D354–D359 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh046

PlantGDB, plant genome database and analysis tools

Qunfeng Dong[1], Shannon D. Schlueter[1] and Volker Brendel[1,2,*]

[1]Department of Genetics, Development and Cell Biology and [2]Department of Statistics, Iowa State University, Ames, IA 50011-3260, USA

Received August 15, 2003; Revised and Accepted September 22, 2003

ABSTRACT

PlantGDB (http://www.plantgdb.org/) is a database of molecular sequence data for all plant species with significant sequencing efforts. The database organizes EST sequences into contigs that represent tentative unique genes. Contigs are annotated and, whenever possible, linked to their respective genomic DNA. Genome sequence fragments are assembled similarly. The goal of the PlantGDB web site is to establish the basis for identifying sets of genes common to all plants or specific to particular species by integrating a number of bioinformatics tools that facilitate gene prediction and cross-species comparisons. For species with large-scale genome sequencing efforts, PlantGDB provides genome browsing capabilities that integrate all available EST and cDNA evidence for current gene models (for *Arabidopsis thaliana*, see the AtGDB site at http://www.plantgdb.org/AtGDB/).

INTRODUCTION

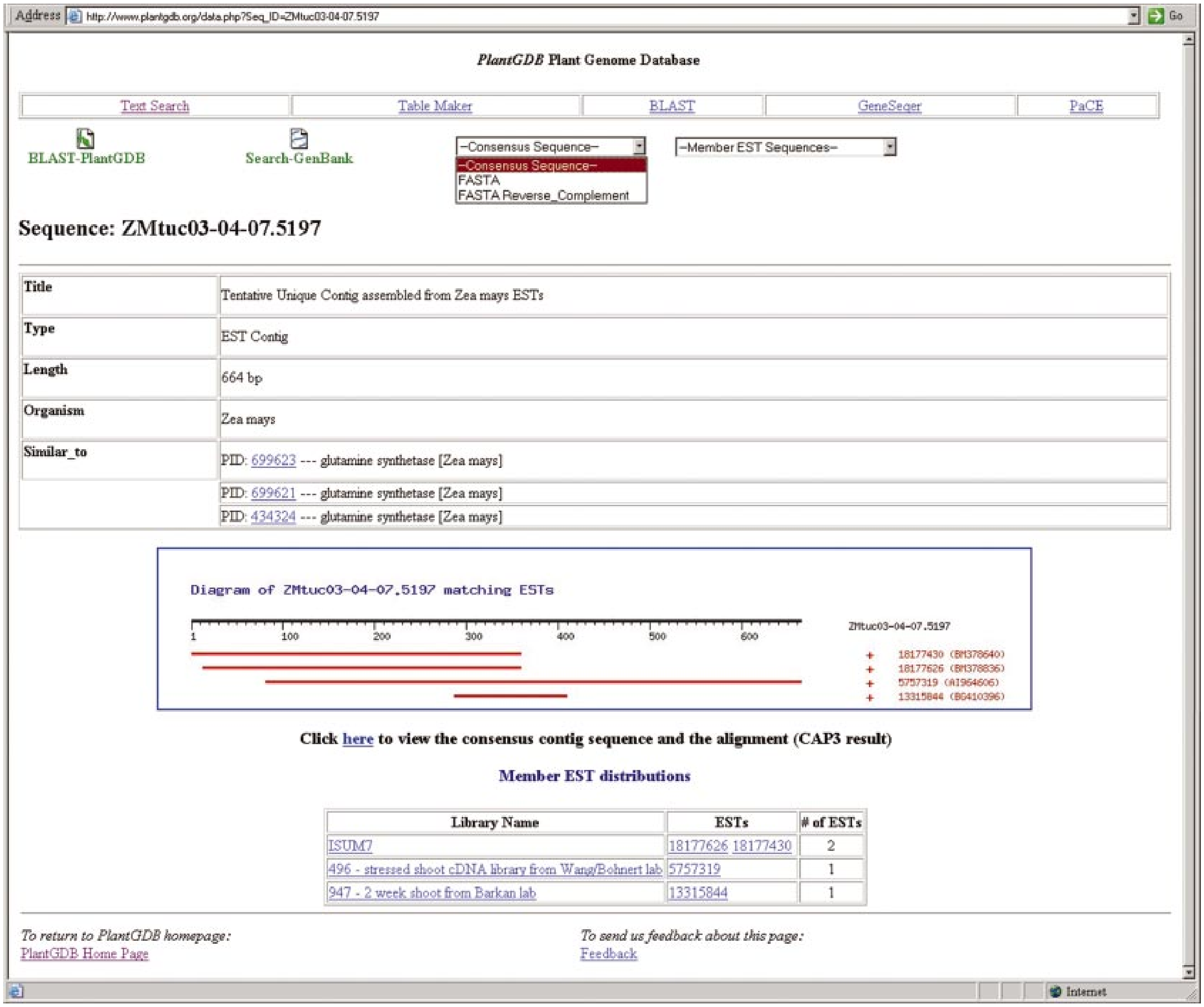
Comparative and functional genomics of (plant) species seek to characterize each species' gene content and chromosomal arrangements, to explain observed similarities and differences within a molecular evolutionary context, and to assess functional significance of distinct genetic blueprints. Ultimately, completion of these tasks will require analysis and comparison of whole genome sequences. Prokaryotes have what are now considered 'small' genomes, and only a few years after sequencing of the first entire bacterial genome (1) more than 100 complete genomes are available (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html) and have been compared in detail (2). Extrapolating from this remarkable progress in sequencing technology in such a short time-span, we can be confident that sequences of complex plant genomes will also become abundantly available. Until such time, researchers must continue to piece together glimpses of plant gene space from the available bits and pieces of sequence and map data, using the nearly complete genomes of *Arabidopsis thaliana* (3) and rice (4,5) as both model and scaffold. The extent to which the compact genomes of *Arabidopsis* (125 Mb) and rice (430 Mb) are adequate models for other dicot and cereal species, respectively, remains an intensely debated issue (6–8).

In addition to whole genome sequencing, plant sequence data have been accumulating from three major sources: sample sequencing of bacterial artificial chromosomes (BACs), genome survey sequencing (GSS) and sequencing of expressed sequence tags (ESTs). We developed PlantGDB (http://www.plantgdb.org/) as an integrated database and suite of analytical tools to organize and interpret these data. As such, PlantGDB will be instrumental for making the difficult decisions concerning which species to select for reference and draft sequencing, which is dependent upon the assessment of the differences in gene space and genome organization between related species. Our resource will help to answer important questions like, 'How representative is *Arabidopsis* of all plants? How good a model is rice for the other cereals, or *Medicago* for other legumes?' The pan-species views and organization of PlantGDB provide a foundation from which to address questions of functional and evolutionary genomics. For example, cross-species EST comparisons help to identify orthologs conserved among different taxonomic groups (9) and contribute to studies of the origin and fate of duplicated genes (10). Clustering and alignment of ESTs (to one another and to genomic DNA) reveals the existence of alternative splicing patterns and the extent to which these alternative splicing patterns are conserved (11). In the following, we first discuss the design principles of PlantGDB and then describe our implementation, including EST contig display, a genome browser for gene structure annotation and spliced alignment, and tutorial and outreach functions.

DATABASE DESIGN

There are several prominent public databases that provide access to plant genome data. These include general repositories, such as GenBank (12), and species-specific or data-type-specific resources, such as TAIR (13), MaizeGDB (14), Gramene (15) and Stanford Microarray Database (16); of course, larger database providers such as NCBI (http://www.ncbi.nlm.nih.gov/) and TIGR (http://www.tigr.org/) function as both repositories and specific resources. Repositories are typically the addresses to visit when one wishes to retrieve records for further analysis to be performed locally. A major limitation, evident in many applications, is the redundancy and the lack of curation in such repositories. For example, retrieving protein records from GenBank by keywords will typically result in a superset of the intended sequence collection that includes duplicated and misannotated records. In contrast, specialized databases provide curated records that have detailed and updated information. However,

[*]To whom correspondence should be addressed. Tel: +1 515 294 9884; Fax: +1 515 294 6755; Email: vbrendel@iastate.edu



PlantGDB **Plant Genome Database**

Sequence: ZMtuc03-04-07.5197

Title	Tentative Unique Contig assembled from Zea mays ESTs
Type	EST Contig
Length	664 bp
Organism	Zea mays
Similar_to	PID: 699623 --- glutamine synthetase [Zea mays]
	PID: 699621 --- glutamine synthetase [Zea mays]
	PID: 434324 --- glutamine synthetase [Zea mays]

Click here to view the consensus contig sequence and the alignment (CAP3 result)

Member EST distributions

Library Name	ESTs	# of ESTs
ISUM7	18177626 18177430	2
496 - stressed shoot cDNA library from Wang/Bohnert lab	5757319	1
947 - 2 week shoot from Barkan lab	13315844	1

To return to PlantGDB homepage:
PlantGDB Home Page

To send us feedback about this page:
Feedback

Figure 1. EST contig display at PlantGDB. The screenshot represents a typical display of an EST contig record at PlantGDB, assembled from species-specific EST collections using the PaCE (19) and CAP3 (20) programs. The central panel provides basic annotation for the sequence, including tentative functional annotation based on significant protein-level similarities. Sequence similarity search tool functions (BLAST) are linked via the icons on top (selection pastes the sequence into the query screen of the respective tool), as are download functions for the sequences. The diagram in the lower panel displays a schematic of the multiple sequence alignment derived with the CAP3 program. The black line represents the contig consensus sequence, and the red lines indicate each member EST. The actual alignments can be viewed by clicking on the link below the diagram. The table at the bottom links to the library sources of member ESTs and individual sequence records. GSS contigs are displayed similarly (not shown).

it is usually impossible to perform more analyses directly at these resource sites: often they only support browsing of pre-calculated data points, and further analysis would require raw data download and local analysis. PlantGDB is in part a resource for similar look-up methods, but our emphasis is on its dual design as a web-based research tool. Thus, all records in PlantGDB are linked to tools that allow any user immediate use of the data in other applications and the ability to re-calculate curated records using updated or proprietary data. For example, PlantGDB includes EST clusters and assemblies for all major plant species [similar to the TIGR gene indices (17) available for some of the species]. The consensus sequence representing an assembly is automatically annotated by reporting three significant matches to protein records in GenBank, based on BLAST (18) results. Both assembly and annotation can be easily repeated on site using tools built into the sequence record display pages with current or additional data, or other than default parameters (Fig. 1).

PlantGDB also seeks to make database technology available to individual molecular biology and genomics research groups. This aim is based on the premise that software tools should be as widely distributed as are significant laboratory techniques. For example, it is taken for granted that almost any molecular biology laboratory today should be able to conduct experiments in house based on, for example, semi-quantitative PCR, the yeast two-hybrid system, or even microarray technology. Biologists' software literacy is comparatively poor. In standard use are query searches, multiple sequence alignments and molecular phylogeny studies. We firmly believe that database technology is essential to modern genome research and must be (and eventually will be) acquired as a research tool in common use (rather than remaining within the domain of specialists). After all, biological research often relies upon mapping relationships between different data points, the forte of modern relational databases. Thus, PlantGDB is committed to providing domain-specific database implementations that are entirely portable, with extensive documentation. Currently, our *Arabidopsis* gene structure database and genome browser (AtGDB, described below) has already been successfully copied to several stand-alone locations, enabling genome-scale biological research on alternative splicing and distribution and conservation of U12-type introns (11,21).

DATABASE IMPLEMENTATION

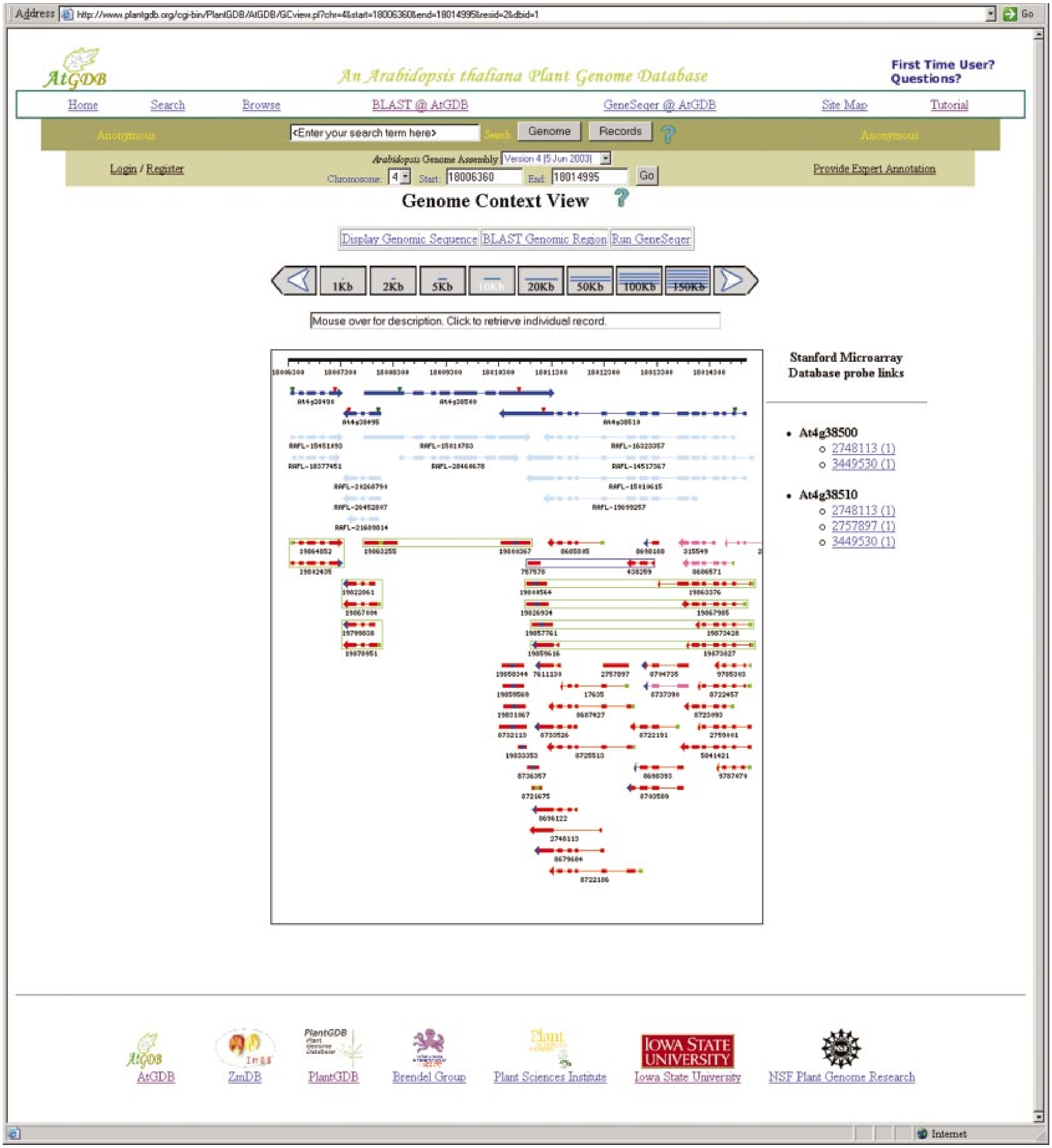
Currently, all raw sequence data in PlantGDB are obtained by periodic upload from GenBank. Sequence types include protein, EST (http://www.ncbi.nlm.nih.gov/dbEST), GSS (http://www.ncbi.nlm.nih.gov/dbGSS), STS (http://www.ncbi.nlm.nih.gov/dbSTS), HTG (http://www.ncbi.nlm.nih.gov/HTGS), cDNA and other genomic DNA sequences from 50 plant species with major sequencing efforts. However, these data types are not stored in PlantGDB by way of simple duplication of GenBank records; rather, storage involves reorganization, curation and processing. For example, the GenBank EST records, originally stored in Abstract Syntax Notation One (ASN.1) format, are processed to extract selected fields such as library information including tissue type, developmental stage, etc. The extracted information is then stored in PlantGDB MySQL (http://www.mysql.com) relational database tables. Such information extraction and reorganization enables meaningful biological queries that may be difficult or even impossible to carry out at GenBank. For example, it is currently impossible to execute a simple query like 'display all the leaf ESTs generated from maize inbred B73' at GenBank, even though the necessary information is actually embedded in the GenBank EST records. By contrast, it is easy to obtain such results by querying PlantGDB using the TableMaker tool (http://www.plantgdb.org/TableMaker.php) described below.

Because ESTs usually correspond to only partial cDNA sequences, and because EST samples typically are highly redundant, the PlantGDB EST processing pipeline includes frequently updated EST clustering and assembly into putative unique transcript contigs (http://www.plantgdb.org/ESTCluster/progress.php). Similarly, PlantGDB also assembles GSS sequences, although display is currently limited to maize GSS contigs assembled from the *RescueMu* transposon tagged GSS sequences (22). Both EST and GSS assembly are critical for plant gene discovery and many other important applications. For example, our maize EST assembly has been used for developing maize microarray cDNA probes (23). The ESTs and EST contigs are also integrated to our PlantGDB-specific GeneSeqer server (24) for gene structure annotation by spliced alignment to genomic sequences (11). In addition, all ESTs, EST contigs, as well as other types of sequences such as GSS are consistently going through our functional annotation pipeline to have putative protein function assigned by running BLAST against public protein databases. Finally, all information is interconnected and collectively displayed on the web (Fig. 1).

The web-based query capabilities of current biological databases are often quite limited, in part because of the practical necessity to restrict SQL access. For example, a typical text search for *myb* genes at GenBank allows users to retrieve a list of records matching the *myb* keyword. However, biologists often demand more than just a list. Depending on their purpose, some may need a simple two-column table, with one column displaying 'Sequence ID' and the other displaying 'Organism' to indicate the source of the sequence. Others may need a more complex table with columns 'Sequence ID', 'Organism', 'Tissue Type' and 'Developmental Stage'. The complex nature of biological research implies that simple text search capabilities and fixed table reporting are not sufficient, presenting a key challenge for biological database development. More specifically, such tasks require a good front-end interface system, because the actual table-making requests are usually easily implemented for a modern back-end relational database system. At PlantGDB we seek to provide dynamic query tools over the web so that users are able to easily construct their own queries based on the table structures we provide. The PlantGDB TableMaker tool, inspired by the legacy AceDB TableMaker tool (http://www.acedb.org/Software/whelp/Table_Maker.html), is our attempt to meet this challenge. The tool allows users to specify criteria on different columns and to select which columns to incorporate into a final report table. Then it translates the user's specifications to the back-end SQL query statements and presents the results back to the user in a tabular form. This enables flexible queries such as 'Show me all the maize ESTs, library name, tissue type, and EST contigs they belong to' by simply filling out web forms, not requiring any knowledge of SQL.

Our development of analysis tools has focused on cross-species sequence comparisons. In particular, we have set up BLAST (http://www.plantgdb.org/cgi-bin/PlantGDBblast) and GeneSeqer (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi) web services that allow users to compare sequences of interest (using BLAST) and 'thread' EST/cDNA into genomic DNA (using GeneSeqer) across all species. The BLAST@PlantGDB server implements the standard NCBI BLAST as its search engine (18). The unique feature of BLAST@PlantGDB is its flexibility in database selection. Users are able to search against single or multiple BLAST databases of their choice simultaneously (e.g. only rice or maize ESTs, or both; or all monocot ESTs; or rice ESTs and all cereal EST contigs). The GeneSeqer@PlantGDB server implements the GeneSeqer spliced alignment program (24). It produces plant gene structure models based on spliced alignment to genomic sequences of both native and homologous ESTs, cDNAs and protein sequences. By integrating the stand-alone GeneSeqer program with back-end database operations, users can conveniently align ESTs of specific quality or origin: for example ESTs generated from a specific tissue can be evaluated to study tissue-specific gene expression.

The PlantGDB genome browsing capabilities are exemplified for *Arabidopsis* by way of the companion *A.thaliana* Genome Database (AtGDB; http://www.plantgdb.org/AtGDB/).



Address http://www.plantgdb.org/cgi-bin/PlantGDB/AtGDB/GCview.pl?chr=4&start=18006360&end=18014995&resid=2&dbid=1

An Arabidopsis thaliana Plant Genome Database

First Time User? Questions?

Home | Search | Browse | BLAST @ AtGDB | GeneSeqer @ AtGDB | Site Map | Tutorial

Anonymous | <Enter your search term here> | Search | Genome | Records | Anonymous

Login / Register

Arabidopsis Genome Assembly: Version 4 (5 Jun 2003)

Chromosome: 4 | Start: 18006360 | End: 18014995 | Go

Provide Expert Annotation

Genome Context View

Display Genomic Sequence | BLAST Genomic Region | Run GeneSeqer

Mouse over for description. Click to retrieve individual record.

Stanford Microarray Database probe links

- At4g38500
  - 2748113 (1)
  - 3449530 (1)
- At4g38510
  - 2748113 (1)
  - 2757897 (1)
  - 3449530 (1)

Figure 2. AtGDB visualization of current genome annotation and cDNA and EST spliced alignments. The figure illustrates AtGDB visualization of cDNA and EST spliced alignments compared with current GenBank gene structure annotation for a 9 kb region on chromosome 4. Exons are indicated by solid boxes, connected by lines representing intron sequences. The arrows denote 3′-ends. Dark blue, current GenBank mRNA annotation. Light blue, cognate cDNA spliced alignments. Red and pink, cognate and non-cognate EST spliced alignments. Green and blue indicators within the EST structures represent 5′- and 3′-clone end designators, respectively. Green boxes surrounding EST structures associate members of a clone pair. cDNAs and ESTs are labeled by their GenBank GI numbers. Note the example of typical difficulties of automated gene structure annotation: the 3′-ends of genes At4g38500 and At4g38510 are annotated as overlapping at GenBank (dark blue), although several full-length cDNAs clearly indicate the correct 3′-ends of both genes. The erroneous annotation is likely caused by assignment of single-exon 3′-ESTs from At4g38500 transcripts to pseudo-transcripts of both At4g38500 and At4g38510. The importance of correct gene structure annotations (and visualization of all evidence) is underscored by the links to gene expression data accessible at the Stanford Microarray Database: EST probes 2748113 and 3449530 could not be resolved to a single gene based on the current genome annotation, and probe 2757897 is seen to contain an intron.

AtGDB stores EST and cDNA spliced alignments along with current *Arabidopsis* genome annotation. An elaborate web interface was designed for the database to allow users to browse the genome and query the database by sequence similarity, identifiers or descriptions (Fig. 2). In general, the web interface is composed of three parts: the genomic context view, the query view and the sequence view. The genomic context view allows users to browse a specific genomic region in the context of multiple annotation resources. The region graphic displays these multiple sources of alignment information relative to one another. Each is colored with respect to its specific annotation source. The query view allows users to view and interact with the results of a user query. Each stored EST/cDNA alignment and annotated transcript has an individual page with a sequence view, which glues together sequence data, analysis tools and related external links. This web interface efficiently presents the database entries on-the-fly and facilitates data access and utilization. Development of analogous genome browsers for rice and other plant species with substantial BAC sequencing projects is currently under way. Our development strategy facilitates the sharing of user-contributed gene structure annotations. For example, for *Arabidopsis*, a specialized GeneSeqer server can be accessed from any genome region display window, run with *Arabidopsis* and other EST or protein targets, and the resulting gene structure predictions, if unambiguous, can be contributed to the database for general display (after a curator's approval). In our view, a community effort is essential for comprehensive and accurate genome annotation given the inevitable difficulties with automated annotation pipelines (cf. Fig. 2).

SERVICE

As a supplement to our PlantGDB grant, we have set up a Plant Genome Research Outreach Portal (PGROP) to provide a centralized repository of various NSF-sponsored Plant Genome Research 'outreach' programs and activities. The website (http://www.plantgdb.org/outreach/) lists a variety of resources organized by subject and type and provides entry forms for contributors to publish their programs. The intended user audience includes minorities, students with disabilities, undergraduates, high-school students and teachers, as well as the public at large. PGROP seeks to broaden participation of these user groups in plant genome research topics by making information on appropriate programs, materials and guidance available.

AVAILABILITY

PlantGDB is accessible at the URL http://www.plantgdb.org/. Data files and source code for the programs used at PlantGDB can be downloaded from links on the home page. The manager of the database can be contacted by email at plantgdb@iastate.edu.

ACKNOWLEDGEMENTS

PlantGDB is supported by USA National Science Foundation Plant Genome Research Projects grants DBI-0110254 and DBI-0321600.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Microevolutionary genomics of bacteria. *Theor. Popul. Biol.*, **61**, 435–447.
3. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
4. Goff,S.A., Ricke,D., Lan,T.-H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
5. Yu,J., Hu,S., Wang,J., Wong,G.K.-S., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
6. Zhu,H., Kim,D.-J., Baek,J.-M., Choi,H.-K., Ellis,L.C., Küster,H., McCombie,W.R., Peng,H.-M. and Cook,D.R. (2003) Syntenic relationships between *Medicago truncatula* and *Arabidopsis* reveal extensive divergence of genome organization. *Plant Physiol.*, **131**, 1018–1026.
7. VandenBosch,K.A. and Stacey,G. (2003) Summaries of legume genomics projects from around the globe. Community resources for crops and models. *Plant Physiol.*, **131**, 840–865.
8. Chandler,V.L. and Brendel,V. (2002) The maize genome sequencing project. *Plant Physiol.*, **130**, 1594–1597.
9. Fulton,T.M., Van der Hoeven,R., Eannetta,N.T. and Tanksley,S.D. (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, **14**, 1457–1467.
10. Meyers,B.C., Morgante,M. and Michelmore,R.W. (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J.*, **32**, 77–92.
11. Zhu,W., Schlueter,S.D. and Brendel,V. (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.*, **132**, 469–484.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
13. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *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.
14. Lawrence,C.J., Dong,Q., Polacco,M.L., Seigfried,T.E. and Brendel,V. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **32**, D393–D397.
15. Ware,D.H., Jaiswal,P., Ni,J., Yap,I.V., Pan,X., Clark,K.Y., Teytelman,L., Schmidt,S.C., Zhao,W., Chang,K. *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol.*, **130**, 1606–1613.
16. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al.* (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.
17. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2003) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
18. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Kalyanaraman,A., Aluru,S., Kothari,S. and Brendel,V. (2003) Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res.*, **31**, 2963–2974.
20. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
21. Zhu,W. and Brendel,V. (2003) Identification, characterization, and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.*, **31**, 4561–4572.

22. Dong,Q., Roy,L., Freeling,M., Walbot,V. and Brendel,V. (2003) ZmDB, an integrated database for maize genome research. *Nucleic Acids Res.*, **31**, 244–247.
23. Fernandes,F., Brendel,V., Gai,X., Lal,S., Chandler,V.L., Elumalai,R.P., Galbraith,D.W., Pierson,E.A. and Walbot.V. (2002) Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. *Plant Physiol.*, **128**, 896–910.
24. Schlueter,S.D., Dong,Q. and Brendel,V. (2003) GeneSeqer@PlantGDB: gene structure prediction in plant genomes. *Nucleic Acids Res.*, **31**, 3597–3600.