries are updated weekly, and PIR entries updated with each major release.

A third dataset has recently been added. This comprises the BAC clone end-sequences being produced by TIGR, ATGC and CNS (Table 2). BAC end sequences have become an integral part of the AGI strategy to create minimum tiling paths of BAC clones to span the entire genome. At the time of writing, over 17 000 end sequences are available with many hundreds being added every week. Once an initial BAC sequence is known, the end-sequence dataset can be used to determine the neighbouring BACs with a simple BLAST search. The end sequences are also useful to the community because they can be used to identify a BAC of interest that is in the early stages of sequencing, but for which no sequence has yet been released (apart from the end-sequence).

In an improvement to the display of the BLAST result pages, an on-the-fly graphical summary of the query matches is now produced and displayed with each query run. This cartoon shows the query sequence and the position and size of the matches to it. Users can specify the percentage minimum match cut-off to increase or reduce the number of hits shown on the graphic. With queries to the BAC end-sequence dataset, the displayed clones are clickable and link to the end-sequence within AtDB's database. From that page, there are links to GenBank and, if the end-sequence is from a clone being fully sequenced by AGI, to that clone's database entry and graphical display. For the *Arabidopsis* GenBank DNA and non-redundant protein datasets, hot-links to the appropriate database will soon be added.

ACKNOWLEDGEMENTS

The authors thank Gail Juvik and Mark Schroeder for software support and Mark Mooney, Gail Juvik and Laura Donohue for helpful comments on the manuscript. Funding for AtDB is provided by the National Science Foundation, grant numbers DBI-9503776 and DBI-9714160.

REFERENCES

1 Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S., Hester,E., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) *Nucleic Acids Res.*, **26**, 73–79.
2 Meyerowitz,E.M. and Somerville,C.R. (eds) (1994) *Arabidopsis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3 Lister,C. and Dean,C. (1993) *Plant J.*, **4**, 745–750.
4 Anderson,M. (1997) *WeedsWorld*, **4** http://nasc.nott.ac.uk:8300/home.html
5 Bevan,M., Ecker,J., Federspiel,N., Davis,R., McCombie,D., Martienssen,R., Chen,E., Waterson,B., Wilson,R., Rounsley,S., *et al.* (1997) *Plant Cell*, **9**, 476–478.
6 Meinke,D. (1997) http://mutant.lse.okstate.edu/comm_linkage_list.html
7 Liu,Y.G., Mitsukawa,N., Lister,C., Dean,C. and Whittier,R.F. (1996) *Plant J.*, **10**, 733–736.
8 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
9 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.