Annotation of the Arabidopsis Genome[1]

Jennifer R. Wortman, Brian J. Haas, Linda I. Hannick, Roger K. Smith, Jr., Rama Maiti, Catherine M. Ronning, Agnes P. Chan, Chunhui Yu, Mulu Ayele, Catherine A. Whitelaw, Owen R. White, and Christopher D. Town*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850

The Arabidopsis Genome Sequencing Project was officially completed in late 2000, leading to the publication of a landmark paper describing, in broad outline, many salient features of the Arabidopsis genome (Arabidopsis Genome Initiative [AGI], 2000). However, the genome annotation, generated by the individual sequencing centers, was heterogeneous, both in terms of gene structure predictions and terminology used in their description. In response, The Institute for Genomic Research (TIGR) was funded by the National Science Foundation to carry out a whole-genome reannotation that would be of a uniform high standard and consistency, eliminating the heterogeneity that had accumulated over time in the public databases.

The first phase of this process took place in the latter half of 2000, as the sequencing itself was drawing to a close. At that time, it was agreed that two centers, The Munich Information Center for Protein Sequences (MIPS) and TIGR, would carry out the bulk of the analyses for the whole-genome publication. As part of this effort, all publicly available sequence and associated annotation was retrieved from various databases and loaded into the TIGR annotation database. Bacterial artificial chromosomes (BACs) that either lacked annotation or required review were examined by TIGR annotators to provide as complete and nonredundant a data set as possible for publication. When the whole-genome analysis, carried out in close collaboration with MIPS, was completed, TIGR embarked upon the reannotation proper, using the integrated annotation data set as a starting point. The purpose of this article is to describe the goals, chronology, and accomplishments to date of this reannotation effort and to communicate the nature, quality, and basis of the latest TIGR data release to the plant research community.

A CENTRALIZED REANNOTATION EFFORT

The stated goals of the Arabidopsis reannotation effort are the comprehensive identification of protein coding genes, the elucidation of accurate gene structures, and the assignment of function to each gene product in the predicted proteome. The approach taken at TIGR involves both automated annotation and manual curation, including the development of novel algorithms and custom software interfaces to facilitate the generation of data that are complete, thorough, and of high quality.

The first challenge is to generate and maintain an integrated, nonredundant set of gene models across the Arabidopsis genome. This involves incorporating data from multiple original sources and recognizing and correcting for duplicate and partial gene models on overlapping BACs. The growing availability of full-length cDNA (FL-cDNA) and expressed sequence tag (EST) data enables robust computational corrections to the original annotated gene structures. Protein data and genomic reads from an evolutionary neighbor, such as *Brassica oleracea*, can also inform computational gene prediction.

With a comprehensive set of gene models available, the predicted proteome can be clustered, making it possible to examine, model, and describe genes in the context of gene families. This can highlight inconsistencies in previous annotations and aids in achieving consistency and accuracy in the manual curation of both gene structure and function. Annotation updates are incorporated weekly into the TIGR Arabidopsis Annotation Web site.

TIGR publishes its annotation data in the context of Arabidopsis chromosome sequences, generated based on the available tiling path information. Where tiling path data are incomplete or inconsistent, overlaps are being validated experimentally. Updated chromosome sequences along with annotation are generated and released twice a year to the Arabidopsis Information Resource (TAIR) and the Genome Division of the National Center for Biotechnology Information (NCBI).

GENERATING A COMPREHENSIVE GENE SET

To support the reannotation effort, an in-house database was populated with a complete set of BAC sequences representing the published Arabidopsis

[1] This work was supported by the National Science Foundation (Cooperative Agreement no. DBI 9813586).

* Corresponding author; e-mail cdtown@tigr.org; fax 301–838–0208.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.103.022251.

tiling path, along with the genes and other biological features identified on these genomic sequences by the respective sequencing centers. These sequences were then processed through the TIGR annotation pipeline, a collection of software known as Eukaryotic Genome Control (EGC) that serves as the central data management system.

EGC processes each BAC sequence through a series of algorithms for predicting genes (Genscan+, Genemark.hmm, and Glimmer; Burge and Karlin, 1997; Lukashin and Borodovsky, 1998; Salzberg et al., 1999), splice sites (Hebsgaard et al., 1996; Pertea et al., 2001), and tRNAs (Lowe and Eddy, 1997). Homology to nucleotide and protein databases is computed using the AAT package (Huang et al., 1997), which utilizes a two-step approach consisting of a fast database search step to identify the boundaries of the sequence match, followed by a rigorous alignment step which takes into account consensus splice signals. Data sets include Arabidopsis-specific cDNA and EST sequences, TIGR gene indices for Arabidopsis and other plants (Quackenbush et al., 2000), a nonredundant amino acid database filtered from public sources, and SwissProt (Bairoch and Apweiler, 2000).

The output of the gene prediction algorithms and homology-based computes were compared with the preexisting gene models. New gene models were created, and existing gene models were updated based on different classes of evidence. The process of automating gene structure updates is continually being improved as new data becomes available and pipeline components become more robust. Current annotation data is based on the following data sets: Arabidopsis and plant cDNAs and ESTs downloaded from GenBank on January 28, 2003, the February 2003 release of TIGR gene indices, and a nonredundant protein set generated on February 1, 2003. Table I describes the numbers of gene models supported by different types of evidence in the current annotation data set.

Gene Models Supported by Arabidopsis FL-cDNA and EST Sequences

Early in 2001, we performed a detailed analysis with approximately 5,000 FL-cDNA sequences provided by Ceres Inc. (Malibu, CA) and evaluated the performance of a number of programs for their suitability in generating gene models based on the genomic alignments of these data. We found that approximately 35% of the cognate gene models with cDNA support required some kind of modification. During this pilot study, 240 new genes were discovered and modeled based upon FL-cDNA alignments. This analysis also supported the documentation of many instances of alternative splicing (Haas et al., 2002) and mini-exons (N. Volfovsky, B.J. Haas, and S.L. Salzberg, unpublished data) and has allowed us to develop a pipeline in which almost all new FL-cDNAs can be used to validate or update a gene model automatically. We have continued to download FL-cDNAs from GenBank. Currently, there are 24,839 cDNA accessions supporting 12,569 gene models for 11,960 genes. From the end user's perspective, gene models and predicted coding sequence based on FL-cDNAs can be regarded as high confidence, although a small percentage may be truncated or contain unspliced introns or genomic DNA.

Arabidopsis ESTs also provide strong support for components of gene structure, including exons, splice sites, and untranslated regions. Many of the original gene models relied, at least in part, on ESTs for their support. However, due to the distributed nature of the annotation and the gradual accumulation of sequences, EST evidence was not uniformly or systematically incorporated into gene models. We recently have developed an algorithmic approach to incorporate all EST evidence into gene models and, where appropriate, provide evidence for alternative splicing. As a result, the current annotation data set contains approximately 16,000 genes whose models incorporate all available EST evidence (B.J. Haas,

Table I. *Evidence supporting annotated Arabidopsis gene models*

Support was determined by BLAST-based similarity between the current, complete set of gene models/proteins, and the following datasets: non-Arabidopsis proteins parsed from the TIGR nonredundant protein set (February 1, 2003); the Arabidopsis protein set as represented in TIGR's annotation database (February 1, 2003); and Arabidopsis cDNAs, Arabidopsis ESTs, and non-Arabidopsis plant ESTs downloaded from GenBank (January 28, 2003). Note that matches between an Arabidopsis protein and itself were excluded from the count for Arabidopsis protein support.

Each gene was counted only once, in the category associated with the highest overall confidence. Gene models supported by both Arabidopsis cDNA and protein similarity to another organism are considered to be highest confidence. Genes with no EST and no protein support are based solely on gene predictions and are the lowest confidence set.

	Non-Arabidopsis Protein	Arabidopsis Protein	No Protein	Total
Arabidopsis cDNA	9,017	2,055	1,664	12,736
Arabidopsis EST	2,607	1,041	1,033	4,681
Other Plant EST	3,691	1,365	1,032	6,088
No cDNA/EST	247	2,280	1,352	3,879
Total	15,562	6,741	5,081	27,384

A.L. Delcher, J.R. Wortman, R.K. Smith Jr, L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, S.L. Salzberg, and O.R. White, unpublished data). Similar algorithms have been described previously (Kan et al., 2001), and an independent annotation of the Arabidopsis genome using similar principles has been performed by others (W. Zhu, S.D. Schlueter, and V. Brendel, unpublished data). Users should be aware that the biological relevance of the splice isoforms predicted by these analyses remains to be determined.

Gene Models Supported by Database Matches to Proteins from Other Species

The EGC pipeline generates gapped alignments (DNA Protein Search and Nucleotide Amino Acid Alignment Program; Huang et al., 1997) between the finished Arabidopsis genome sequence and a nonredundant amino acid database. Typically, these alignments show great consistency both with one another and with ab initio gene predictions. In the latest annotation release, 57% of gene models have strong support from non-Arabidopsis protein database matches. Although subsequent cDNA information may reveal minor inaccuracies in gene structure, the predicted proteins encoded by this category of gene models are generally very similar to their homologs from other species. Protein alignments, together with EST and FL-cDNA support, provide strong evidence for gene identification and structure elucidation and are given greater weight than computational gene predictions.

Gene Models with No Database Support

The third category of gene models in the Arabidopsis annotation are those that are based only upon ab initio computer predictions. Under TIGR annotation standards, these are described as "hypothetical" genes or proteins. In previous rounds of annotation involving multiple centers, these have also been described as "putative" or "predicted," but these terms are being systematically replaced to eliminate ambiguity. For a hypothetical gene to be included in our annotation, it must be supported by at least two concordant gene predictions.

How reliable are the structures of these hypothetical genes? At TIGR, a pilot study showed that expression could be detected by PCR for 138 of 169 hypothetical gene models tested (Xiao et al., 2002), suggesting that as many as 80% of these gene predictions could correspond to a novel gene. FL-cDNA sequences were generated for 16 of these genes, with 14 of these entirely consistent with the original gene prediction.

Extending Gene Detection and Validation by Comparative Genomics

Comparison between genome sequences of evolutionarily related species is emerging as a powerful tool for the identification of functionally important regions of a genome (Carlton et al., 2002; Mural et al., 2002). Over the course of evolution, functional regions of the genome tend to be more conserved than nonfunctional regions; thus, local sequence similarity suggests biological functionality. We have begun to use comparative genomics both to validate and to extend our annotation process. By fall of 2002, we had generated over 400,000 sequences from *B. oleracea* whole-genome shotgun libraries accounting for a total of 274 Mb of sequence. The *B. oleracea* sequences were aligned against the Arabidopsis genome using BlastZ (Schwartz et al., 2003), and conserved genomic segments were identified by filtering and collapsing overlapping BlastZ alignments. To facilitate the identification of novel genes, conserved sequences that are colocated and fall into intergenic regions were chained together, and each such chain was assumed to represent a potential gene. Under stringent filtering conditions, this analysis generated over 2,000 potential genes. Manual curation of more than 500 of these resulted in both the creation of new gene models and extension of existing genes. A sample of 200 genes from the total pool was tested for expression using PCR and diverse cDNA libraries. The results indicated that approximately 30% of these genes are expressed. Overall, it appears that *B. oleracea*-Arabidopsis sequence conservation will lead to the identification of a significant number of novel genes.

ANNOTATION OF RNA GENES

Since the first cognate BACs were sequenced, ribosomal RNAs have been annotated by homology, and transfer RNAs have been recognized and accurately annotated using tRNAScan-SE (Lowe and Eddy, 1997). We have confirmed these annotations across the whole genome and assigned AGI identifiers. Other classes of RNAs:snoRNAs (Brown et al., 2001; http://rna.wustl.edu/snoRNAdb), micro-RNAs (Llave et al., 2002), and other non-coding RNAs (MacIntosh et al., 2001) will be incorporated in the near future.

GENOME COMPLETENESS, MISSING GENES, AND UNANCHORED CONTIGS

In terms of DNA sequencing, AGI (2000) measured genome completeness as coverage of the region of the chromosomes extending from either the telomeres or ribosomal DNA repeats to the 180-bp repeats typical of centromeres. Consequently, these regions of highly repetitive DNA, which amount to approximately 10 Mb (Round et al., 1997), remained largely unsequenced. Of more than 20,000 BACs fingerprinted (Mozo et al., 1999), approximately 350 BACs,

which can be assembled into 26 contigs ranging in size from two to 163 BACs per contig, did not anchor to the tiling path constructed from the "complete" genome sequence. Therefore, in collaboration with groups from the University of Chicago, Cold Spring Harbor Laboratories, and the University of North Carolina, we have been survey sequencing BACs selected from each of the unanchored contigs, analyzing them for uniqueness, and mapping them to the genome. The unanchored contigs contain a number of gene rich BACs, two of which (F26J21 and T13I7) have been tentatively mapped to chromosome arms (G.P. Copenhaver, personal communication), suggesting that the present tiling path may require adjustment.

In addition, the quest for a more complete genome extends to locating the genomic origins of cDNAs and ESTs that possess no cognate genomic sequence in the current tiling path. To maintain an up-to-date list of these "missing genes," all cDNAs and ESTs in GenBank are searched periodically against the current version of the genome. cDNAs without a stringent genomic alignment are being systematically investigated to determine whether they exist in the Columbia accession and to identify BACs to which they hybridize. In addition, recent submissions of BAC sequences to GenBank permitted the localization of some previously unmapped cDNAs. Currently, there are 21 cDNAs that cannot be found in the current chromosome and unanchored contig sequences, seven of which have been confirmed by PCR to be present in the Columbia accession (http://www.tigr.org/tdb/e2k1/ath1/missing_genes.shtml). In addition, 2% of approximately 180,000 Arabidopsis ESTs cannot be found in the available genome sequence, but no experimental investigation has been initiated in these cases.

MANUAL CURATION OF GENE STRUCTURE AND FUNCTION

One of the major goals of the reannotation effort is to provide the community with a consistently and uniformly annotated genome. Even as gene prediction programs mature and automated pipelines become more robust, limitations still exist due to the complexity of the biology and the heterogeneity of the data sources that are used; thus, the reannotation effort has relied quite extensively on manual curation.

An important difference between the previous BAC by BAC annotation and this current effort is the availability of the complete proteome, which allows annotation of individual gene products in the context of related genes in the genome. To facilitate this process, the approximately 27,000 gene products in the genome have been organized into protein groupings designed to approximate paralogous families. Paralogous proteins exist due to gene duplications that evolved from a single ancestral gene. TIGR's paralogous family groupings are based on conserved domain composition, taking into account both previously identified domain signatures (Pfam; Sonnhammer et al., 1998) and potential novel domains identified in the Arabidopsis proteome. The current implementation of this family building process has produced a set of 2,780 protein families containing approximately 19,000 proteins (Table II). The grouping of these related proteins enables consistent, uniform annotation and allows better evaluation of the function of predicted gene products.

TIGR employs a versatile computer interface (Annotation Station, Affymetrix, Santa Clara, CA) to visually inspect and modify gene structure. Gene models and the associated computational evidence are presented in a graphical display allowing for easy assessment. This software supports editing of intron-exon boundaries and designation of open reading frames, committing the resulting gene models and tracking information to the annotation database. Computational evidence, including cDNA, EST, and protein alignments, gene model predictions, and domain information, is examined and integrated into existing gene models.

Once gene structures have been computationally and manually validated, gene products are given a descriptive name based on database matches to functionally characterized gene products and protein domains. The annotator is presented with a compact summary of the computational analysis performed on each gene product through MANATEE (http://manatee.sourceforge.net/), a Web-based interface that also supports the addition of functional annotation to the database. To maximize the accuracy and consistency of the naming process, we have developed a set of guidelines based on the annotator's confidence in the computational evidence. If a gene product is identical to an experimentally characterized protein in Arabidopsis, it is named for that protein. If a gene product shares significant sequence similarity with a characterized protein in any species or to a protein domain associated with protein function, the gene product is named based on the implied

Table II. *Paralogous family groupings of Arabidopsis proteins*

The Arabidopsis proteome was clustered into paralogous family groupings using a conservative algorithmic approach. The nos. of proteins and families represented at different family size cutoffs are listed. The five families consisting of more than 100 members include kinases, G proteins, zinc finger proteins, MYB family transcription factors, and cytochrome P450.

Proteins Per Family	No. Of Proteins	No. Of Families
>100	757	5
50–99	2,002	27
20–49	3,353	120
10–19	3,124	236
2–9	9,500	2,392
Total	18,736	2,780

function, with a modifier describing the confidence of the match. The modifiers, in order of confidence, are "putative," "family," and "-related."

A nomenclature standard has also been established for gene products that do not have a good database match to characterized proteins or protein domains. If either cDNA or stringently matched EST evidence supports all or part of the gene model, then the gene product is designated as an "expressed protein." The previous practice of using the term "unknown protein" for gene products supported by EST(s) but lacking database matches has been abandoned. Information regarding cDNA and EST support of named gene products is captured in a public comment field, and the word "expressed" is not appended to the name. If there are no good database matches of any kind, the gene/protein is designated "hypothetical." Note that some proteins designated as "expressed" or "hypothetical" have paralogs in Arabidopsis or protein matches to uncharacterized proteins (hypothetical) in other species. This conservation will also be captured in a public comment field.

In addition to adopting and implementing naming standards, TIGR (in association with TAIR) has begun using the Gene Ontology (GO) to classify Arabidopsis genes. The GO Consortium is an international effort to produce dynamic controlled vocabularies that can be applied across organisms (The Gene Ontology Consortium, 2000; http://www.geneontology.org). GO is used to organize and define gene products based on molecular function, biological process, and cellular component. Ideally, Arabidopsis annotation would include GO assignments for as many proteins as possible, especially because this is the first plant model organism to be fully sequenced. Currently, GO associations are available for approximately 8,000 genes.

CHROMOSOME ASSEMBLY FROM BAC SEQUENCES

One of the goals of TIGR's reannotation effort is to provide continuous genomic sequences representing each of the five Arabidopsis chromosomes, interrupted only by centromeres, and to map all identifiable genes and other features to these sequences. Because both the original annotation data and subsequent curated data are represented at the level of individual BACs, these BACs must be merged at the sequence level to provide representations of whole chromosomes. Once these composite sequences are generated and validated, all annotated features are mapped onto the chromosomes by coordinate transformation, and conflicts and redundancies at BAC overlap regions are resolved.

The original tiling paths for each chromosome were developed and maintained by the individual sequencing groups during the lifetime of the project. We collected this tiling path information from the various Web sites and incorporated the information into our database. We then rigorously examined the quality and extent of sequence similarity in each of the approximately 1,500 BAC overlaps using standard alignment programs. These alignment data were then used to identify the longest region of perfect match within each sequence overlap. Rather than use a traditional "left greedy" approach for chromosome generation, in which the entire sequence of the left BAC is used before switching to the next BAC in the tiling path, we used these high-quality match regions to define the point of transition from one BAC sequence to the next.

Several chromosomes contain regions where it is difficult to reconstruct the tiling path from existing data either because BACs have been trimmed or artificially extended to facilitate annotation. There are 12 junctions where overlaps range from 0 to 6 bp but were reported as being adjacent (http://mips.gsf.de/proj/thal/db/gv/gv_frame.html). Because these junctions cannot be validated by sequence alignment, we are currently using PCR across the junctions to verify their accuracy.

We also examined by PCR approximately 30 cases in which sequence discrepancies between two overlapping BACs led to the existence of different gene models on the two BACs (one often a corrupt version of the other). By sequencing PCR products from genomic DNA across the discordant region, we could determine which version of the sequence should be supported. In every case, we found that PCR supported the sequence that produced the better gene model and, coincidentally, that this same correct sequence was incorporated into the tiling path by the transition selection strategy described above V. Subbu, C. Yu, and C.D. Town, unpublished data).

Independent of sequence consistencies, many gene models in regions of BAC overlaps were either discordant or incomplete. We examined approximately 1,800 gene models in regions of overlaps and made modifications as appropriate to ensure consistency. Genes spanning the junction between two overlapping BACs are modeled partially and merged into complete gene products upon chromosome assembly.

AVAILABILITY OF ARABIDOPSIS REANNOTATION DATA

The TIGR Arabidopsis Annotation Web site (http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml) allows users to access current annotation data, updated on a weekly basis. There are a number of search options, including BAC clone name, original BAC locus, AGI identifier, text-based common name, and sequence search. Once a user narrows down a search to a specific gene, detailed gene and gene product information are displayed on newly designed gene detail pages (Fig. 1). Computational and experimental evidence is shown, and GO assignments are displayed where

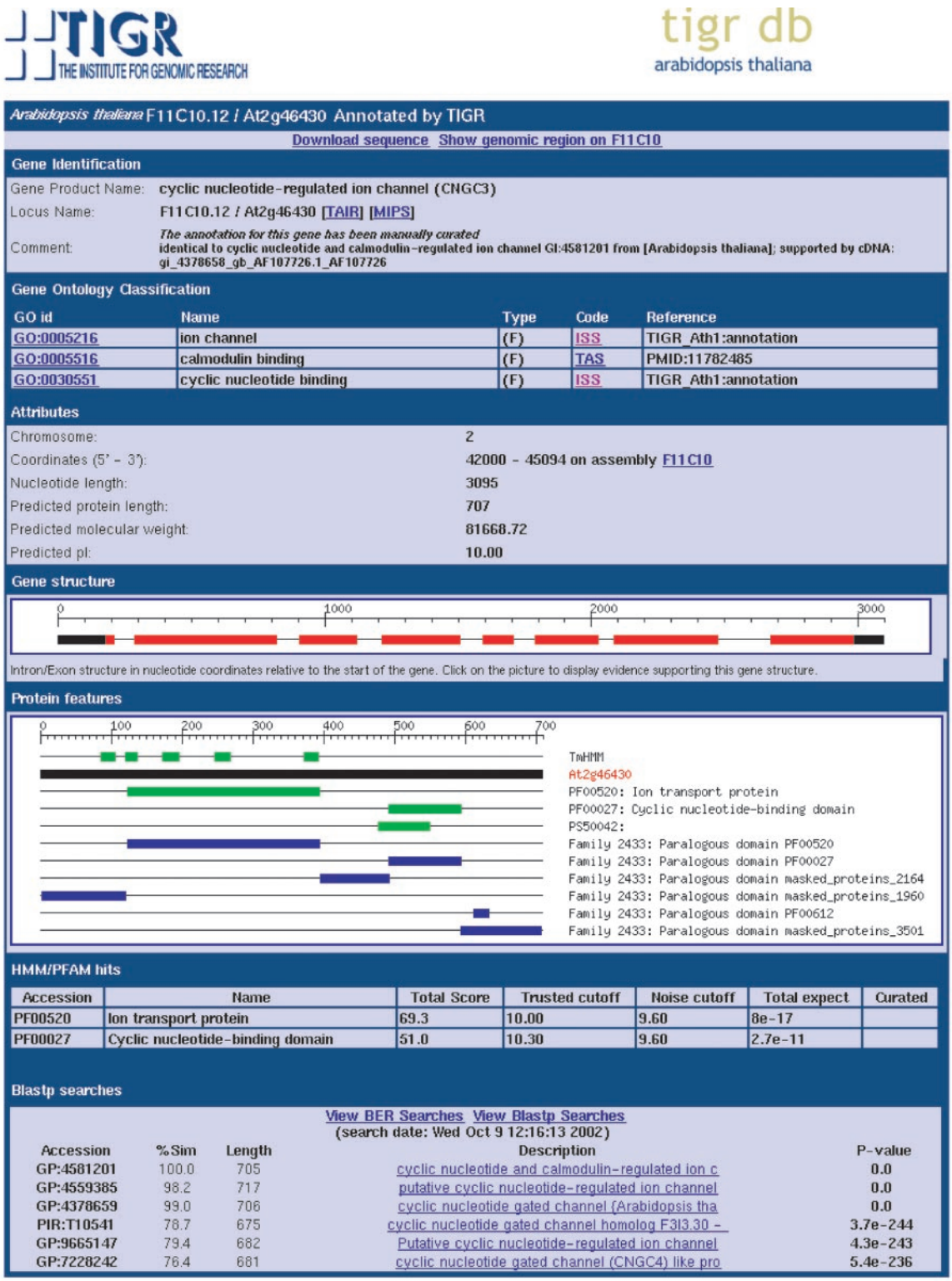

Figure 1. A typical TIGR gene detail page is shown. At the top, locus and the original annotation group are shown. In this section, genomic, gene, and gene product sequence can be obtained, as well as additional information about the BAC genomic region. The "Gene Identification" section contains gene identifiers, gene product name, annotation status, and comments included in the GenBank release. Links to TAIR and MIPS gene annotation are also provided. GO assignments (if made) and GO evidence are shown in the "Gene Ontology Classification" section. Links to the GO graph for assigned terms are also provided. The "Attributes" section shows features of the gene product, and a cartoon representing the gene

(*Legend continues on facing page.*)

available. Users have access to a wide variety of information used in the reannotation process and are able to critically evaluate the resulting annotations. In addition to the annotation pages, the TIGR Web site provides many specialized pages with information on the tiling path and its support, segmental duplications, splicing variants, etc. Summary statistics describing currently available annotation data are presented and compared with the original annotation in Table III.

TIGR generates whole-genome annotation releases approximately twice each year with release 4.0 scheduled for April 2003. These releases involve rebuilding chromosomes from updated tiling paths, mapping current annotation from BACs to chromosome sequences, assigning AGI identifiers to new loci, and validating the entire data set for accuracy and lack of redundancy. Assignment of AGI identifiers (i.e. At5g66770) is coordinated with ongoing curation efforts at MIPS and TAIR (http://mips.gsf.de/proj/thal/db/about/codes.html). These unique and consistent gene identifiers allow users world wide to easily access and track genes. Complete data sets, in XML format, are available at the TIGR ftp site (ftp://ftp.tigr.org/pub/data/a_thaliana/).

There are several different versions of Arabidopsis annotation in public databases around the world. The original BAC-based annotation resides in the Plant division of GenBank (as well as EMBL and DDBJ) and belongs to the sequencing group that generated and submitted it. Because these groups seldom reanalyze or update their annotation, these entries are often outdated and stale. Thus, the National Center for Biotechnology Information hosts third party annotation, including the latest TIGR whole-genome annotation release, in the Genomes Division of GenBank (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html). TIGR data releases are also incorporated into the TAIR database and displayed at their Web site (Rhee SY et al., 2003; http://www.arabidopsis.org). In 2003, the TIGR reannotation effort will conclude, and TAIR will take over the responsibility for annotation updates and data submission to GenBank.

MIPS has been continuing its own annotation effort in parallel to TIGR and also maintains a separate Arabidopsis database, MATDB (Schoof et al., 2002; http://mips.gsf.de/proj/thal/db/).

Table III. *Comparison of Arabidopsis genome statistics*

Summary statistics comparing features of the Arabidopsis genome annotation published in 2000 (Arabidopsis Genome Initiative, 2000) and the present annotation data set are shown.

Feature	October, 2000	February, 2003
Length of sequence in chromosomes 1–5 (Mb)	115.4	119.0
No. of protein-coding genes	25,498	27,384
Gene density (kb gene^{-1})	4.5	4.4
Average gene length (bp)	2,011	2,195
Average peptide length (residues)	434	426
Total no. of exons	132,982	155,190
Average exons per gene	5.2	5.4
Average exon size (bp)	250	276
Average intron size (bp)	168	166

COMMUNITY INPUT

Many user comments containing corrections or suggestions for the improvement of the current annotation are directed to the TAIR curators (curator@arabidopsis.org) because this is a site heavily visited by the research community. Relevant communications are forwarded from TAIR to TIGR. A smaller number are sent directly to TIGR (at@tigr.org). All e-mails are logged into a data management system to track the messages and responses. Last year, approximately 300 e-mails were handled requiring changes to annotation, help with understanding data or data retrieval, as well as general questions. We strongly encourage and welcome continued community input because this further enhances and improves the annotation.

ACKNOWLEDGMENTS

We would like to thank past members of the Arabidopsis annotation group at TIGR for their contributions, the TIGR informatics and information technology staff for their software and hardware support, and members of the Arabidopsis lab group for their experimental contributions. In addition, we would like to thank our colleagues at TAIR and MIPS for many fruitful interactions.

Figure 1. (*Legend continued from facing page.*)
structure is shown in the "Gene Structure" section. Selecting the gene structure cartoon opens a page displaying the evidence supporting the gene structure. In the "Protein Features" section, an image of various types of domain evidence is displayed aligned to the gene product. Domain evidence includes Pfam, Interpro, and Prosite domains, as well as putative Arabidopsis-specific domains, which are generated in-house. Selecting a particular domain opens a page displaying more information about the domain as well as other Arabidopsis gene products that share the domain. Trusted HMM/Pfam domain hits to the gene product are displayed in the HMM/Pfam hits section. Finally, the top Blastp search results are shown in the "Blastp Searches" section. The Blast alignments can be obtained by selecting "View Blastp Searches." In addition to Blast, alignments calculated using Blast Extend Repraze (BER), a modified Smith Waterman algorithm, can also be obtained. Users can access the gene detail page from the Arabidopsis Annotation Web site (http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml) using BAC clone names, original BAC loci, AGI identifiers, text-based gene names, or sequence searches. Because this page is constantly undergoing revisions, changes and/or additions may be made in the future.

Received February 15, 2003; returned for revision March 7, 2003; accepted March 18, 2003.

LITERATURE CITED

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

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res **28:** 45–48

Brown JW, Clark GP, Leader DJ, Simpson CG, Lowe T (2001) Multiple snoRNA gene clusters from Arabidopsis. RNA **7:** 1817–1832

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol **268:** 78–94

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature **419:** 512–519

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996) Splice site prediction in *Arabidopsis thaliana* pre mRNA by combining local and global sequence information. Nucleic Acids Res **24:** 3439–3452

Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL (2002) Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol **3:** reaearch 0029.1–research 0029.12

Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. Genomics **46:** 37–45

Kan Z, Rouchka EC, Gish WR, States DJ (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res **11:** 889–900

Llave C, Kasschau KD, Rector MA, Carrington JC (2002) Endogenous and silencing-associated small RNAs in plants. Plant Cell **14:** 1605–1619

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:** 955–964

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26:** 1107–1115

MacIntosh GC, Wilkerson C, Green PJ (2001) Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. Plant Physiol **127:** 765–776

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S et al. (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat Genet **22:** 271–275

Mural RJ, Adams MD, Myers EW, Smith HO, Gabor Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J et al. (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science **296:** 1661–1671

Pertea M, Lin X, Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res **29:** 1185–1190

Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res **28:** 141–145

Rhee SY, Beavis W, Berardini TZ, 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

Round EK, Flowers SK, Richards EJ (1997) *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. Genome Res **7:** 1045–1053

Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. Genomics **59:** 24–31

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF (2002) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30:** 91–93

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res **13:** 103–107

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res **26:** 320–322

The Gene Ontology Consortium (2000) Gene Ontology: tool for unification of biology. Nat Genet **25:** 25–29

Xiao Y-L, Malik M, Whitelaw CA, Town CD (2002) Cloning and sequencing of cDNAs for hypothetical proteins from chromosome 2 of *Arabidopsis thaliana*. Plant Physiol **130:** 2118–2128