Improvement of whole-genome annotation of cereals through comparative analyses

Wei Zhu and C. Robin Buell

Genome Res. 2007 17: 299-310 originally published online February 6, 2007
Access the most recent version at doi:10.1101/gr.5881807

Letter

Improvement of whole-genome annotation of cereals through comparative analyses

Wei Zhu and C. Robin Buell[1]

The Institute for Genomic Research, Rockville, Maryland 20850, USA

Rice is an important model species for the Poaceae and other monocotyledonous plants. With the availability of a near-complete, finished, and annotated rice genome, we performed genome level comparisons between rice and all plant species in which large genomic or transcriptomic data sets are available to determine the utility of cross-species sequence for structural and functional annotation of the rice genome. Through comparative analyses with four plant genome sequence data sets and transcript assemblies from 185 plant species, we were able to confirm and improve the structural annotation of the rice genome. Support for 38,109 (89.3%) of the total 42,653 nontransposable element-related genes in the rice genome in the form of a rice expressed sequence tag, full-length cDNA, or plant homolog from our comparative analyses could be found. Although the majority of the putative homologs were obtained from Poaceae species, putative homologs were identified in dicotyledonous angiosperms, gymnosperms, and other plants such as algae, moss, and fern. A set of rice genes (7669) lacking a putative homolog was identified which may be lineage-specific genes that evolved after speciation and have a role in species diversity. Improvements to the current rice gene structural annotation could be identified from our comparative alignments and we were able to identify 487 genes which were mostly likely missed in the current rice genome annotation and another 500 genes for structural annotation review. We were able to demonstrate the utility of cross-species comparative alignments in the identification of noncoding sequences and in confirmation of gene nesting in rice.

[Supplemental material is available online at www.genome.org.]

The Poaceae (or grass family) is the most economically important family of plants as the majority of food for human diet or feed food are obtained from species within the family including rice (*Oryza sativa*), maize (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), sorghum (*Sorghum bicolor*), oats (*Avena sativa*), millet (*Eleusine coracana*), and rye (*Secale cereale*) (http://faostat.fao.org). At the genome level, gene content and gene order are well conserved among the Poaceae (Gale and Devos 1998; Goff et al. 2002; Sorrells et al. 2003; The Rice Chromosome 3 Sequencing Consortium 2005), and colinearity at the sequence level (i.e., micro-colinearity) is conserved in spite of gene loss, inversion, duplication, and local genome rearrangements (Bennetzen 2000).

The map-based sequence of the rice genome (*O. sativa* ssp. *japonica* var. Nipponbare) was completed in 2005 (International Rice Genome Sequencing Project 2005) and a draft sequence of the *indica* subspecies (var. 93–11) has also been made available (Yu et al. 2002, 2005). Annotation of the rice genome has been performed by a number of laboratories and consortia (Sakata et al. 2002; Zhao et al. 2004; Ito et al. 2005; Yuan et al. 2005; Ohyanagi et al. 2006). Yuan et al. (2005) reported a structural and functional annotation pipeline in which 43,719 nontransposable element (TE)-related genes were described. This annotation has been updated in which 42,653 non-TE-related genes representing 49,472 gene models have been annotated (Release 4, January 2006, http://rice.tigr.org). Of these, 21,403 have a putative or known function, 6913 were annotated as encoding an expressed protein (transcript support only), and 14,337 annotated as encoding a hypothetical protein. In addition, 13,237 TE-related genes were identified.

As rice is the first finished cereal genome, the rice genome annotation will be used extensively in the annotation of genes in other cereals and grass species. As with other eukaryotic species, annotation of the rice genome was initiated using gene predictions from ab initio gene finders and further improved by using cDNA and expressed sequence tags (ESTs) (Yuan et al. 2005). Currently, more than 30,000 full-length cDNAs (Kikuchi et al. 2003; Xie et al. 2005) and ~1.2 million ESTs are available for rice. Even with the availability of a large transcript sequence data set, a subset of predicted rice genes still lack transcript data support and are derived solely from ab initio gene predictions. Other evidence such as protein similarity and comparative alignments can be used to either support or amend gene models predicted by the ab initio gene finders and has been demonstrated to be useful in genome-wide analysis in metazoan eukaryotes (Ureta-Vidal et al. 2003).

In this study, we performed large-scale comparative genome analyses with rice using all available major plant sequence data sets. Genome sequence data sets include the finished sequence of the model dicotyledonous (dicot) plant, *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), a draft sequence of poplar (*Populus trichocarpa*; Tuskan et al. 2006), a woody dicot perennial, as well as gene-rich genomic sequences of two monocotyledonous (monocot) Poaceae species, maize (Palmer et al. 2003; Whitelaw et al. 2003) and sorghum (Bedell et al. 2005), which were generated using the strategies of methylation filtration (Rabinowicz et al. 1999; Rabinowicz and Bennetzen 2006) and/or high C_0t selection (Peterson et al. 2002; Yuan et al. 2003). In addition, transcript sequence data in the form of ESTs are available from over 400 plant species (9.8 million ESTs in total, dbEST Release dated on 6/26/2006) which provide an estimation of the tran-

[1]Corresponding author.
E-mail rbuell@tigr.org; fax: (301) 838-0208.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5881807. Freely available online through the *Genome Research* Open Access option.

scriptome from a wide range of plant species. To improve the structural annotation of the rice genome and further extend our understanding of sequence conservation among plants and, specifically, the Poaceae, we compared the rice genome and its predicted proteome with the genomes or the genomic sequences of *Arabidopsis*, maize, sorghum, and poplar along with transcript data from 185 plant species. One goal in our study was to improve the current rice genome annotation while the other goal was to build linkages between rice genome annotations and other plant species, especially other cereal species, to facilitate the annotation of those species.

Results

Support for current rice genome annotation

Support in the form of rice transcripts or putative homologs of the 55,890 total rice genes were identified by searching against sequence data sets from 185 plant species which collectively represents 2670 Mb of sequence. The sequence data included (1) genomic sequences from *A. thaliana*, *P. trichocarpa*, *Z. mays*, and *S. bicolor*, (2) the *Arabidopsis* proteome, and (3) 185 plant transcript data sets which are clustered assemblies of ESTs, mRNAs, and full-length cDNAs (Table 1; http://plantta.tigr.org; Childs et al. 2007). The numbers of transcript sequences used in the build of the TIGR Plant Transcript Assemblies (TAs) were variable; of these 185 species, 17 species have more than 100,000 transcript sequences included in the build, which cumulatively represent 72.7% of the total plant transcript sequences (Supplemental Table 1). The Poaceae represent 44% of all of the transcripts in the 185 species TA collection, with rice having more transcript sequences than any other plant species, representing ~16% of all the plant transcripts (Table 1; Supplemental Fig. 1). Clearly, the numbers of putative rice homologs within the Plant TAs will vary based on both the representation of the transcriptome and the evolutionary distance between rice and each species.

In this study, the TA and genomic sequences were placed into 10 groupings based on the type of data source and taxonomic distance relative to rice: (1) rice TA, (2) Other Poaceae TAs (excluding *Oryza sativa*; 23 species), (3) Other Monocot TAs (excluding Poaceae species; eight species), (4) Eudicotyledons (Eudicot) TAs (121 species), (5) Other Plant TAs (32 species, such as basal angiosperms, algae, mosses, and ferns), (6) Assembled *Zea mays* (AZMs) genomic sequences, which are assembled methylation filtration and high C_0t reads from the pilot maize gene enrichment sequencing project (Whitelaw et al. 2003), (7) Assembled *Sorghum bicolor* (ASBs) genomic sequences, which are assembled methylation filtration reads from the sorghum gene enrichment sequencing project (Bedell et al. 2005), (8) near-complete, finished *Arabidopsis* genome sequence, (9) poplar genomic assemblies, and (10) *Arabidopsis*-predicted proteome. Rice genes supported by cognate rice transcripts were identified using the Program to Assemble Spliced Alignments (PASA2) program (Haas et al. 2003). As ~780,000 ESTs have been released since functional annotation of Release 4 of our annotation and Massively Parallel Signature Sequencing (MPSS), Serial Analysis of Gene Expression (SAGE), and proteomic data were utilized in functional annotation of Release 4 models (http://rice.tigr.org), there are some inconsistencies between the function assignment of the gene models in Release 4 and the data presented in this study. For example, hypothetical genes should lack transcript support. However, in this study, we identified 520 (3.6%) hypothetical genes with cognate transcripts due to the recent rice EST release (Table 1), which should be promoted in their annotation to "expressed gene". In this study, only 83.3% of the rice genes annotated with expression support in Release 4 have cognate EST and/or full-length cDNA transcript support, indicating that the remaining 16.7% genes annotated with expression support in Release 4 were obtained through MPSS, SAGE, and peptide evidence data types.

Among the 10 groupings, homologs for rice genes in all gene categories (i.e., known/putative, expressed, hypothetical, and TE-related; Table 1) were most frequently identified within the Other Poaceae TA data sets, which is consistent with previous reports of high sequence identity among the Poaceae (Ware and

Table 1. Plant homologs of rice genes

Data set			No. of sequences	No. of assemblies	Total length ($\times 10^6$ bp)	Percentage of rice homologs[a]			
						Putative	Expressed	Hypothetical	TE-related
Monocot TAs	Poaceae TAs	Rice TA	1,205,038	247,516	158	77.49[b]	83.34	3.63	5.55
		Other Poaceae TAs	2,136,897	740,019	473	99.59	84.83	47.56	93.92
	Other monocot TAs		58,953	36,041	24	87.81	40.85	7.83	59.61
Eudicot TAs			3,567,550	1,245,716	776	98.00	66.48	23.37	81.62
Other plant TAs			663,273	235,470	153	93.57	48.47	9.57	60.62
Arabidopsis genome				5	119	92.07	50.77	11.26	66.00
Poplar genome				22,012	486	94.65	56.66	15.30	72.27
AZM				275,904	316	98.58	77.03	37.23	88.82
ASB				163,908	153	98.45	78.84	41.48	88.63
Arabidopsis proteome				30,690	13	97.12	64.83	20.02	55.25

[a]Rice homologs were determined by searching against the predicted rice proteome (which consists of 21,403 putative genes, 6913 expressed genes, 14,337 hypothetical genes, and 13,237 TE-related genes) using TBLASTN and a cutoff of *E*-value of $<1 \times 10^{-5}$.
[b]For the search with the rice TAs, the percentage of putative, expressed, and hypothetical genes solely reflect high identity alignments to the rice genes using PASA2 and does not include all expression support used in functional annotation of rice genes (MPSS, SAGE, proteomic). Furthermore, ~780,000 rice ESTs have been released since the functional annotation assignments were made in Release 4, and 3.63% (520) of the hypothetical genes now have cognate rice transcripts support. The annotation of these genes will be promoted to "expressed genes" in the next release of our annotation.

Stein 2003). By combining the Other Poaceae TAs with the cereal genomic sequence data sets (i.e., AZM and ASB), more putative homologs for the rice loci could be identified than with any other data set combination. Fewer putative homologs for the known/putative rice genes were identified with the Other Monocot TA data sets compared to the Eudicot TA data sets, which was mainly attributable to the relatively low abundance of the sequence data from non-Poaceae monocots (Table 1; Supplemental Fig. 1). Surprisingly, over 50% of hypothetical genes could be supported by sequence data from the other species, primarily Poaceae sequence data (Table 1). Although we did employ a flexible cutoff of the TBLASTN/BLASTP *E*-value in this study, these data suggest that many of the hypothetical genes encode "real" genes for which cognate transcript evidence in rice is currently lacking.

The prevalence of homologs from diverse clades of the plant kingdom suggests that most of these "core plant genes" may be important housekeeping genes that are not only constitutively expressed and detectable through EST sampling methods but also conserved in function. In contrast, the inability to detect a homolog for 7669 rice genes (including 30 known genes, 895 expressed genes, and 6744 hypothetical genes) in the 2512 Mb of non-rice genomic and transcriptomic sequence available to date suggests the presence of lineage specific genes in rice, which may have evolved after speciation and have a role in species diversity. Alternatively, these, or a subset of these genes, may be artifacts of our annotation methods or encode pseudogenes or transposable elements that we have failed to identify properly.

Distribution of support for rice gene models throughout the plant kingdom

The above analysis suggests that, with the exception of rice itself, Poaceae sequences provide the best support of the rice genes compared to sequence data from the other plant species. However, it is not clear how many of the rice loci are supported solely by Poaceae sequences, non-Poaceae monocots, or other plant species. To identify the breakdown of support, the plant sequence data sets (TAs, genomic assemblies, predicted proteome) were divided into three groups: Poaceae (including 24 Poaceae TAs, AZMs, and ASBs), non-Poaceae monocots (eight non-Poaceae monocot TAs), and all other plant species group (153 plant TAs, genomic sequences of *Arabidopsis* and poplar, and the *Arabidopsis* predicted proteome).

As shown in Table 2, the majority of the rice genes have Poaceae evidence support and only a very small number (i.e., $P^-M^-O^+ + P^-M^+O^- + P^-M^+O^+ = 43 + 0 + 0 = 43$) of rice genes are supported solely by non-Poaceae sequence data. Noticeably, a mere 4544 (10.6%) of the total 42,653 non-TE-related loci have no evidence support under the significance level (*E*-value cutoff $<1 \times 10^{-5}$) used in this study, of which 4384 of these unsupported loci are hypothetical genes. Overall, evidence support could be identified for 69.4% (Total $- P^-M^-O^-$ = 14,337 – 4384 = 9953; Table 2) of the 14,337 hypothetical genes using an *E*-value cutoff of $<1 \times 10^{-5}$ and 2116 loci (= 14,337 – 12,221) had distinct support under the more stringent *E*-value cutoff of $<1 \times 10^{-50}$. All of the hypothetical genes are the result of the prediction of the program FGENESH (Salamov and Solovyev 2000). Preliminary analysis of rice genes using full-length cDNA-supported gene models showed that the accuracy of FGENESH is <82% at the exon level and 45% at the whole-gene level (B. Haas and W. Zhu, unpubl.). While this level of specificity could be improved, the identification of putative homologs of a large percentage of the hypothetical genes in other Poaceae species suggests that the structure of those hypothetical genes could be improved by homologous evidence using similarity-based gene finders (Mathe et al. 2002).

Table 2. Rice homologs in Poaceae, non-Poaceae monocots, and other plant species

Annotation category	*E*-value	$P^-M^-O^-$	$P^-M^-O^+$	$P^-M^+O^-$	$P^-M^+O^+$	$P^+M^-O^-$	$P^+M^-O^+$	$P^+M^+O^-$	$P^+M^+O^+$	Total
Putative	1.00×10^{-5}	9	1	0	0	275	2,323	8	18,787	21,403
	1.00×10^{-10}	42	4	0	0	670	2,938	6	17,743	21,403
	1.00×10^{-20}	212	10	0	0	1,750	3,564	29	15,838	21,403
	1.00×10^{-50}	1,934	12	0	0	3,836	5,379	70	10,172	21,403
	1.00×10^{-100}	6,057	55	0	1	5,934	6,516	41	2,799	21,403
Expressed	1.00×10^{-5}	151	4	0	0	1,957	1,977	8	2,816	6,913
	1.00×10^{-10}	265	1	0	0	2,470	1,754	19	2,404	6,913
	1.00×10^{-20}	585	1	0	0	3,027	1,457	30	1,813	6,913
	1.00×10^{-50}	2,163	1	0	0	2,843	1,066	8	832	6,913
	1.00×10^{-100}	4,020	8	0	0	2,012	723	11	139	6,913
Hypothetical	1.00×10^{-5}	4,384	38	0	0	6,090	2,702	25	1,098	14,337
	1.00×10^{-10}	6,520	12	0	0	5,705	1,448	21	631	14,337
	1.00×10^{-20}	9,080	8	0	0	4,345	636	11	257	14,337
	1.00×10^{-50}	12,212	9	0	0	1,902	172	2	40	14,337
	1.00×10^{-100}	13,575	1	0	0	693	67	0	1	14,337
TE-related	1.00×10^{-5}	112	0	0	0	2,024	3,211	0	7,890	13,237
	1.00×10^{-10}	413	3	0	0	3,283	2,505	4	7,029	13,237
	1.00×10^{-20}	1,075	3	0	0	4,008	2,255	0	5,896	13,237
	1.00×10^{-50}	2,774	13	0	1	4,353	2,714	4	3,378	13,237
	1.00×10^{-100}	5,308	10	0	0	3,558	4,082	4	275	13,237
All	1.00×10^{-5}	4,656	43	0	0	10,346	10,213	41	30,591	55,890
	1.00×10^{-10}	7,240	20	0	0	12,128	8,645	50	27,807	55,890
	1.00×10^{-20}	10,952	22	0	0	13,130	7,912	70	23,804	55,890
	1.00×10^{-50}	19,083	35	0	1	12,934	9,331	84	14,422	55,890
	1.00×10^{-100}	28,960	74	0	1	12,197	11,388	56	3,214	55,890

Rice homologs were determined using variable *E*-value cutoffs and binned into eight bins based on evidence support from the Poaceae (P), non-Poaceae monocots (M), and other plant species (O) sequence datasets. For example, the column "$P^+M^-O^+$" represents the number of the rice homologs found in the Poaceae (P) and the other plant species (O) data sets but not in the non-Poaceae monocots (M) data set under the specified *E*-value cutoff. Similarly, the column "$P^-M^-O^-$" shows the number of rice genes which have no homologs in any plant species.

Frequency of homologs in gene–enriched genomic versus transcript Poaceae sequences

The above analyses indicated that Poaceae sequence data are a valuable resource for annotating the rice genome. The Poaceae data set contains 24 TAs, AZMs, and ASBs. Although it is well known that transcript sequence data are the most important resource in the gene identification, we were interested in ascertaining the contribution of the Poaceae (excluding rice) TAs and maize/sorghum genomic sequences relative to the rice TAs in providing support for genome annotation. Not surprisingly, the rice TA data set yielded the best contribution among the three major Poaceae data resources (Table 3). Interestingly, the non-rice Poaceae TAs had a comparable number of rice homologs as the maize and sorghum genome assemblies, suggesting broad representation of the Poaceae transcriptome in the collective Poaceae TA data set. Indeed, ~93.8% [(19,699 + 369)/21,403] of the known/putative rice genes have a potential homolog in both the non-rice Poaceae TAs and AZM/ASB sequences at a high significance level (*E*-value cutoff of $<1 \times 10^{-20}$). AZMs and ASBs have homologs in 92.4% (19,780) and 89.1% (19,070) of the known/putative rice genes, respectively. Over 98% (21,071) coverage would be reached if the significance was lowered to 1×10^{-5}, consistent with reports that the gene-rich sequencing strategy provides significant coverage of the maize and sorghum gene space (Palmer et al. 2003; Whitelaw et al. 2003; Bedell et al. 2005). Certainly, the homologs with lower *E*-value are more likely to offer better support in the gene structure identification. Using the significance *E*-value of $<1 \times 10^{-100}$, as shown in Table 3, a total of 1224 (i.e., $R^-P^-G^+ + R^-P^+G^- + R^-P^+G^+$ = 414 + 374 + 336) known/putative rice genes are supported not by a rice TA sequence but by sequences from the other Poaceae species, suggesting that cross-species comparative analysis could provide additional support for this subset of known genes.

Overall, 90,039 (32.6%) of the total 275,904 AZMs had BLASTZ alignments with the rice genome. Of these, 51,403 AZMs have representative alignments that cover 54.5 Mb of the rice genome. Similarly, 65,233 (39.8%) of the total 163,908 ASBs had BLASTZ alignments to the rice genome with representative alignments from 39,885 ASBs spanning 69 Mb of the rice genome. In total, the genic regions of over two thirds of the total rice genes were covered at least partially by the AZM and ASB genomic alignments, including 31,690 (74%) non-TE-related genes.

Noncognate transcripts

Comparative analyses can be applied not only across species but also within the species. In Release 4 of the TIGR rice genome annotation, only the best hit of each rice transcript sequence in the rice genome with ≥95% sequence identity and ≥90% coverage was used in the annotation process to ensure that only the cognate transcript was associated with its respective gene. A large portion of rice genes originated from the large-scale segmental duplication (or polyploidization) that occurred in the rice genome about 70 MYA (Paterson et al. 2004; Wang et al. 2005a), and a recent study showed that 95% of the introns have been conserved among segmentally duplicated rice genes (Lin et al. 2006). These noncognate rice transcripts provide a valuable resource that can be exploited to improve the structural annotation of paralogous genes. Using an *E*-value cutoff of $<1 \times 10^{-5}$, approximately one-quarter [$R^+P^-G^-$/(Total − $R^-P^-G^-$) = 2395/(14,337 – 4422) = 24.2%] of the hypothetical rice genes supported by Poaceae data were supported by rice transcripts only (Table 3). Despite the 3.6% (520) hypothetical rice genes with cognate transcripts from the recent rice EST release, the remain-

Table 3. Rice homologs in the rice TA, non-rice Poaceae TAs, and cereal genomic sequences

Annotation category	*E*-value	$R^-P^-G^-$	$R^-P^-G^+$	$R^-P^+G^-$	$R^-P^+G^+$	$R^+P^-G^-$	$R^+P^-G^+$	$R^+P^+G^-$	$R^+P^+G^+$	Total
Putative	1×10^{-5}	10	26	8	96	28	24	59	21,152	21,403
	1×10^{-10}	46	41	8	210	85	66	126	20,821	21,403
	1×10^{-20}	222	164	59	369	256	172	462	19,699	21,403
	1×10^{-50}	1,946	271	231	476	1,066	432	2,412	14,569	21,403
	1×10^{-100}	6,113	414	374	336	2,349	1,165	4,063	6,589	21,403
Expressed	1×10^{-5}	155	20	16	142	770	104	227	5,479	6,913
	1×10^{-10}	266	32	23	168	1,116	133	252	4,923	6,913
	1×10^{-20}	586	37	44	223	1,396	157	374	4,096	6,913
	1×10^{-50}	2,164	62	94	163	1,348	202	815	2,065	6,913
	1×10^{-100}	4,028	85	80	61	1,150	231	652	626	6,913
Hypothetical	1×10^{-5}	4,422	324	349	1,295	2,395	378	658	4,516	14,337
	1×10^{-10}	6,532	341	244	750	2,560	345	575	2,990	14,337
	1×10^{-20}	9,088	292	134	389	2,078	255	448	1,653	14,337
	1×10^{-50}	12,221	158	69	105	1,135	199	157	293	14,337
	1×10^{-100}	13,576	65	21	12	561	64	19	19	14,337
TE-related	1×10^{-5}	112	38	14	243	519	136	622	11,553	13,237
	1×10^{-10}	416	58	42	249	889	142	596	10,845	13,237
	1×10^{-20}	1,078	122	38	286	1,300	388	177	9,848	13,237
	1×10^{-50}	2,788	454	44	246	1,630	1,289	591	6,195	13,237
	1×10^{-100}	5,318	642	108	199	2,040	1,906	49	2,975	13,237
All	1×10^{-5}	4,699	408	387	1,776	3,712	642	1,566	42,700	55,890
	1×10^{-10}	7,260	472	317	1,377	4,650	686	1,549	39,579	55,890
	1×10^{-20}	10,974	615	275	1,267	5,030	972	1,461	35,296	55,890
	1×10^{-50}	19,119	945	438	990	5,179	2,122	3,975	23,122	55,890
	1×10^{-100}	29,035	1,206	583	608	6,100	3,366	4,783	10,209	55,890

The information displayed in this table is in the same format as Table 2. Rice homologs were binned into eight bins based on evidence support from the rice TA (R), non-rice Poaceae TAs (P), and AZM/ASB (G) data sets. For example, the column "$R^+P^-G^+$" represents the number of the rice homologs found in the rice TA (R), the genomic sequences from maize or sorghum (G) data sets, but not the non-rice Poaceae TAs (P) data set under the specified *E*-value cutoff.

ing 20.6% (2395 – 520 = 1875 $R^+P^-G^-$) rice hypothetical genes can be supported by the transcripts from putative rice paralogs.

Improvement of gene prediction using comparative alignments

The rice genes in Release 4 were identified on the basis of either coding potential using the program FGENESH and/or the spliced alignment of the cognate transcripts by PASA2 (Yuan et al. 2005), and thus those rice genes with limited transcript data or low coding potential were not likely to be identified in the process of the genome annotation. Use of genomic comparisons between rice and other plant genomic sequences can contribute to the identification of newly identified genes in conserved intergenic regions and could be utilized to improve the structure of those rice genes lacking cognate rice transcript data, especially hypothetical genes. The refinement could be achieved in two nonexclusive ways: by generating gene predictions using the similarity-based program like TWINSCAN (Korf et al. 2001), N-SCAN (Gross and Brent 2006), AUGUSTUS+ (Stanke et al. 2006), and EUGENE'HOM (Foissac et al. 2003), or by manual inspection using a genome annotation tool. We utilized cross-species spliced and genomic alignments to improve our structural annotation as well as to identify "unannotated" genes.

Cross-species spliced alignments can be used to corroborate predicted gene structures (Fig. 1) and amend gene predictions (Brendel et al. 2004). A total of 217,269 filtered cross-species alignments were assembled into 40,935 assemblies using PASA2 (Haas et al. 2003). PASA2 was utilized to compare the assemblies with the gene models, and 25,268 (61.7%) alignments could be successfully incorporated into 15,654 distinct rice genes. Among the 15,654 rice genes supported by these transcript assemblies, 12,548 are known/putative genes, 2460 are expressed genes, 314 are hypothetical genes, and 332 are TE-related genes. While these data strongly support our annotation, many of the assemblies were not completely consistent with the existing gene models.

As exon-intron boundaries defined by spliced alignments of heterologous transcripts are not as reliable as those by cognate transcripts, more stringent criteria were employed to refine our analysis. First, a putative exon had to be supported by at least three alignments. As a result, 75,111 putative exons were predicted. Second, we compared these cross-species putative exons with existing exons within our annotation. To avoid confounding our results due to alternative splicing and UTR exons, we focused on "novel" exons in genic regions, which did not overlap with existing exons in Release 4, i.e., "novel" exons in annotated intronic regions and in which the annotated intron is not supported by any transcript (rice or heterologous). A total of 500 genes (including 395 known genes, 66 expressed genes, and 39 hypothetical genes) with potential new exons were identified through cross-species alignments in which the exon had to be supported by at least three independent alignments. For 477 of the 500 genes with new exons, cross-species alignments were from more than one species. Manual inspection showed that most of these genes have incorrect gene structures (see Figs. 2, 3; Supplemental Figs. 2, 3), suggesting that known genes and expressed genes can be improved through comparative analyses. In addition to refinement of gene structures, comparison using PASA2 identified 1854 assemblies, which supported unannotated or "missed" genes. Using a stringent set of criteria (≥300 bp in length and ≥3 exons), we conservatively identified 388 assemblies located in 255 distinct intergenic regions as candidate unannotated genes.

Using BLASTZ alignments of the AZMs and ASBs to the rice genome, many alignments between rice genes and putative homologs were identified that spanned multiple exons. Not surprisingly, the identity of the alignment in intronic regions was significantly lower than the flanking exonic regions, appearing as a banded pattern in the genome browser display (Fig. 1). For example, maize, sorghum, and even *Arabidopsis* genomic comparisons indicated a potential gene upstream of LOC_Os04g45820 that was not predicted by FGENESH and in which only short cognate rice EST sequences and two cross-species spliced alignments are available (Supplemental Fig. 4). By combining the partial gene structure provided by rice EST spliced alignments, exon patterns in the genomic alignments with AZM5_17958, AZM_5_84956, ASB44489, ASB71162 and ASB45539, and cross-species spliced alignments from wheat and maize, we can construct a gene model consistent with the gene prediction from TWINSCAN (Korf et al. 2001). In addition, the last exon of the new gene model could be further extended in the 3′-UTR exon region as indicated by the rice transcript assembly TA26377_4530.

Genomic comparisons can also indicate the existence of novel genes. Using the AZMs and ASBs, numerous BLASTZ alignments were located in "intergenic" regions, which may lead to the identification of the unannotated genes. Each continuous intergenic region was regarded as one unit in the analysis to simplify the computation (which may contain more than one gene). In total, there were 1145 and 830 intergenic regions over 1000 bp length containing alignments with AZMs and ASBs, respectively. Overall, 1614 distinct intergenic regions were covered and 361 of them were covered by matches from both an AZM and an ASB sequence. The conserved regions were then searched against the TIGR *Oryza* Repeat and the UniProt databases, resulting in 493 and 339 non-TE-related conserved intergenic sequences identified from maize and sorghum, respectively. Even when the significance was increased to an *E*-value cutoff of $<1 \times 10^{-50}$, there were still 291 and 175 potentially new genes identified from maize and sorghum, respectively, which could be merged into 324 distinct intergenic regions. Further analyses showed that many of those regions encode genes not contained in the current rice genome annotation (data not shown). As our filtering criteria were stringent in that they required similarity to annotated proteins, other conserved regions may also encode genes which have not been previously identified. Indeed, by removing the filter of UniProt similarity yet retaining the repetitive sequence filter, we identified 800 additional candidate new genes. Some conserved regions may contain multiple genes (Supplemental Fig. 5), while others may contain coding regions of the neighboring genes missed in the annotation process and not new genes. Nevertheless, these conserved "intergenic regions" can be used to improve the current rice genome annotation.

Conserved noncoding regions

The analyses described above concentrated on the identification of protein coding regions; however, conservation is not restricted within protein coding regions but also exists in noncoding regions. In addition to the regular protein coding genes, there are numerous non-protein coding RNA (ncRNA) molecules encoded in the genome that have been found in eukaryotes, eubacteria, archaebacteria, and viruses (Eddy 2001; Dennis and Omer 2005; Liu et al. 2005). ncRNAs can be classified into different groups

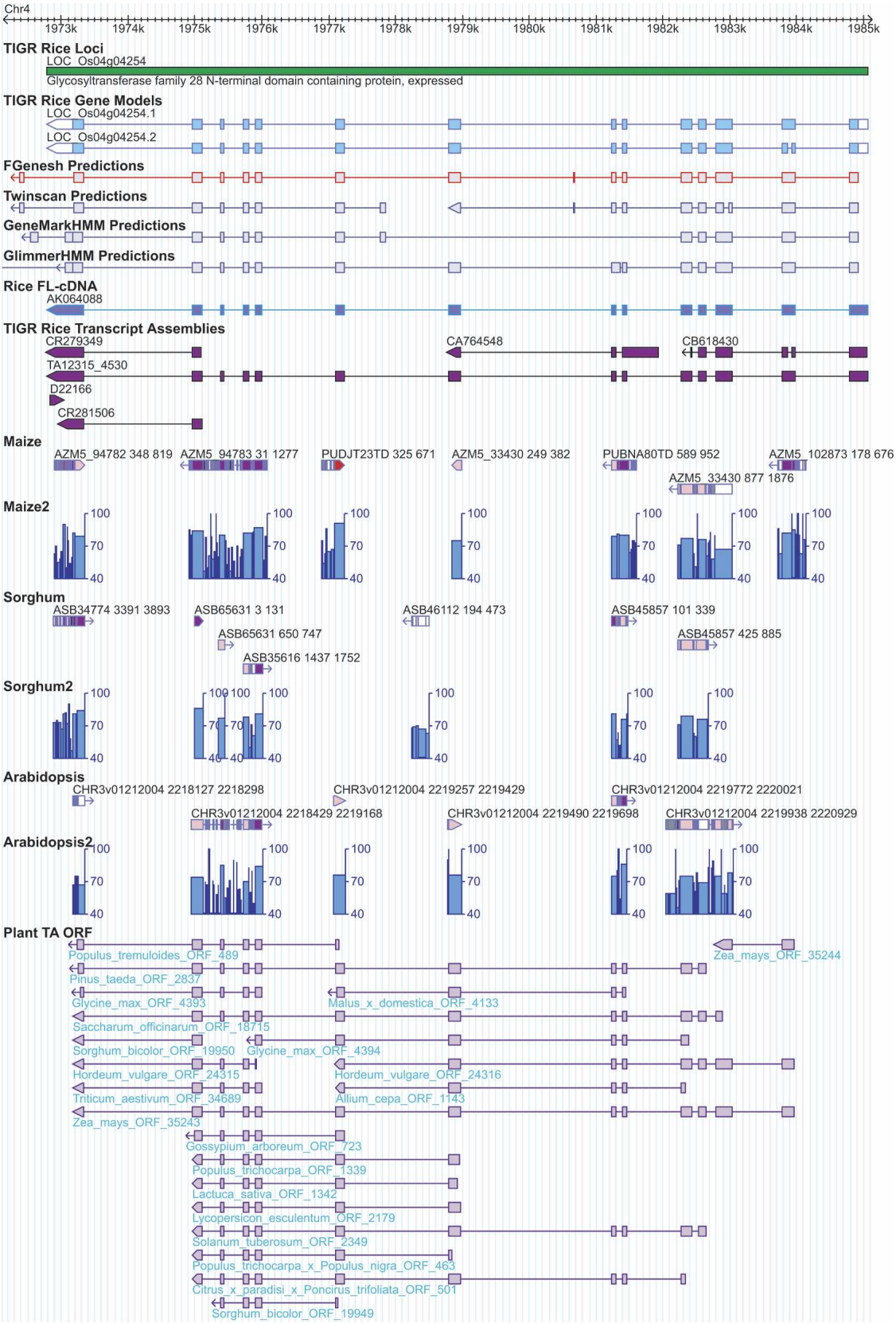

Figure 1. (Legend on next page)

such as transfer RNA (tRNA), small nucleolar RNAs (snoRNA), ribosomal RNAs (rRNA), and microRNAs (miRNA). miRNAs are a recently identified type of ncRNA, a short (~21 nucleotides) single-stranded RNA excised from a long self-complementary precursor (Bartel 2004). A majority of miRNAs are conserved between *Arabidopsis* and rice (Reinhart et al. 2002; Sunkar and Zhu 2004; Sunkar et al. 2005; Jones-Rhoades et al. 2006; Zhang et al. 2006) and, in this study, we examined whether conserved regions identified among the cereal genomes could be ascribed to ncRNA genes. We downloaded all 153 rice miRNAs from the miRBase Sequence Database (Release 8.1; ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/genomes/osa.gff). Ninety-eight of the 153 miRNAs were found in the alignments between rice and the AZMs while 85 were detected in alignments between rice and the ASBs. Of these, 73 were commonly shared among rice, maize, and sorghum. Furthermore, 13 out of the 153 miRNAs could be found in the alignments between *Arabidopsis* and rice, and 12 were shared with either maize or sorghum. Intriguingly, nine miRNAs were conserved between the three cereal genomes and the *Arabidopsis* genome.

Although our study utilized simple genomic alignments and did not employ algorithms dedicated to finding miRNAs (Adai et al. 2005; Wang et al. 2005b; Berezikov et al. 2006), we were able to demonstrate that miRNAs loci are more conserved compared to their neighboring regions (Fig. 4A). Indeed, sequence conservation was observed in the miRNA target sites between rice, sorghum, and *Arabidopsis*. Six target sites of the miRNA family miR399 were found in the 5′-UTR region of LOC_Os05g48390 and four target sites were found in the 5′-UTR region of the *Arabidopsis* orthologous gene At2g33770 (Supplemental Fig. 6; see also Fig. 4 in Sunkar and Zhu 2004). In addition to the conservation examined in the miRNA genes and miRNA target sites, conservation was also observed for other ncRNAs such as tRNAs and snoRNAs (Fig. 4B,C).

Discussion

Our analyses show that comparative analyses are extremely useful in the annotation of the rice genome even when more than one million rice transcript sequences are available. Furthermore, we show that the completed rice genome sequence and its annotation provide a valuable data resource for genomic research in other grass species and will certainly facilitate the ongoing maize genome annotation or play an even more important role for those cereal species with only limited sequence data such as oat and rye.

Through our comparative analyses, we were able to identify 255 and 324 unannotated candidate genes which were missed in Release 4, by cross-species spliced alignments and genomic comparison, respectively, of which, 92 were found by both methods. In total, 487 distinct candidates were identified. Further analysis showed that, although there are FGENESH predictions in 350 (72%) of these conserved "intergenic regions", in most cases, the FGENESH algorithm predicted a single, long gene model that spanned two valid neighboring genes with an intron in a relatively short intergenic region, i.e., a merged model. As full-length cDNAs are available to support one gene in the merged FGENESH model, the long FGENESH prediction is truncated by the PASA2 program which heavily weights full-length cDNA evidence over ab initio gene finder output. Consequently, the other exons within the merged FGENESH model that lack cDNA support are deleted and not included in the final model or gene set. Of the remaining 137 unannotated gene candidates, 43 (31%) likely originate from organellar insertions (data not shown; Supplemental Fig. 5). This analysis suggests that a modified update strategy for the PASA2 program to capture the deprecated portion of merged FGENESH models, coupled with integration of an organellar gene finder into our annotation pipeline, should undercover a majority (80.7%) of these two classes of missed genes.

The BLASTZ alignments between rice and maize or sorghum were able to span short introns and clear, distinct alignments were apparent; however, these alignments might be split by long introns. Clearly, these genomic comparisons are able to reveal gene structures, although it may be still difficult for a curator to determine the exact exon-intron boundaries without additional information. To address this problem, spliced alignments from paralogous and heterologous transcripts could be employed to identify the exact exon-intron boundaries. It was shown that 25,258 (61.7%) of the cross-species spliced alignment assemblies can be incorporated into the genes annotated in Release 4. Many assemblies may reveal the right gene structure (Figs. 2, 3; Supplemental Figs. 2, 3); however, most of them are problematic due to low sequence similarity or gene structure alternation subsequent to speciation. Therefore, additional filters are needed to improve the quality of the spliced alignments. For example, establishing a

Figure 1. Example of a rice gene (LOC_Os04g04254) supported by additional evidence derived from the comparative analyses. The green bar represents the TIGR rice locus with the locus name *above* and the putative function assignment *below* the track. There are two gene models (i.e., alternative splicing isoforms) predicted for this locus as shown in the track "TIGR Rice Gene Models," in which exons are represented by rectangles and introns by the intervening horizontal thin lines. The coding regions of the gene model are in light blue, distinguished from the untranslated regions, which are in white. The genes prediction program output (FGENESH, TWINSCAN, GeneMark.hmm, and GlimmerHMM) and the spliced alignments of the rice cognate full-length cDNAs and TAs are displayed in a similar manner as the gene models. *Above* the transcript alignment tracks are the GenBank accession numbers or TA identifiers to indicate the sequence data source. *Below* the transcript alignment tracks are the BLASTZ alignments between rice and maize, sorghum, *Arabidopsis*. Each BLASTZ alignment is composed of one or more gap-free blocks interspersed with gaps (or insertion/deletions). Each block is represented by a rectangle with the color for the average sequence identity within the block: red color (>90% identity), purple (80%–90% identity), pink (70%–80% identity), light gray (60%–70% identity), gray (50%–60% identity), and black (<50% identity). The gaps (not introns) between the flanking blocks are represented by the horizontal thin line. Each BLASTZ alignment feature is also labeled with the sequence identifier, which is followed by two numbers marking the start and end points of the match in the assembly. For example, ASB assembly ASB45857 has two matches with the genic region of the rice gene LOC_Os04g04254 which are separated by a 1-kb intron in the rice genome. The labels "ASB45857 101 339" and "ASB45857 425 885" indicate that the two matched regions have a gap of 85 (= 425 – 339 – 1) bp in the sorghum genome. The tracks "Maize2," "Sorghum2," and "Arabidopsis2" displayed the exact same BLASTZ alignment data but in alternative manners: The sequence identity is represented by the height of the column and the block size is represented by the column width. The cross species spliced alignments are displayed in the last track "Plant TA ORF" (for details, see Methods). As shown in the figure, the exonic regions have higher sequence identity than the neighboring intronic regions in the BLASTZ alignments. The gene structure of the gene LOC_Os4g04254 is well supported by rice full-length cDNA and rice TAs as well as the TAs from other plant species such as apple (*Malus x domestica*), barley (*Hordeum vulgare*), loblolly pine (*Pinus taeda*), lettuce (*Lactuca sativa*), maize (*Zea mays*), onion (*Allium cepa*), soybean (*Glycine max*), and sugarcane (*Saccharum officinarum*).

Figure 2. Gene structure of the expressed gene LOC_Os12g10200 corrected by the cross-species spliced alignments. Symbols are as in Fig. 1. Several rice transcripts cover a portion of the expressed gene LOC_Os12g10200 which were stitched into the FGENESEH prediction to create the existing gene model by the program PASA2. The cross-species spliced alignments suggest that two exons have been missed.

requirement that each exon-intron boundary in cross-species spliced alignment assemblies be supported by at least three or more alignments may permit more automated incorporation of cross-species alignments data into an annotation pipeline.

We also show that genomic comparisons can shed light on the evolution of gene structure and organization. Some rice genes are intervened by a short intergenic region and synteny of not only gene order but also intergenic regions, which can be seen with rice, maize, and sorghum (data not shown). However, it is unclear whether the conservation of short intergenic regions has a biological function role. Alternative splicing is a common feature in plants (Wang and Brendel 2006), and conservation of alternative splicing isoforms across species may indicate that some of the alternative splicing events have biological significance and therefore are preserved after species divergence. For example, the rice gene LOC_Os07g43950 has a dominant exon-skipping isoform and the skipped exon was conserved in both maize and sorghum. Intriguingly, cross-species alignments indicate that the same alternative splicing event may occur in maize and sugarcane (Supplemental Fig. 7), consistent with the report that 25% of human alternatively skipped exons are alternatively spliced in mouse (Sorek et al. 2004).

Comparative analyses can also be applied to the study of the transposable elements. Some mutator-like transposable elements (MULEs), for example, can capture fragments from host gene and are referred to as Pack-MULEs (Jiang et al. 2004). Recent studies reported that there are several thousand Pack-MULEs in the rice genome (Jiang et al. 2004; Juretic et al. 2005). There is an interesting example of gene nesting which is likely caused by a MULE. Cross-species alignments confirmed that the gene structure supported by the full-length cDNA AK059758 is likely to be accurate, which contains an intron ~5 kb in length (Supplemental Fig. 8). Rice EST evidence clearly indicates that there is a gene encoded within the long intron, suggestive of gene nesting. The alignments of the ASB assembly ASB37 indicate that the rice gene encoded by the AK059758 is conserved in sorghum while the length of the corresponding intron is only 100 bp in sorghum (Supplemental Fig. 8). Additional information from rice repetitive sequences and trans-duplicated MULEs (Juretic et al. 2005)

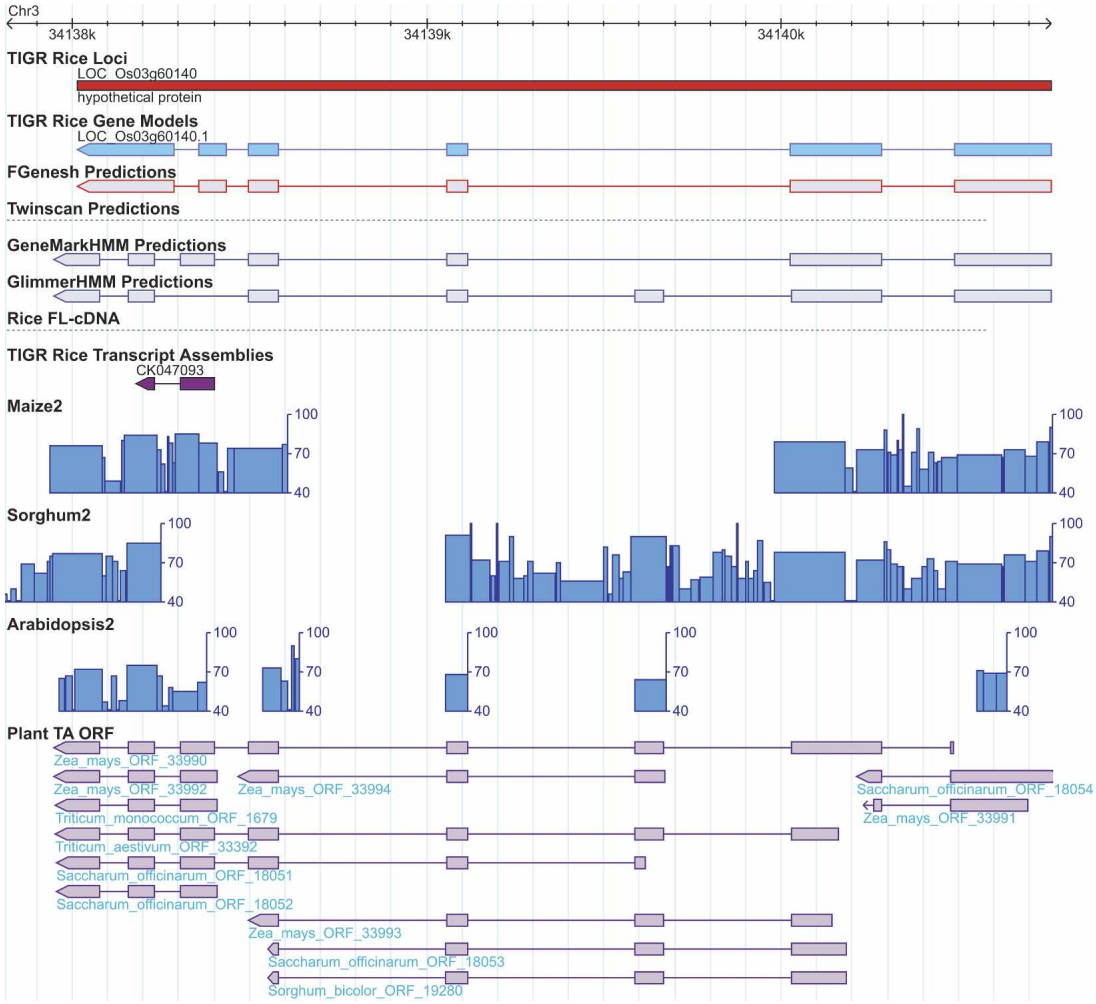

Figure 3. Gene structure of the hypothetical gene LOC_Os03g60140 corrected by the cross-species spliced alignments. Symbols are as in Fig. 1. It appears that there is an exon missed in the second intron of LOC_Os03g60140 and the correct gene structure is shown via cross-species alignments but not correct by any gene finders.

suggest that this particular case of the nested gene was likely to be mediated by the mutator-like elements.

In this study, we have shown the value of comparative alignments in improving structural and functional annotation of the rice genome, which can be attributed in large part to the deep representation of genomic and transcriptomic sequence for the Poaceae. Clearly, the depth of sequence data is not evenly distributed among taxa in the plant kingdom, and increased efforts in sequencing non-Poaceae monocots may shed light not only on the evolution of the Poaceae genome but also on the divergence of monocots from eudicots.

Methods

TIGR rice genome annotation

Release 4.0 of the TIGR rice genome annotation (available at http://rice.tigr.org/; Yuan et al. 2005) was utilized in this study in which genes were identified using the ab initio gene finder FGENESH and amended based on rice transcript evidence (full-length cDNAs and ESTs) with PASA2 (Haas et al. 2003). Transposable element-related genes in Release 4 were identified as described in Yuan et al. (2005). Putative, known, and hypothetical genes were annotated as described previously (Yuan et al. 2005). Genes were annotated as encoding expressed proteins if the gene lacked protein support but had expression support from MPSS, SAGE, EST/full-length cDNA, or proteomic data sets.

Other plant genomes

Maize genomic assemblies (Release 5.0), derived from methylation filtration and high C_0t reads and termed AZMs were downloaded from TIGR (ftp://ftp.tigr.org/pub/data/MAIZE/AZMs/release_5.0; Chan et al. 2006). There are 275,904 assemblies in the AZM data set (Table 1), with lengths ranging from 65 bp to 16,340 bp. Over 90% of AZM assemblies are shorter than 2200 bp. The sorghum genomic sequences derived from methylation filtration reads (Bedell et al. 2005) were downloaded from GenBank and assembled into ASBs in a similar way as the maize assemblies at TIGR (available at ftp://ftp.tigr.org/pub/data/MAIZE/Sorghum_assembly/ASB.gz). The length distribution of ASBs is similar to that of AZMs. The *Populus trichocarpa* assembled scaffolds (Release 1.0) were downloaded from the Joint Genome Institute (JGI, ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/assembly/v1.0/poplar.masked.fasta.gz). The *Arabidopsis* pseudo-

Figure 4. Conserved noncoding RNAs: (*A*) miRNA, (*B*) snoRNA, and (*C*) tRNA. (*A*) The precursor miRNA transcript osa-miR394 is represented by the red bar with the arrow for the orientation and the mature miRNA sequence is highlighted in yellow color which is significantly more conserved in maize, sorghum, and *Arabidopsis* than the neighboring intergenic or intronic regions. The snoRNA (*B*) and tRNA (*C*) features are represented by arrowed bars in orange and blue, respectively.

molecules and the genome annotation (Release 6) were downloaded from TAIR (ftp://ftp.arabidopsis.org/home/tair).

Plant transcript assemblies

Transcript assemblies (TAs; Release 1.0; 185 total) were downloaded from the Web site of the TIGR plant transcript assembly (ftp://ftp.tigr.org/pub/data/plantta/update_08152005), where transcript assemblies were constructed for all plant species with >1000 ESTs in NCBI dbEST (as of August 15, 2005). The rice proteome was searched against the non-rice plant genome sequence data sets using TBLASTN.

Cross-species spliced alignments

To annotate rice gene structures using the Plant TA sequences, all Plant TAs (except rice) were aligned to the TIGR Release 4 pseudomolecules using the program *GeneSeqer* (Usuka et al. 2000; Brendel et al. 2004). In order to improve the efficiency, the source code of the program *GeneSeqer* was modified and a new command-line option for the minimum coverage was inserted so that the optimal spliced alignment would be generated only when the high-scoring segment pairs span a specified length coverage of the transcript sequence (Brendel et al. 2004). In addition to the default *GeneSeqer* setting, we empirically set the minimum coverage (i.e., the new command option) at 40%. Due to the low sequence similarity between divergent species, the exon-intron boundaries defined by those cross-species spliced alignments are likely to be partially, if not completely, erroneous. Further complicating the interpretation is the possibility that there may have been an alteration of the gene structure after speciation and/or gene duplication. Therefore, three stringent criteria were applied to remove regions of poor quality from spliced alignments: (1) The matched rice genome segments had to contain a long, open reading frame (≥150 bp), (2) the open reading frame had to contain at least one intron, and (3) all intron boundaries had to contain the canonical splices sites GT/AG. The qualified regions were further assembled using the PASA2 program (Haas et al. 2003) and compared to the current rice genome annotation.

Genomic comparisons

The TIGR Release 4 pseudomolecules were aligned with the genomic assemblies of maize, sorghum, *Arabidopsis*, and poplar using the program BLASTZ (Schwartz et al. 2003). The BLASTZ options "H = 2200 C = 2" were employed for maize and sorghum, whereas slightly different options "H = 2200 C = 0" were used for *Arabidopsis* and poplar. A simple algorithm was applied to identify "representative alignments" for each species. For each species, by sorting the BLASTZ alignment scores in descending order, if the rice genomic region lacked any BLASTZ alignments, the alignment was selected as the "representative alignment" and no further alignments were allowed in this region. If an alignment was already "assigned" to this genomic region, the alignment was skipped. Therefore, each selected alignment represents the best alignment for a specified rice genomic region with no overlap permitted among the representative alignments from the same species. Those representative alignments were displayed as feature tracks in the TIGR Rice Genome Browser (http://www.tigr.org/tigr-scripts/osa1_web/gbrowse/rice/; also see Fig. 1).

Unannotated genes

To identify "intergenic" regions that might represent candidate unannotated genes, a lower cutoff for the match length between cross-species alignments was empirically set to 1 kb and the putative conserved intergenic regions were further searched using the BLAST package against the TIGR *Oryza* Repeat Database (Ouyang and Buell 2004) and the UniProt database (http://www.ebi.uniprot.org/index.shtml). The conserved regions lacking a significant match in the TIGR *Oryza* Repeat Database but having a significant match with an entry in the UniProt database were selected as candidates for newly identified genes. We also searched for unannotated genes using cross-species spliced alignments. We conservatively selected each spliced alignment in the intergenic regions with at least 300 bp length and three exons as an unannotated gene candidate.

Data availability

All alignments are viewable on the TIGR Rice Genome Browser (http://www.tigr.org/tigr-scripts/osa1_web/gbrowse/rice/) through selection of the appropriate tracks. Additional data sets are made available through supplemental files associated with the on-line version of this manuscript. Supplemental File 1 contains a list of potential new genes identified in intergenic regions using comparative alignments. Supplemental File 2 contains a list of potential new exons within gene models identified through cross-species spliced alignments. Supplemental File 3 lists the loci that are supported or unsupported by our comparative analysis. These are binned into putative/known, expressed, hypothetical, and transposable-element related. They are further separated into two sets: those with rice TA support and those without rice TA support. We have provided a set of files of interest for download at ftp://ftp.tigr.org/pub/data/rice/GENOME_2006_058818/, including a FASTA formatted file of the CDS and proteins of the new exons added to genes.

Acknowledgments

We thank members of the rice annotation team at TIGR for critical comments on the manuscript, and B. Haas for technical assistance on the configuration of the PASA2 pipeline. This work was supported by a National Science Foundation Plant Genome Research Program grant to C.R.B. (DBI-0321538).

References

Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V., and Sundaresan, V. 2005. Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res.* **15:** 78–91.

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

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bedell, J.A., Budiman, M.A., Nunberg, A., Citek, R.W., Robbins, D., Jones, J., Flick, E., Rholfing, T., Fries, J., Bradford, K., et al. 2005. Sorghum genome sequencing by methylation filtration. *PLoS Biol.* **3:** e13.

Bennetzen, J.L. 2000. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* **12:** 1021–1029.

Berezikov, E., Cuppen, E., and Plasterk, R.H. 2006. Approaches to microRNA discovery. *Nat. Genet.* **38 (Suppl. 1):** S2–S7.

Brendel, V., Xing, L., and Zhu, W. 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* **20:** 1157–1169.

Chan, A.P., Pertea, G., Cheung, F., Lee, D., Zheng, L., Whitelaw, C., Pontaroli, A.C., SanMiguel, P., Yuan, Y., Bennetzen, J., et al. 2006. The TIGR Maize Database. *Nucleic Acids Res.* **34:** D771–D776.

Childs, K.L., Hamilton, J., Zhu, W., Ly, E., Cheung, F., Hank, W., Rabinowicz, P.D., Town, C.D., Buell, C.R., and Chan, A.P. 2007. The TIGR Plant Transcript Assemblies Database. *Nucleic Acids Res.* **35 (Database issue):** D846–D851.

Dennis, P.P. and Omer, A. 2005. Small non-coding RNAs in Archaea. *Curr. Opin. Microbiol.* **8:** 685–694.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.

Foissac, S., Bardou, P., Moisan, A., Cros, M.J., and Schiex, T. 2003.

EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **31:** 3742–3745.
Gale, M.D. and Devos, K.M. 1998. Comparative genetics in the grasses. *Proc. Natl. Acad. Sci.* **95:** 1971–1974.
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.
Gross, S.S. and Brent, M.R. 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13:** 379–393.
Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr., R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31:** 5654–5666.
International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793–800.
Ito, Y., Arikawa, K., Antonio, B.A., Ohta, I., Naito, S., Mukai, Y., Shimano, A., Masukawa, M., Shibata, M., Yamamoto, M., et al. 2005. Rice Annotation Database (RAD): A contig-oriented database for map-based rice genomics. *Nucleic Acids Res.* **33:** D651–D655.
Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569–573.
Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. 2006. MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* **57:** 19–53.
Juretic, N., Hoen, D.R., Huynh, M.L., Harrison, P.M., and Bureau, T.E. 2005. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **15:** 1292–1297.
Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., et al. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301:** 376–379.
Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 (Suppl. 1):** S140–S148.
Lin, H., Zhu, W., Silva, J.C., Gu, X., and Buell, C.R. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol.* **7:** R41.
Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., and Chen, R. 2005. NONCODE: An integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* **33:** D112–D115.
Mathe, C., Sagot, M.F., Schiex, T., and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30:** 4103–4117.
Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., Fujii, Y., Antonio, B.A., Nagamura, Y., Imanishi, T., et al. 2006. The Rice Annotation Project Database (RAP-DB): Hub for *Oryza sativa* ssp. *japonica* genome information. *Nucleic Acids Res.* **34:** D741–D744.
Ouyang, S. and Buell, C.R. 2004. The TIGR Plant Repeat Databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* **32:** D360–D363.
Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M., Martienssen, R.A., and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. *Science* **302:** 2115–2117.
Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.* **101:** 9903–9908.
Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., and Paterson, A.H. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12:** 795–807.
Rabinowicz, P.D. and Bennetzen, J.L. 2006. The maize genome as a model for efficient sequence analysis of large plant genomes. *Curr. Opin. Plant Biol.* **9:** 149–156.
Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R., and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23:** 305–308.
Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & Dev.* **16:** 1616–1626.
The Rice Chromosome 3 Sequencing Consortium. 2005. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. *Genome Res.* **15:** 1284–1291.
Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., et al. 2002. RiceGAAS: An automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30:** 98–102.
Salamov, A.A. and Solovyev, V.V. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10:** 516–522.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20:** 68–71.
Sorrells, M.E., La Rota, M., Bermudez-Kandianis, C.E., Greene, R.A., Kantety, R., Munkvold, J.D., Miftahudin, Mahmoud, A., Ma, X., Gustafson, P.J., et al. 2003. Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.* **13:** 1818–1827.
Stanke, M., Schoffmann, O., Morgenstern, B., and Waack, S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7:** 62.
Sunkar, R. and Zhu, J.K. 2004. Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *Plant Cell* **16:** 2001–2019.
Sunkar, R., Girke, T., Jain, P.K., and Zhu, J.K. 2005. Cloning and characterization of microRNAs from rice. *Plant Cell* **17:** 1397–1411.
Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596–1604.
Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251–262.
Usuka, J., Zhu, W., and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16:** 203–211.
Wang, B.B. and Brendel, V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci.* **103:** 7175–7180.
Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. 2005a. Duplication and DNA segmental loss in the rice genome: Implications for diploidization. *New Phytol.* **165:** 937–946.
Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., and Li, Y. 2005b. MicroRNA identification based on sequence and structure alignment. *Bioinformatics* **21:** 3610–3614.
Ware, D. and Stein, L. 2003. Comparison of genes among cereals. *Curr. Opin. Plant Biol.* **6:** 121–127.
Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302:** 2118–2120.
Xie, K., Zhang, J., Xiang, Y., Feng, Q., Han, B., Chu, Z., Wang, S., Zhang, Q., and Xiong, L. 2005. Isolation and annotation of 10828 putative full length cDNAs from indica rice. *Sci. China C Life Sci.* **48:** 445–451.
Yu, J., Hu, S., Wang, J., Wong, G.K., 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.
Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., et al. 2005. The Genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3:** e38.
Yuan, Y., SanMiguel, P.J., and Bennetzen, J.L. 2003. High-Cot sequence analysis of the maize genome. *Plant J.* **34:** 249–255.
Yuan, Q., Ouyang, S., Wang, A., Zhu, W., Maiti, R., Lin, H., Hamilton, J., Haas, B., Sultana, R., Cheung, F., et al. 2005. The Institute for Genomic Research Osa1 rice genome annotation database. *Plant Physiol.* **138:** 18–26.
Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. 2006. Conservation and divergence of plant microRNA genes. *Plant J.* **46:** 243–259.
Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., et al. 2004. BGI-RIS: An integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.* **32:** D377–D382.

Received August 23, 2006; accepted in revised form December 20, 2006.