

# Rice Bioinformatics. Analysis of Rice Sequence Data and Leveraging the Data to Other Plant Species<sup>1</sup>

Qiaoping Yuan, John Quackenbush, Razvan Sultana, Mihaela Pertea, Steven L. Salzberg, and C. Robin Buell\*

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

Rice (*Oryza sativa*) is a model species for monocotyledonous plants, especially for members in the grass family. Several attributes such as small genome size, diploid nature, transformability, and establishment of genetic and molecular resources make it a tractable organism for plant biologists. With an estimated genome size of 430 Mb (Arumuganathan and Earle, 1991), it is feasible to obtain the complete genome sequence of rice using current technologies. An international effort has been established and is in the process of sequencing *O. sativa* spp. *japonica* var "Nipponbare" using a bacterial artificial chromosome/P1 artificial chromosome shotgun sequencing strategy. Annotation of the rice genome is performed using prediction-based and homology-based searches to identify genes. Annotation tools such as optimized gene prediction programs are being developed for rice to improve the quality of annotation. Resources are also being developed to leverage the rice genome sequence to partial genome projects such as expressed sequence tag projects, thereby maximizing the output from the rice genome project. To provide a low level of annotation for rice genomic sequences, we have aligned all rice bacterial artificial chromosome/P1 artificial chromosome sequences with The Institute of Genomic Research Gene Indices that are a set of nonredundant transcripts that are generated from nine public plant expressed sequence tag projects (rice, wheat, sorghum, maize, barley, Arabidopsis, tomato, potato, and barrel medic). In addition, we have used data from The Institute of Genomic Research Gene Indices and the Arabidopsis and Rice Genome Projects to identify putative orthologues and paralogues among these nine genomes.

## CURRENT STATUS OF PLANT GENOMICS

The advancement of sequencing technologies within the last decade has been capitalized on by plant biologists. As of October 27, 2000, there are over 102 plant species (among 278 species total) represented in the expressed sequence tag (EST) division (dbEST) of GenBank ([http://www.ncbi.nlm.nih.gov/dbEST/dbEST\\_summary.html](http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)). These various EST projects collectively represent over 835,884 entries (among 6,259,492 total) in GenBank. At the forefront of plant genomics is the model dicotyledonous plant, Arabidopsis. Starting with the Arabidopsis EST project in the early 1990s (Hofte et al., 1993; Newman et al., 1994) and culminating with the first complete plant genome in 2000 (Arabidopsis Genome Initiative, 2000), Arabidopsis has led the plant community in capitalizing on genomic technologies. Of the top 20 organisms in

GenBank, Arabidopsis ranks fourth in the number of total bases of DNA/RNA (<ftp://ncbi.nlm.nih.gov/GenBank/gbrel.txt>; GenBank release 120.0).

With the completion of the Arabidopsis genome, plant biologists will have the opportunity to assess the entire gene complement of a plant for the first time. New avenues of research have begun that will culminate in determining the function of every gene in Arabidopsis (<http://www.nsf.gov/pubs/2001/nsf0113/nsf0113.htm>). Although the complete sequence and the subsequent analyses are an immense achievement in plant biology, Arabidopsis cannot be utilized to address all aspects of plant growth, development, and reproduction. For other plant species that represent diverse physiological and developmental programs, complete genomic sequencing is unlikely to be completed in the foreseeable future. Thus, sequencing of ESTs remains the primary tool for genomic exploration and for functional genomics analyses. The value of EST resources can be greatly enhanced if the data are used to reconstruct a high-fidelity set of nonredundant transcripts such as gene indices (Liang et al., 2000a; Quackenbush et al., 2000; <http://www.tigr.org/tdb/tgi.shtml>). Gene indices are constructed by assembling ESTs after filtering for possible sequence contaminants. This has several advantages over simple clustering approaches: It separates closely related genes into distinct consensus sequences, it separates splice variants, and it produces longer representations of the underlying gene sequences. The resulting tentative consensus sequences (TCs) can be used for eukaryotic genome

---

<sup>1</sup> This work was supported in part by the U.S. Department of Agriculture (grant no. 99-35317-8275 to C.R.B.), by the National Science Foundation (grant no. DBI998282 to C.R.B.), and by the U.S. Department of Energy (grant no. DE-FG02-99ER20357 to C.R.B.). This work was also supported by the U.S. Department of Energy (grant no. DE-FG02-99ER62852 to J.Q.) and by the U.S. National Science Foundation (grant nos. DBI-9983070, DBI-9813392, and DBI-9975866 to J.Q.). J.Q. was also supported in part by the National Science Foundation (grant no. KDI-9980088). S.L.S. and M.P. were supported in part by the National Institutes of Health (grant no. R01-LM06845) and by the National Science Foundation (grant nos. KDI-9980088 and IIS-9902923).

\* Corresponding author; e-mail [rbuell@tigr.org](mailto:rbuell@tigr.org); fax 301-838-0208.

sequence annotation (Lin et al., 1999; Liang et al., 2000b), integration of complex mapping data, and identification of orthologous genes. Following assembly TCs are annotated to provide a provisional functional assignment. A TC containing a known gene is assigned the function of that gene; TCs without assigned functions are searched using DPS, a program to compare DNA with a protein database (Huang et al., 1997) against a nonredundant protein database; high-scoring hits are assigned a putative function. An example TC from the Rice gene index is shown in Figure 1. A summary of the currently available Institute of Genomic Research (TIGR) gene indices can be found in Table I.

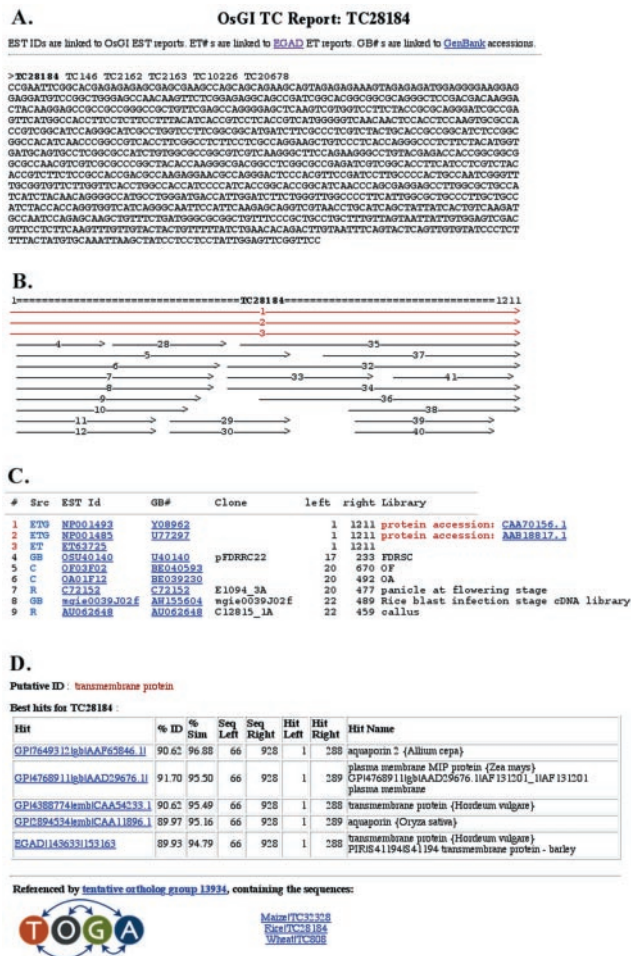
## SELECTION OF RICE (*Oryza sativa*) AS A MODEL SPECIES FOR MONOCOTYLEDONOUS PLANTS

A major classification of plants not represented by Arabidopsis are the monocotyledonous plants. Development of a model monocot species to parallel the achievements in Arabidopsis would have a tremendous impact on plant biology. One of the most important families of monocots is the Gramineae family as it includes major agricultural crop species such as maize, wheat, barley, sugarcane, sorghum, and rice. These grass species share extensive synteny across their genomes, allowing for one species to serve as the base for comparative genomics within the family (Moore et al., 1995).

Due to several factors, rice presents the most tractable species for genomic applications in a cereal. Perhaps the most significant factor in selecting a model species is the small genome size of rice compared with other members of the Gramineae family. The genome size of rice is estimated to be 431 Mb (Arumuganathan and Earle, 1991). This is approximately one-half of the DNA content of sorghum (760 Mb) and 17%, 8.8%, and 2.7% of the DNA content of maize (2,504 Mb), barley (4,873 Mb), and wheat (15,966 Mb), respectively. The reduced genome size in rice compared with these other grass species is attributable to the diploid nature of rice, along with the reduced repetitive DNA content in rice relative to other grass species (Despande and Ranjekar, 1980; Bennetzen et al., 1998). A second criterion in selection of a model species is the availability of genetic and molecular resources. There are currently over 2,200 mapped markers for rice that were generated in a single mapping population of *O. sativa* spp. japonica and *O. sativa* spp. indica (Harushima et al., 1998). The resulting density of approximately 1 marker per 190 kbp provides a deep resource for current high throughput sequencing strategies. Other molecular resources have been developed for a whole genome sequencing project. These include an EST project (Kurata et al., 1994; Yamamoto and Sasaki, 1997), a yeast artificial chromosome library (Umehara et al., 1995), a P1 artificial chromosome (PAC) library (<http://rgp.dna.affrc.go.jp/>), and several bacterial artificial chromosome (BAC) libraries (Wang et al., 1995; Zhang et al., 1996; [http://www.genome.clemson.edu/where/nippon\\_bac/index.html](http://www.genome.clemson.edu/where/nippon_bac/index.html)).

## CURRENT STATUS OF THE INTERNATIONAL RICE GENOME SEQUENCING PROJECT (IRGSP)

The current strategy employed by the IRGSP (<http://rgp.dna.affrc.go.jp/Seqcollab.html>) is a BAC or PAC shotgun sequencing approach. In this approach a minimally overlapping set or tile of clones is identified that is contiguous with the chromosomes and these are sequenced in a high throughput fashion. The BAC/PAC sequences are assembled individ-



**Figure 1.** An example TC from the Rice Gene Index. A, The consensus sequence of a select rice TC (TC28184) is presented in FASTA format. B, The locations of the gene sequences (red) and ESTs (black) that comprise the TC are shown with their respective locations within the assembly. C, Links are provided to GenBank records and to clones available through a variety of sources. D, This entry in the Rice Gene Index encodes a transmembrane protein and links to tentative orthologues in maize and wheat identified from the TIGR Orthologous Gene Alignment (TOGA) are listed at the bottom.

**Table I.** Statistics on the TIGR gene indices

Species	Entries <sup>a</sup>	TIGR Gene Index	Release Date
Arabidopsis (thale cress)	127,890	AtGI 4.0	July 8, 2000
Tomato	90,346	LGI 5.0	August 4, 2000
Soybean	87,504	GmGI 3.0	July 12, 2000
Maize	68,635	ZmGI 4.0	November 14, 2000
Rice	60,166	OsGI 4.0	August 3, 2000
Wheat	44,579	TaGI 1.0	September 18, 2000
Sorghum	44,149	SbGI 1.0	October 10, 2000
Barrel medic	37,560	MtGI 1.0	July 11, 2000
Potato	7,150	StGI 1.0	July 19, 2000
Total no. of sequences	567,979		

<sup>a</sup> The no. of entries for each gene index reflects the no. of ESTs and the no. of expressed transcripts available in Genbank when the index was built. For the Arabidopsis Gene Index, the entries include genes identified through genomic annotation efforts.

ually and the entire chromosome is then assembled from the overlapping BAC sequences. The sequences for each BAC/PAC are then annotated for gene function and the data are released to the public.

Regions of the rice genome are allocated to the IRGSP participants on a chromosomal level. Table II lists the participants of the IRGSP with the corresponding web site for each sequencing center. IRGSP progress can be monitored at several levels. BAC/PAC sequences can be submitted as unfinished sequences to the HTGS division of GenBank at the Phase 1, Phase 2, and Phase 3 level. Phase 1 submissions consist of unordered, unoriented assemblies greater than 2 kbp in length. Phase 2 submissions consist of ordered, oriented assemblies greater than 2 kbp. Phase 3 submissions contain no gaps and may contain annotation. Upon completion, the BAC/PAC sequence is moved to PLN at which time annotation may be added.

Table III lists the basepairs of rice DNA in GenBank as of November 9, 2000. A total of 65,681 entries comprising 28,282,731 bases of sequence have been

submitted from the rice EST sequencing projects. Over 35 Mb of sequence is from the IRGSP and represents BAC/PAC shotgun sequencing efforts. Another 56.9 Mb is in the Genome Sequence Survey division and is derived from BAC end sequencing. Together, with directed rice sequencing efforts by individual labs, rice ranks seventh in total number of bases of DNA/RNA in GenBank, representing the most sequence for any plant species, with the exception of Arabidopsis (<ftp://ncbi.nlm.nih.gov/GenBank/gbrel.txt>; GenBank Release 120.0).

In addition to the public BAC/PAC strategy of the IRGSP, Monsanto has sequenced 3,391 rice BACs; however, these clones were sequenced at a lower coverage than that of the IRGSP and thus are of reduced quality (<http://www.rice-research.org>). From this partial coverage, 259 Mb of assembled sequence data is present in 52,202 contigs. The Monsanto draft of the rice genome is available for basic local alignment searches to academic researchers through a licensing agreement (<http://www.rice-research.org>). As with the public IRGSP approach, Monsanto used *O. sativa*

**Table II.** Participants and chromosomal allocations of the IRGSP

Chromosome <sup>a</sup>	Participant	URL
1, 6, 7, 8	Rice Genome Program (Japan)	<a href="http://rgp.dna.affrc.go.jp/">http://rgp.dna.affrc.go.jp/</a>
1	Korea Rice Genome Research Program	<a href="http://bioserver.myongji.ac.kr/ricemac.html">http://bioserver.myongji.ac.kr/ricemac.html</a>
2	John Innes Centre (UK)	
3, 10	TIGR (USA)	<a href="http://www.tigr.org/tdb/rice">http://www.tigr.org/tdb/rice</a>
3, 10	Clemson University, Cold Spring Harbor Laboratory, Washington University Consortium	<a href="http://www.genome.clemson.edu/projects/rice/chr3_10/index.html">http://www.genome.clemson.edu/projects/rice/chr3_10/index.html</a>
10	Plant Genome Initiative at Rutgers	<a href="http://nucleus.cshl.org/riceweb/clonestatus.htm">http://nucleus.cshl.org/riceweb/clonestatus.htm</a> <a href="http://pgir.rutgers.edu/News.html">http://pgir.rutgers.edu/News.html</a>
4 (indica)	National Center for Gene Research, Chinese Academy of Sciences (China)	<a href="http://www.ncgr.ac.cn/index.html">http://www.ncgr.ac.cn/index.html</a>
5	Rice Genome Project in Republic of China (Taiwan)	<a href="http://genome.sinica.edu.tw/">http://genome.sinica.edu.tw/</a>
9	National Center for Genetic Engineering and Biotechnology (Thailand)	<a href="http://www.cs.ait.ac.th/nstda/biotec/biotec.html">http://www.cs.ait.ac.th/nstda/biotec/biotec.html</a>
9	McGill University (Canada)	
11	University of Wisconsin, Indian Rice Genome Program (India),	
12	Genoscope (France), Universidad Federal de Pelotas (Brazil)	<a href="http://www.genoscope.cns.fr/">http://www.genoscope.cns.fr/</a>
3, 10, 11	All U.S. groups	<a href="http://www.usricegenome.org">http://www.usricegenome.org</a>

<sup>a</sup> A graphical allocation of the chromosomes can be viewed at <http://rgp.dna.affrc.go.jp/rgp/chromosome-share-200010.gif>.

**Table III.** Rice entries in Genbank

Division	Type of Entry	No. of Entries <sup>a</sup>	Total Basepairs <sup>a</sup>
dbEST	ESTs	65,681	28,282,731
Genome sequence survey	BAC end sequences	93,100	56,918,609
High throughput genome sequencing (HTGS)	Phase 1	72	11,218,086
HTGS	Phase 2	32	4,448,878
Plant division of GenBank (PLN)	BACs/PACs (without annotation)	48	5,904,594
PLN	BACs/PACs (with annotation)	100	13,890,936
PLN/others	Other	1,647	2,708,376
All divisions	–	160,680	123,372,210

<sup>a</sup> Rice sequences were downloaded from Genbank on November 9, 2000, and were categorized according to the respective division of Genbank. The ESTs reflect the most recent release of dbEST.

spp. *japonica* variety “Nipponbare,” allowing for integration of the Monsanto partial sequence with the public IRGSP effort. In addition, Monsanto has analyzed the draft sequence for simple sequence repeats, which are invaluable in mapping studies. A file of approximately 7,000 simple sequence repeats with flanking DNA and putative map location (if known) is available to researchers through a free download on the Monsanto web site (<http://www.rice-research.org>).

## COMPUTATIONAL RESOURCES FOR RICE GENOMICS

### Integration of Rice Genetic and Physical Maps

The data from the IRGSP can be utilized even at this early stage. BAC end sequences were used to accelerate the identification of BAC clones that are anchored to the rice genetic map. In Yuan et al. (2000; <http://www.tigr.org/tdb/rice/mappedbacends/>), a series of filtering and search processes were used to align BAC end sequences with rice genetic markers. A total of 418 markers were anchored to the rice physical map that will serve as seed BACs for the

IRGSP and as reagents for positional cloning efforts in rice.

Another important tool in positional cloning is the rapid identification of new reagents. We have implemented an automated process to download rice sequences from GenBank and search against the rice genetic markers. Using a high stringency cutoff we display the alignment of the markers with the BACs/PACs on the TIGR web site (<http://www.tigr.org/tdb/rice/BACmapping/description.shtml>). Through this *in silico* alignment, we have generated a high-resolution map of rice for positional cloning purposes.

### Rice Databases

In a genome project it is imperative that the sequence information is rapidly integrated with available genetic, marker, clone, and other resources, as this will maximize the ability of researchers to utilize the data. This was invaluable in the Arabidopsis Genome Initiative where a single database, initially Arabidopsis Database (*AtDB*) and then The Arabidopsis Information Resource (<http://www.Arabidopsis.org>),

**Table IV.** Additional bioinformatic resources for rice and grass genome projects

Site	URL	Contents
Rice databases		
RiceGenes	<a href="http://ars-genome.cornell.edu/rice/">http://ars-genome.cornell.edu/rice/</a>	Map data, marker data, links
Oryzabase	<a href="http://www.shigen.nig.ac.jp">http://www.shigen.nig.ac.jp</a>	Rice strains, mutants, links
International Rice Research Institute	<a href="http://www.cgiar.org/irri/">http://www.cgiar.org/irri/</a>	Rice strains, mutants
Rice genome information		
Clemson University Genomics Institute	<a href="http://www.genome.clemson.edu/projects/rice/index.html">http://www.genome.clemson.edu/projects/rice/index.html</a>	BAC contig information, BAC end sequence information, BAC clone ordering
Monsanto Rice Genome Sequence Data	<a href="http://www.rice-research.org">http://www.rice-research.org</a>	BLAST server of genome sequencing project, SSRs
Rice Genome Program	<a href="http://rgp.dna.affrc.go.jp">http://rgp.dna.affrc.go.jp</a>	Map data, marker data, sequence data, IRGSP information
TIGR Rice Genome Project	<a href="http://www.tigr.org/tdb/rice">http://www.tigr.org/tdb/rice</a>	Rice Repeat Database, Rice Gene Index, <i>In silico</i> mapping of BAC/PACs to genetic map, GlimmerR, Annotation of rice BACs/PACs via TIGR Gene Indices
Grass genome databases		
USDA-ARS Center for Agricultural Bioinformatics	<a href="http://ars-genome.cornell.edu/grasses.html">http://ars-genome.cornell.edu/grasses.html</a>	Databases on several grasses
Maize DB	<a href="http://www.agron.missouri.edu/">http://www.agron.missouri.edu/</a>	Web site for maize information

served as a central repository for molecular, genetic, clone, and sequence data for Arabidopsis biologists. Likewise, with the generation of a vast amount of rice genomic sequence data, the necessity to integrate rice sequence data with other information from rice genetics, breeding, physiology, and biochemistry is apparent. Several centers have developed databases to integrate rice data from multiple sources (Table IV) and will be working to integrate the sequence data from the IRGSP.

One valuable tool in analyzing rice genomic sequences is the identification of repetitive sequences as it can lead to false associations. To provide a tool to identify repetitive sequences in rice genomic DNA we constructed a rice repeat database using known, curated rice sequences available from GenBank (Yuan et al., 2000; <http://www.tigr.org/tdb/rice/blastsearch.shtml>). This curated database contains satellite DNAs, mobile elements, centromeric repeat sequences, telomeric repeat sequences, and rDNA sequences. In addition, we have identified repetitive sequences using a computational approach using the MUMmer (Delcher et al., 1999) and REPuter (Kurtz and Schleiermacher, 1999) programs. Using available rice BAC end sequences, a total of 8,118 classes of repeats have been identified using this computational approach (N. Volfovsky and S.L. Salzberg, unpublished data). The curated and the MUMmer-identified repeat databases can be searched on the TIGR Rice Genome web site using BLAST and are available for download via anonymous FTP (<http://www.tigr.org/tdb/rice/blastsearch.shtml>).

**Figure 2.** Annotation of a chromosome 10 rice BAC. A chromosome 10 BAC (OSJNBa0051D19) was annotated using output from gene prediction and homology-based searches and a single model (OSJNBa0051D19.18) is displayed. Gene prediction programs used include Genemark.Hmm, Genscan, and Genscan+. Output from the database searches include matches to proteins from the nonredundant amino acid database (prefixed with EGAD or GP) and gene indices (prefixed with TC). For exons that were consistent with the working model, the edges of the exons were highlighted in red. Based on high similarity with the database matches, model OSJNBa0051D19.18 was annotated as a putative alpha-galactosidase.

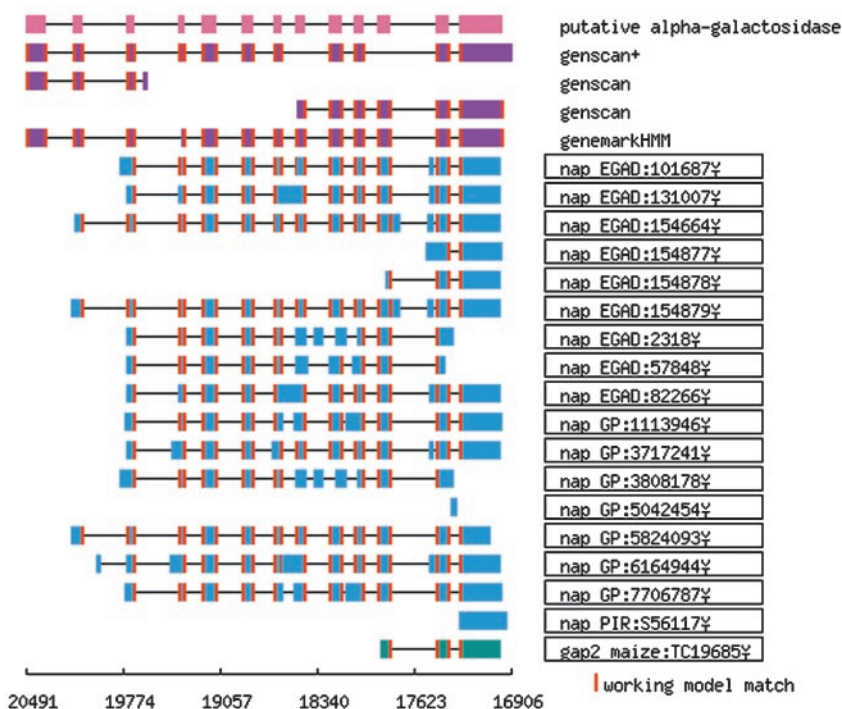
## ANNOTATION OF RICE GENOMIC SEQUENCES: ADDING VALUE TO THE SEQUENCE

### Annotation of Completed BAC/PAC Sequences

Annotation involves adding biological information to DNA sequences and includes identification of genes as well as other miscellaneous features. The annotation system takes a DNA sequence and searches it against a comprehensive protein sequence database to identify similar proteins and against DNA databases such as EST and BAC databases to identify similar gene structures. It also uses ab initio gene finders to identify potentially new genes. The results of these searches are displayed in a graphical genome viewer for a human annotator to analyze. The annotator can modify the model of exon-intron structure for any gene, and can assign a gene name. After careful curation, these data are submitted to public databases (e.g. GenBank, EMBL, and DDBJ) and are displayed on the appropriate web sites. An example of a single-gene model from an annotated rice BAC is shown in Figure 2. The evidence from the database searches and the output from the gene prediction programs are shown along with the final working model constructed by the annotator. This model, OSJNBa0051D19.18, encodes a putative alpha-galactosidase.

### Features of Annotated Rice Genomic Sequences

We have annotated 1.33 Mb of rice genomic sequence from completed BACs on the lower arm of



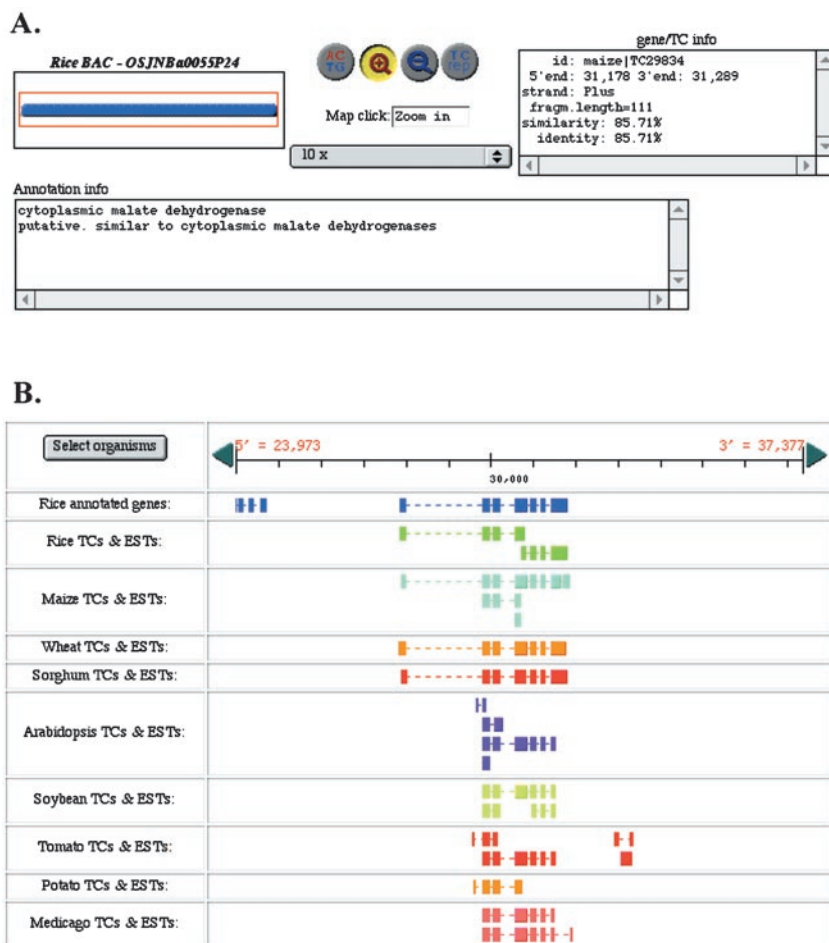
chromosome 10 and thus far we have identified a total of 235 genes. The average rice gene is 2.2 kbp with 3.9 exons and 2.9 introns. The density of rice genes is one gene per 5.7 kbp. The G/C content of rice differs in the coding region (59.1% G/C) from that of the intergenic region (39.1% G/C). We could identify a putative function for 116 of the genes (49.4%), leaving 50.6% of the genes as unknown (having similarity to transcripts with no known function) or hypothetical (having no evidence other than the prediction of a gene finder).

### Preliminary Annotation of Rice BACs

Valuable insight into the gene content of BAC/PACs can be obtained prior to closure of the clone, which is the rate-limiting step in high throughput sequencing. We have developed an automated mechanism to search our unfinished rice BACs (Phases 1 and 2 of HTGS) using available gene prediction and homology-based programs to provide preliminary annotation of rice BACs prior to their completion. The output from the searches is parsed and displayed on the TIGR web page (<http://www.tigr.org/tdb/edb2/osa1/htmls/osa1.html>). An automated mechanism is used to parse the "top hit" from the database

search results and assign a putative name to the working model. This level of preliminary annotation provides researchers a first glimpse of the gene content in rice genome sequences.

Additional interspecies information can be gained by examining the alignment of the EST and gene sequence data in the TIGR Gene Indices with reference to the rice genome. Using all publicly available rice BAC sequence data, including finished, phase 1, and phase 2 sequences in the HTGS division of GenBank, we tabulated the alignment of the TCs and singletons from the nine TIGR Plant Gene Indices with all rice BAC/PAC sequences (<http://www.tigr.org/tdb/ogi/alignTC.html>). These alignments not only provide a low-level functional annotation of all available rice genomic sequence, but also insight into conservation of gene structure across the plant kingdom. An example alignment of an annotated gene on chromosome 10 can be seen in Figure 3. The coding sequence of a putative rice cytoplasmic malate dehydrogenase (OSJNBa0055P24.3) shows significant similarity in gene structure from other plant species as is evident from an alignment between the rice genomic sequence and the various plant TCs. The multiple hits seen in some species may represent paralogues, gene families, alternative



**Figure 3.** Alignment of TCs and singletons from the TIGR Plant Gene Indices with sequences from rice chromosome 10. The rice BAC OSJNBa0055P24 was searched against all of the plant gene indices and matches with >75% identity were displayed graphically. A, Features on the web page include adjustable magnification of the alignments, text display of rice BAC annotation results when available, and text display of gene indices annotation results. B, Alignments from the search of OSJNBa0055P24.3, single gene within a rice BAC, against nine gene indices are shown. Annotation results obtained from the rice BAC sequence are also shown. OSJNBa0055P24.3 encodes a putative cytoplasmic malate dehydrogenase.

splice forms, or partial TC assemblies. However, a clear separation of gene structure is evident between the transcripts identified in the monocot versus the dicot species, as all monocot transcripts and no dicot transcripts contain the more distal 5' exon.

### Improvement of Rice Annotation Tools

One current difficulty with rice annotation is the lack of accurate gene prediction programs. In some instances the gene prediction programs and the homology searches indicate a clear choice for the working model. However, in most cases the gene prediction programs do not agree perfectly with one another and are often in conflict with evidence from sequence homology search results. When similar protein sequences exist, the annotators almost always prefer this evidence over the output of gene finders. However, when sequence homology is very faint or nonexistent, gene prediction programs provide the only available information. Similar to all completed genomes, rice has a substantial number of genes that are hypothetical in that they are predicted solely on the basis of gene prediction programs. Thus, it is imperative that the quality of gene prediction programs be improved for rice.

### GlimmerR, a Rice Gene Finder

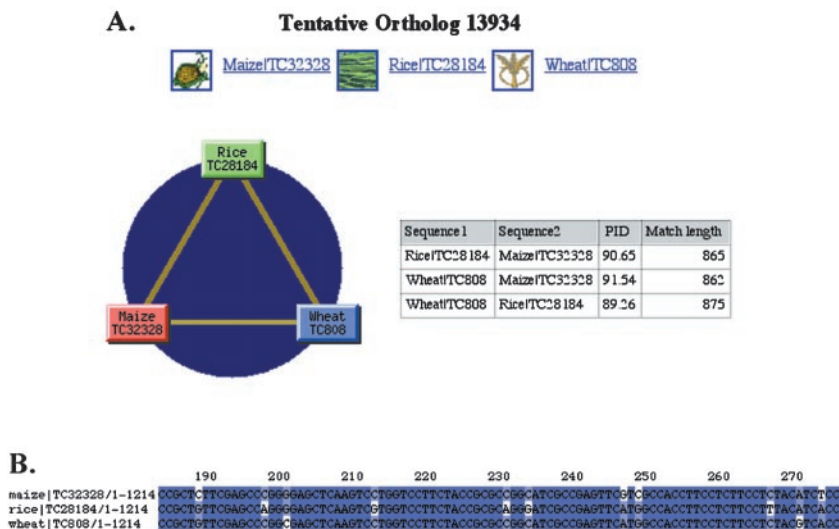
A special version of GlimmerM (Salzberg et al., 1999), known as GlimmerR, was trained to recognize genes in the rice genome. Similar to the original GlimmerM system, which was developed for the malaria parasite *Plasmodium falciparum*, GlimmerR uses interpolated Markov models to score potential coding regions. The score of a second order Markov chain is combined with the score of the splice site, as computed by the maximal dependence decomposition algorithm (Burge and Karlin, 1997). The difference between the coding and non-coding regions on

either side of a splice site is measured by two second order Markov models (using 80-bp windows). These two scores are combined with the maximal dependence decomposition score. False positives are further reduced by only retaining those splice sites whose scores are maximal within a 60-bp window. In addition, GlimmerR has a new dynamic programming algorithm that constructs the gene models. The algorithm prunes the large number of possible exon-intron combinations by scoring the exons and introns of a putative gene with decision trees, constructed using the method of Salzberg et al. (1998). Each gene model is only accepted if the interpolated Markov model score computed over all coding sequences exceed a fixed threshold.

### Training of GlimmerR

Training the system required a set of confirmed genes from rice. The training database was assembled as follows. Thirty-eight complete rice genes were found in GenBank and had Medline references indicating that genomic DNA and cDNA sequences were available. Next, all rice entries with "complete coding regions" in the definition line were downloaded from GenBank and separated into cDNA and non-cDNA groups. To match each cDNA to its corresponding genomic DNA, the two sets of sequences were aligned using DDS (Huang et al., 1997) and manually inspected. These database searches resulted in a total of 159 genes, of which 149 were "complete," meaning that the alignment extended from the start to the stop codon. In addition, 26 rice BAC and PAC sequences were searched against the TIGR plant gene indices. Gene models were selected if they matched at least two exons at >95% identity for rice, or >75% for other plant ESTs. A total of 147 gene models were created from this process: 24 complete coding sequences, 62 partial sequences that included the start or stop codon, and 61 partial coding

**Figure 4.** An example TOG from the TOGA database containing rice, wheat, and maize TCs. Tentative orthologues of the rice TC28184 were identified in the maize and wheat gene indices using the cutoff criteria described in the text. A, The percentage of identity and length of match for all of the pairwise comparisons within the TOG are presented, along with links to the originating gene index. B, A Jalview alignment of the TOG is presented. Alignment of the TOG members can also be viewed using ClustalW.



sequences containing neither start nor stop. In total, the data collection effort from GenBank and BAC/PAC sequences resulted in a set of 172 complete genes that were used for training GlimmerR. The complete set of 133 partial genes plus 172 complete genes was used to construct a training set of confirmed splice sites, which consisted of 1,153 acceptor sites and 1,147 donor sites.

The accuracy of the splice site module on the training set indicates that splice site detection is quite good; for the integrated system, a detection threshold had to be chosen for donor and acceptor sites. Not to miss many true sites, the system was set with a donor site false negative rate of 0.44%, which corresponds to a 6.2% false positive rate and an acceptor site false negative rate of 0.7%, which corresponds to a 9.5% false positive rate. (The false positive rate refers to the percentage of GT/AG dinucleotides that are mistakenly labeled as true splice sites.) When tested on the set of 172 complete genes, GlimmerR's predictions were exactly correct on 100 genes (58%). At the nucleotide level, the sensitivity was 94% (measured as the percentage of true coding bases correctly identified) and specificity was 97% (measured as the percentage of bases labeled as coding that were truly coding). The system exactly predicted 755 exons out of the total number of 921 true exons. One GenBank entry, accession number AF013580, was not included in the training data, but was added later. For this gene, with 3 exons and a total length of 870 bp, GlimmerR predicted all three exons correctly. As sequencing progresses and further genes are found and validated, the system's accuracy can be measured more precisely and new genes will be added to the training set to improve its performance.

#### LEVERAGING THE RICE GENOME SEQUENCE TO OTHER PLANT GENOMES

To provide additional information for functional genomic analysis we have established the TOGA database (<http://www.tigr.org/tdb/toga/toga.shtml>). Homologous genes can be separated into two classes, orthologues and paralogues (Fitch, 1970). Orthologues are homologous genes that perform the same biological function in different species, but have diverged in sequence due to evolutionary separation; paralogues are homologous genes within a species that are the result of a gene duplication event within the lineage. The study of orthologues is of particular importance because it is assumed that these genes play similar developmental or physiological roles and, consequently, should share conserved functional and regulatory domains. For each species included in TOGA, the TCs contained within the respective Gene Indices are compared pairwise. Tentative orthologue groups (TOGs) are identified by requiring reciprocal best hits across three or more species. High-scoring hits that did not meet the re-

ciprocal best hit criteria, but which matched members of existing TOGs were classified as tentative paralogues. Using these criteria we were able to identify 3,074 orthologues from among the eight plant species surveyed. Like the TIGR Gene Indices, TOGA is a relational database that maintains the TOGs as accessionable objects that can be tracked across subsequent releases. TOGs can be searched using a name-based search that allows users to enter a gene name and look for approximate matches or by using a WU-BLAST (Altschul et al., 1990; <http://blast.wustl.edu>) to search the data set.

An example plant TOG can be seen in Figure 4. This TOG (13934) contains rice TC28184 from Figure 1 along with two putative orthologues, one from maize and one from wheat. The high degree of sequence identity between the members of the TOG is apparent in the JALview alignment shown in Figure 4. Although orthologues are properly defined using functional information and protein sequence data, the stringent overlap criteria used to generate the TOGs provide a degree of confidence in the assignment. Further, the progress of the IRGSP, combined with mapping data to be generated by numerous plant EST projects, will provide additional data on syntenic relationships for sequences in the plant TOGA database that will assist in validating the TOG assignments.

#### SUMMARY

Rice has numerous features that make it a model species for the grasses. Its small genome size is compatible with current genomic technologies and sequencing efforts are under way to determine the complete sequence of the rice genome. Tools and resources are being developed to maximally interpret the rice genome sequence. These include improvement of gene prediction programs, expansion of the rice EST and cDNA resources, and identification of molecular resources for mapping. As the data from rice genomics can be leveraged rapidly to other grass species, it is imperative that resources be developed to exploit rice in this manner. Current efforts to extend the rice sequence data to other genomes includes the identification of putative orthologues, alignment of rice BAC/PAC sequences with plant gene indices, and integration of rice sequence data into comparative genetic maps. All of these efforts will continue to be refined as more sequence information is collected. As evidenced by the accomplishments of the Arabidopsis Genome Initiative, knowledge of rice and its close relatives in the grass family will be exponentially increased in the next few years.

#### ACKNOWLEDGMENTS

The authors are indebted to Anna Glodek for database development. The authors also wish to thank Michael



Heaney and Susan Lo for database support, and Vadim Sapiro, Billy Lee, Sonja Gregory, Corey Irwin, Rajeev Kramchedu, Jackie Neubrech, Mark Sengamalay, and Eddie Arnold for computer system support. The authors wish to thank Lowell Umayan, Jeremy Peterson, Hanif Khalak, Patee Gesuwan, and Qi Yang for their informatic support. All the TIGR data and databases described in this article, as well as the GlimmerR software, are freely available from the TIGR web site ([www.tigr.org](http://www.tigr.org)) or upon direct request from the corresponding author.

Received November 16, 2000; accepted December 18, 2000.

#### LITERATURE CITED

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815
- Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* **9**: 208–219
- Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, Avramova Z (1998) Grass genomes. *Proc Natl Acad Sci USA* **95**: 1975–1978
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**: 78–94
- Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL (1999) Alignment of whole genomes. *Nucleic Acids Res* **27**: 2369–2376
- Despande VG, Ranjekar PK (1980) Repetitive DNA in three Gramineae species with low DNA content. *Hoppe-Seyler Z Physiol Chem* **261**: 1223–1233
- Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99–113
- Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin S, Antonio BA, Parco A et al. (1998) A high-density rice genetic linkage map with 2275 markers using a single F<sub>2</sub> population. *Genetics* **148**: 479–494
- Hofte H, Desprez T, Amselem J, Chiapello H, Rouze P, Caboche M, Moisan A, Jourjon MF, Charpentreau JL, Berthomieu P et al. (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J* **4**: 1051–1061
- Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. *Genomics* **46**: 37–45
- Kurata N, Nagamura Y, Yamamoto K, Harushima Y, Sue N, Wu J, Antonio BA, Shomura A, Shimizu T, Lin SY et al. (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. *Nat Genet* **8**: 365–372
- Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**: 426–427
- Liang F, Holt I, Perte G, Karamycheva S, Salzberg SL, Quackenbush J (2000a) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**: 239–240
- Liang F, Holt I, Perte G, Karamycheva S, Salzberg SL, Quackenbush J (2000b) An optimized protocol for analysis of EST sequences. *Nucleic Acids Res* **28**: 3657–3665
- Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**: 761–768
- Moore G, Devos KM, Wang Z, Gale MD (1995) Grasses, line up and form a circle. *Curr Biol* **5**: 737–739
- Newman TC, de Bruijn F, Green P, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E et al. (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol* **106**: 1241–1255
- Quackenbush J, Liang F, Holt I, Perte G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* **28**: 141–145
- Salzberg SL, Delcher AL, Fasman K, Henderson J (1998) A decision tree system for finding genes in DNA. *J Comp Biol* **5**: 667–680
- Salzberg SL, Perte M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**: 24–31
- Umehara Y, Inagaki A, Tanoue H, Yasukochi Y, Nagamura Y, Saji S, Otsuki Y, Fujimura T, Kurata N, Minobe Y (1995) Construction and characterization of a rice YAC library for physical mapping. *Mol Breed* **1**: 79–89
- Wang GL, Holsten TE, Song WY, Wang HP, Ronald PC (1995) Construction of a rice bacterial artificial chromosome library and identification of clones linked to the Xa-21 disease resistance locus. *Plant J* **7**: 525–533
- Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. *Plant Mol Biol* **35**: 135–144
- Yuan Q, Liang F, Zismann V, Hsiao J, Benito MI, Quackenbush J, Wing R, Buell CR (2000) Anchoring of rice BAC clones to the rice genetic map *in silico*. *Nucleic Acids Res* **28**: 3636–3641
- Zhang HB, Choi S, Woo SS, Li Z, Wing RA (1996) Construction and characterization of two rice bacterial artificial chromosome library from the parents of a permanent recombinant inbred mapping population. *Mol Breed* **2**: 11–24