Orthologous Comparisons of the *Hd1* Region across Genera Reveal *Hd1* Gene Lability within Diploid *Oryza* Species and Disruptions to Microsynteny in Sorghum

Abhijit Sanyal,†,[1] Jetty S.S. Ammiraju,†,[2] Fei Lu,†,[3] Yeisoo Yu,[2] Teri Rambo,[2] Jennifer Currie,[2] Kristi Kollura,[2] Hye-Ran Kim,[2] Jinfeng Chen,[3] Jianxin Ma,[1] Phillip San Miguel,[1] Chen Mingsheng,*,[3] Rod A. Wing,*,[2] and Scott A. Jackson*,[1]

[1]Department of Agronomy, Purdue University

[2]Department of Plant Sciences, BIO5 Institute, Arizona Genomics Institute, University of Arizona, Tucson

[3]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

†These authors contributed equally to the study.

***Corresponding author:** E-mail: mschen@genetics.ac.cn; rwing@ag.arizona.edu; sjackson@purdue.edu.

Associate editor: Naruya Saitou

Abstract

Heading date is one of the most important quantitative traits responsible for the domestication of rice. We compared a 155-kb reference segment of the *Oryza sativa* ssp. *japonica* cv. Nipponbare genome surrounding *Hd1*, a major heading date gene in rice, with orthologous regions from nine diploid *Oryza* species that diverged over a relatively short time frame (~16 My) to study sequence evolution around a domestication locus. The orthologous *Hd1* region from *Sorghum bicolor* was included to compare and contrast the evolution in a more distant relative of rice. Consistent with other observations at the *adh1/adh2*, *monoculm1*, and *sh2/a1* loci in grass species, we found high gene colinearity in the *Hd1* region amidst size differences that were lineage specific and long terminal repeat retrotransposon driven. Unexpectedly, the *Hd1* gene was deleted in *O. glaberrima*, whereas the *O. rufipogon* and *O. punctata* copies had degenerative mutations, suggesting that other heading date loci might compensate for the loss or nonfunctionality of *Hd1* in these species. Compared with the *japonica Hd1* region, the orthologous region in sorghum exhibited micro-rearrangements including gene translocations, seven additional genes, and a gene triplication and truncation event predating the divergence from *Oryza*.

Key words: comparative genomics, microsynteny, genome evolution, transposons.

Introduction

Over the past two decades, comparative genome studies in plants have led to a better understanding of genome evolution, by gaining insights into genome organization, syntenic relationships to other plant genomes, and the mechanisms that are involved in the maintenance or disruption of micro- and macrolevel synteny (reviewed in Bennetzen and Ma 2003; Kellogg and Bennetzen 2004; Bennetzen 2007). These studies have focused primarily on a few widely divergent genomes (rice, barley, maize, wheat, sorghum) that diverged 50–60 My. Because of the long divergence time, it has been difficult to thoroughly characterize the chronology of the mechanisms leading to genome evolution as well as to clearly explain the unexpected patterns of lineage-specific and/or chromosomal region–specific structural changes (Bennetzen 2002, 2007). With the rapid advancement in whole-genome high-throughput sequencing technologies (Mardis 2008; Shendure and Ji 2008) there has been an explosion of sequencing projects (Harismendy et al. 2009; Varshney et al. 2009). An emerging trend is the comparison of closely related species to streamline the tracking of genome architectural changes within species phylogeny (Ma and Bennetzen 2004; Ammiraju et al. 2007, 2008; Grover et al. 2007, 2008; Wang et al. 2008; Lu et al. 2009).

The genus *Oryza* is an exceptional system for the study of the evolutionary history of recently diverged species and fine mapping of structural rearrangements during species radiation from a common ancestor. *Oryza* consists of 23 wild and two domesticated species that represent six diploid (A, B, C, E, F, and G) and four allopolyploid (BC, CD, HJ, and HK) genome types with a pantropical distribution, a reasonably well-resolved phylogeny (Ge et al. 1999; Zou et al. 2008) and a number of rapid genome diversification events associated with speciation occurring over a relatively short time period (~14 My) (Aggarwal et al. 1997; Khush 1997; Ge et al. 1999; Vaughan et al. 2003). In addition, a wealth of genomic resources such as bacterial artificial chromosome (BAC) libraries (Chen et al. 2002; Ammiraju et al. 2006), linkage maps (Harushima et al. 1998; Cheng, Buell, et al. 2001; Cheng, Presting, 2001; Chen et al. 2002; Kao et al. 2006), microarray gene chips for expression studies (Rensink and Buell 2005), and a genome sequence of the widely cultivated species *Oryza sativa* (Sasaki and Burr 2000; International Rice Genome Sequencing Project 2005) make *Oryza* an ideal platform for

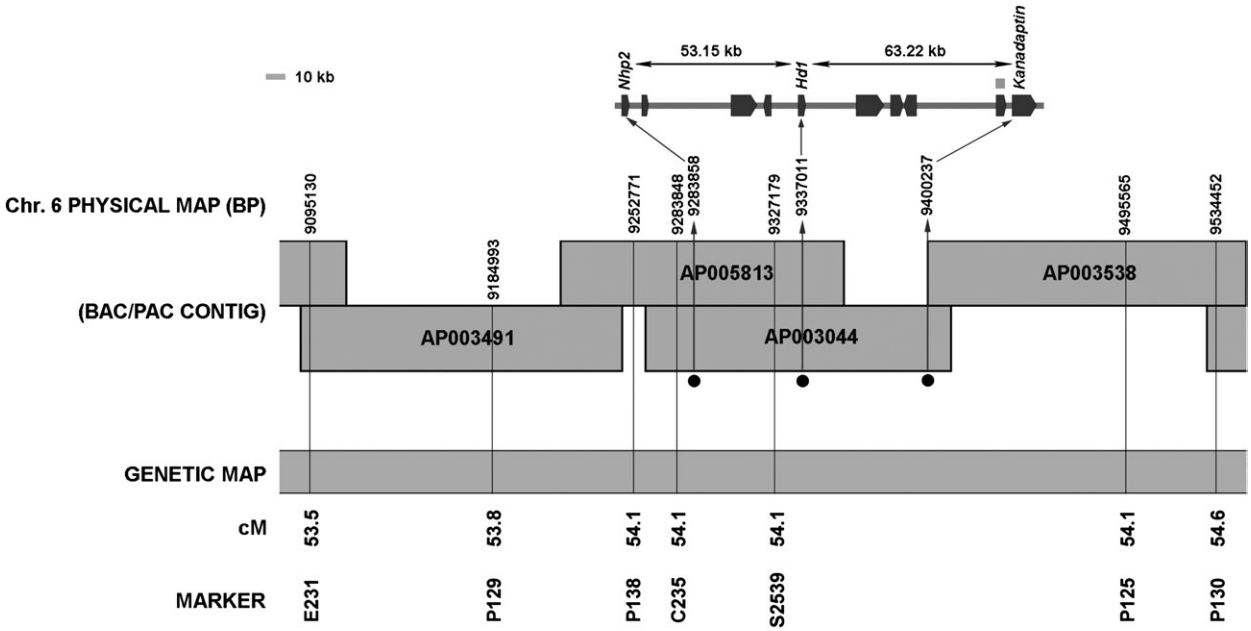

FIG. 1. Positions of the *Hd1* containing P1 derived artificial chromosome (PAC) clones on International Rice Genome Sequencing Project chromosome 6 map of *japonica*. Gray dots = genes selected to prepare radiolabeled hybridization probes; light gray square = re-screening of *glaberrima* library with probe designed from *NAA*.

comparative studies in monocots (Wing et al. 2005; Bennetzen 2007).

The wild germplasm of *Oryza* is a largely untapped reservoir of beneficial genes for rice improvement in terms of yield and nutritional value (Brar and Khush 1997). To establish a genus-wide resource for the study of genome evolution and gene regulatory networks, the *Oryza* Map Alignment Project (OMAP) was initiated in 2003 (see www.omap.org, Wing et al. 2005). As part of this project, 10 representative genome types of *Oryza* were chosen, and deep-coverage BAC libraries were constructed and characterized through BAC end sequencing (BES) and physical map construction (Ammiraju et al. 2006; Kim, Hurwitz, et al. 2008). The BES data coupled with region-specific orthologous BAC sequence data enabled a macrolevel survey of genome size variation (Kim et al. 2007; Zhang et al. 2007), transposable element (TE) dynamics (Piegu et al. 2006; Ammiraju et al. 2007), and centromere evolution (Ma et al. 2007). To investigate local genome organization and gene and TE evolution at a microlevel, we conducted genus-wide orthologous BAC sequence comparisons at a domestication locus, *Hd1* (*Heading date 1*).

Heading date is a quantitative trait important for the adaptation of rice to various growing environments (Yano et al. 2000) and is controlled by at least 14 quantitative trait loci (QTLs) (Yano et al. 1997; Lin et al. 1998, 2000; Yamamoto et al. 1998, 2000). *Hd1*, the gene underlying one of these QTLs, was found to play a central role in the photoperiod pathway of flowering (Lin et al. 2000) and was cloned from the short arm of chromosome 6 (Yano et al. 2000). In this study, we compared the sequence evolution around the *Hd1* locus in nine diploid OMAP species and contrasted the gene and TE dynamics of this locus to the syntenic region from sorghum.

Materials and Methods

Isolation of Orthologous *Hd1* Regions from Diploid Wild Rice

Eight OMAP BAC libraries (*O. nivara*, *O. rufipogon*, *O. glaberrima*, *O. punctata*, *O. officinalis*, *O. australiensis*, *O. brachyantha*, and *O. granulata*) and one unpublished library for *O. sativa* ssp. *indica* cv. 93-11 (www.genome.arizona.edu, Ammiraju et al. 2006) were used to isolate BAC clones orthologous to the *Hd1* locus in the sequenced rice genome (International Rice Genome Sequencing Project 2005). Polymerase chain reaction primers designed from *Hd1* and its flanking genes *Nhp2* (*Nucleolar protein family A member 2*) and *kanadaptin* (*Adaptor protein kanadaptin*) were used to amplify, clone, and generate 200- to 500-bp radioactively-labeled probes from rice genomic DNA using standard methods (Sambrook and Russell 2001). The genetic and physical map of the *Hd1* region on chromosome 6 of rice (fig. 1) and primer sequences are listed in supplementary table S12 (Supplementary Material online). Each probe was hybridized overnight to rice BAC filters using standard protocols (Sambrook and Russell 2001) with modifications (Hass-Jacobus et al. 2006). Filters were exposed to X-ray film overnight at −80 °C. Digital images of the autoradiographs were scored for BACs hybridizing to all three probes using a modified ComboScreen (Jamison et al. 2000; Hass-Jacobus et al. 2006). Validated BACs from each species were extracted and sequenced.

Shotgun Library Creation and Sequencing of BAC Clones

After sizing of the selected BACs, shotgun libraries were constructed by shearing 5 μg of BAC DNA into 3- to 4-kb fragments and cloning these fragments in DH10B

electroporation competent cells following the TOPO TA Cloning Kit protocol (Invitrogen). A total of 1,536 shotgun clones were end sequenced for each BAC using BigDye terminator chemistry on an ABI 3730 automated capillary sequencer (Applied Biosystems). The sequence reads were assembled according to previously described methods (International Rice Genome Sequencing Project 2005), to obtain phase II or III quality sequence for each BAC. The *Hd1* orthologous region between coordinates 12100000 and 12634000 on chromosome 10 of sorghum was identified by homology searches to the sequenced sorghum genome (www.phytozome.net/sorghum) with *Nhp2*, *Hd1*, and *kanadaptin* sequences.

Sequence Annotation

All sequences were masked for repeats using RepeatMasker (www.repeatmasker.org) with a manually curated *O. sativa* TE library (Jiang et al. 2003). This output was then manually compared with structure-based repeat identification methods such as LTR_STRUC (McCarthy and McDonald 2003), LTR_Finder (Xu and Wang 2007), and DOTTER (Sonnhammer and Durbin 1995) to identify complete TEs (all identifiable TE signatures such as terminal inverted repeats [TIR], target site duplication [TSD], polypurine tract [PPT], primer binding site [PBS], and long terminal repeats [LTRs] present). Species-specific repeat libraries generated by RECON (Bao and Eddy 2002) on BAC end sequences of each species (available upon request from Dr Navdeep Gill at Purdue University) were used to identify genome specific repeats in the *Hd1* region. The manually derived repeat file for each sequence was then used to mask the corresponding repeat coordinates.

The sequences were then annotated for genes using a combination of ab initio prediction programs FGENESH (www.softberry.com, Salamov and Solovyev 2000) and GeneMark (Besemer and Borodovsky 2005) and the spliced cDNA/expressed sequence tag (EST) sequence alignment program GeneSeqer (www.plantgdb.org, Usuka et al. 2000). This process (repeat masking followed by gene identification) helped to eliminate TE-related genes (*gag*, *pol*, *env*, and *int*) and improve accuracy of gene calling. In addition, the candidate genes had to meet the following criteria: possession of a known functional domain and availability of a homologue outside of *Oryza*.

Sequence Comparisons

Without masking for repeats, the *Hd1* region sequences from all species were compared with the orthologous region from the reference rice genome (Sasaki and Burr 2000) using the global multiple sequence alignment program LAGAN (Brudno et al. 2003), as implemented in the mVISTA suite of programs (http://genome.lbl.gov/vista/index.shtml, Frazer et al. 2004). Gene annotation files of the compared sequences were uploaded to the mVISTA server to reduce nonsignificant alignments, and the program was run using default parameters. The mVISTA output provided a visual representation of pairwise sequence identity plots and also generated pairwise comparative data for detecting sequence divergence among the *Oryza* species. The rankVISTA feature was used to identify non-exonic segments of the *Hd1* region evolutionarily conserved over approximately 15 My between *O. sativa* and the distal species *O. granulata* and over approximately 50 My between *O. sativa* and sorghum. Sequences that were 10 bp or longer with at least 70% identity between *O. sativa* and sorghum or had more than 90% identity to the distal *Oryza* species were defined as conserved noncoding sequences (CNS) according to previously described criteria (Guo and Moose 2003; Ammiraju et al. 2008).

Phylogenetic Analysis

Phylogenetic analyses were done using a concatenated data set of eight fully conserved (complete coding sequence [CDS] available in all species) and five partially conserved (whole or a part of the CDS is unavailable) genes. Our annotation revealed the presence of alternative noncoding transcripts for two genes in all species. The coding sequences were aligned using RevTrans (Wernersson and Pedersen 2003), whereas the noncoding transcripts were aligned using ClustalW (Thompson et al. 1994). The too data sets were joined and manually edited using Se-Al (http://tree.bio.ed.ac.uk/software/seal/). The aligned DNA sequences were uploaded to PAUP* 4.0b10 (Swofford 1998) for maximum likelihood–based phylogenetic analyses. The best-fit model of evolution was chosen using Modeltest (Posada and Crandall 1998) as implemented in PAUP*. Heuristic searches were performed with 100 replicates of random taxon addition. To evaluate clade support, bootstrap analysis (Felsenstein 1985) with 10,000 replicates was performed using PhyML (Guindon and Gascuel 2003) and graphical representations of phylogenies were made using the tree visualization software FigTree (http://tree.bio.ed.ac.uk/software/figtree/). LTR retrotransposon (RT) and other shared TEs were aligned using ClustalW (Thompson et al. 1994). The dates of insertion for the RTs were calculated using previously described methods (SanMiguel et al. 1998; Ma and Bennetzen 2004). Simple Neighbor-Joining trees were constructed for the shared TEs using Phylo_win (Galtier et al. 1996). The synonymous (Ks) and nonsynonymous (Ka) distance measures of each putative gene were calculated pairwise using the codeml program (runmode = −2, seqtype = 1, CodonFreq = 2, model = 1) as described in PAML (Yang 1997, 2007). The average Ks of the putative genes was used to estimate the divergence time among the *Oryza* species, using a synonymous substitution rate of 6.5×10^{-9} substitutions per site per year (Gaut et al. 1996).

Accession Numbers

The *Hd1 O. sativa* ssp. *japonica* cv. Nipponbare reference sequence was obtained from (GenBank accession numbers: AP005813, AP003044). The sequences of all BACs used in this analysis were deposited to the GenBank data library under the following accession numbers (GenBank accession numbers: FJ581043–FJ581047 and GQ407104–GQ407107].

Table 1. BAC Clones Sequenced and Annotated for the Genus-Wide Comparative Study of the *Hd1* Region in *Oryza*.

Oryza Species	Genome (size Mb)	Clone Address	Insert Size (bp)	Total Length Sequenced[a]	GenBank 2Accession No.
Oryza sativa ssp. *japonica* cv. Nipponbare[b]	AA (389)	P0676F10	142,510	191,872	AP005813
		P0038C05	152,941		AP003044
O. sativa ssp. *indica* cv. 93-11	AA (466)	OSI9Ba0036P05	204,854	204,854	FJ581046
O. nivara	AA (448)	OR_BBa0090L02	153,095	324,233	FJ581047
		OR_BBa0060M01	185,924		
O. rufipogon	AA (439)	OR_CBa0013M07	132,311	225,388	FJ581045
		OR_CBa0017C04	130,939		
O. glaberrima	AA (368)	OG_BBa0059P02	127,473	241,010	FJ581044
		OG_BBa0012J16	157,313		
O. punctata	BB (425)	OP_Ba0017E18	167,491	167,491	FJ581043
O. officinalis	CC (651)	OO_Ba0152I16	217,876	217,876	GQ407106
O. australiensis	EE (965)	OA_ABa0059I14	141,640	576,861	GQ407105
		OA_ABa0091H08	185,468		
		OA_ABa0177P07	246,333		
O. brachyantha	FF (362)	OB_Ba0030H09	112,853	112,853	GQ407107
O. granulata	GG (882)	OG_ABa0076A15	135,193	242,758	GQ407104
		OG_ABa0186L08	158,247		

[a] Sequence size after trimming the BAC overlaps.
[b] Positive control in hybridizations.

Results

Generation of a Comparative Sequence Data Set across Genus *Oryza*

The region containing the *Hd1* gene between genes *lsp* (*leaf senescence protein*) and *kanadaptin* on chromosome 6 of rice (sequence extracted from overlapping genomic clones AP005813 and AP003044 on the build 5 rice pseudomolecule; International Rice Genome Sequencing Project 2005; Rice Annotation Project Database [update] 2008) was selected as *Hd1* is an important domestication gene underlying a major QTL controlling early heading, an important agronomic trait in rice (Yano et al. 1997, 2000; Lin et al. 2000). Eight diploid species from the *Oryza* along with *japonica* rice (*O. sativa* ssp. *japonica* cv. Nipponbare), *indica* rice (*O. sativa* ssp. *indica* cv. 93-11), and sorghum (*Sorghum bicolor* (L.) Moench) were included in the comparative study. The *Hd1* gene plus flanking genes *Nhp2* and *kanadaptin* were used to screen 580,608 clones on 31 BAC filters from *japonica*, *indica*, and eight diploid rice species from OMAP to isolate orthologous BACs. For the *Oryza* species, 15 BAC clones from nine diploids (four AA genome and one species from each of the genomes BB through GG) were isolated and sequenced to generate a comparative data set of about 2.31 Mb in the *Hd1* region (table 1). The orthologous *Hd1* region from sorghum was found on chromosome 10 of the genome sequence (see Materials and Methods). Together with the *japonica* and sorghum sequences, about 3.04 Mb of genome data was annotated. The sequenced *Oryza* genomes and sorghum were compared with the 154,662-bp *japonica* reference sequence from chromosome 6. For simplicity, we refer to the different sequences either by their species names or with the following two-letter codes: *JA*, *japonica*; *IN*, *indica*; *NV*, *nivara*; *RU*, *rufipogon*; *GL*, *glaberrima*; *PU*, *punctata*; *OF*, *officinalis*; *AU*, *australiensis*; *BR*, *brachyantha*; *GR*, *granulata*; and *SB*, *Sorghum bicolor*.

Gene and TE Organization in the *Hd1* Region

The complete sequence of the overlapping genomic clones AP005813 and AP003044 (191,872 bp) from which the *Hd1* reference sequence of *japonica* was derived was re-annotated (see Materials and Methods) for consistency in sequence comparisons across the diploid *Oryza* genomes and sorghum. Of the initial 49 gene models present in the annotation of AP005813 and AP003044 (International Rice Genome Sequencing Project 2005), 32 were removed because they were transposon related. The gene structures of the remaining 17 genes largely agreed with FGENESH predictions (www.softberry.com, Salamov and Solovyev 2000) and with previous annotation. We corrected the gene structure of *importin 9* (GI: 55295981) by merging exons based on EST/cDNA data, FGENESH predictions, and conservation in all *Oryza* species studied. An alternative transcript with full-length cDNA (FL-cDNA) evidence was also identified for this gene. The final re-annotated raw data set for *japonica* consisted of 16 intact genes and one apparent pseudogene formed due to frame-shift mutation.

Gene annotation of the remaining rice genomes and sorghum was done by first repeat masking (www.repeatmasker.org) using a custom database of rice repeats (Jiang et al. 2003; Jurka et al. 2005) and then predicting the gene models using a combination of FGENESH (Salamov and Solovyev 2000), GeneMark (Besemer and Borodovsky 2005), and GeneSeqer (Usuka et al. 2000) (see Materials and Methods). This process enabled us to identify non–transposon-related gene structures (fig. 2*a* and *b* and supplementary tables S1–S11, Supplementary Material online). Table 2 summarizes the general features such as genic, TE, and GC content of the annotated *Hd1* region sequences. Among the species investigated, a positive correlation was found between genome size, GC content, TE content, and sequence length (r^2 between 0.58 and 0.93 for each pairwise comparison). The orthologous gene set occupied approximately the same amount of DNA with no significant differences in the

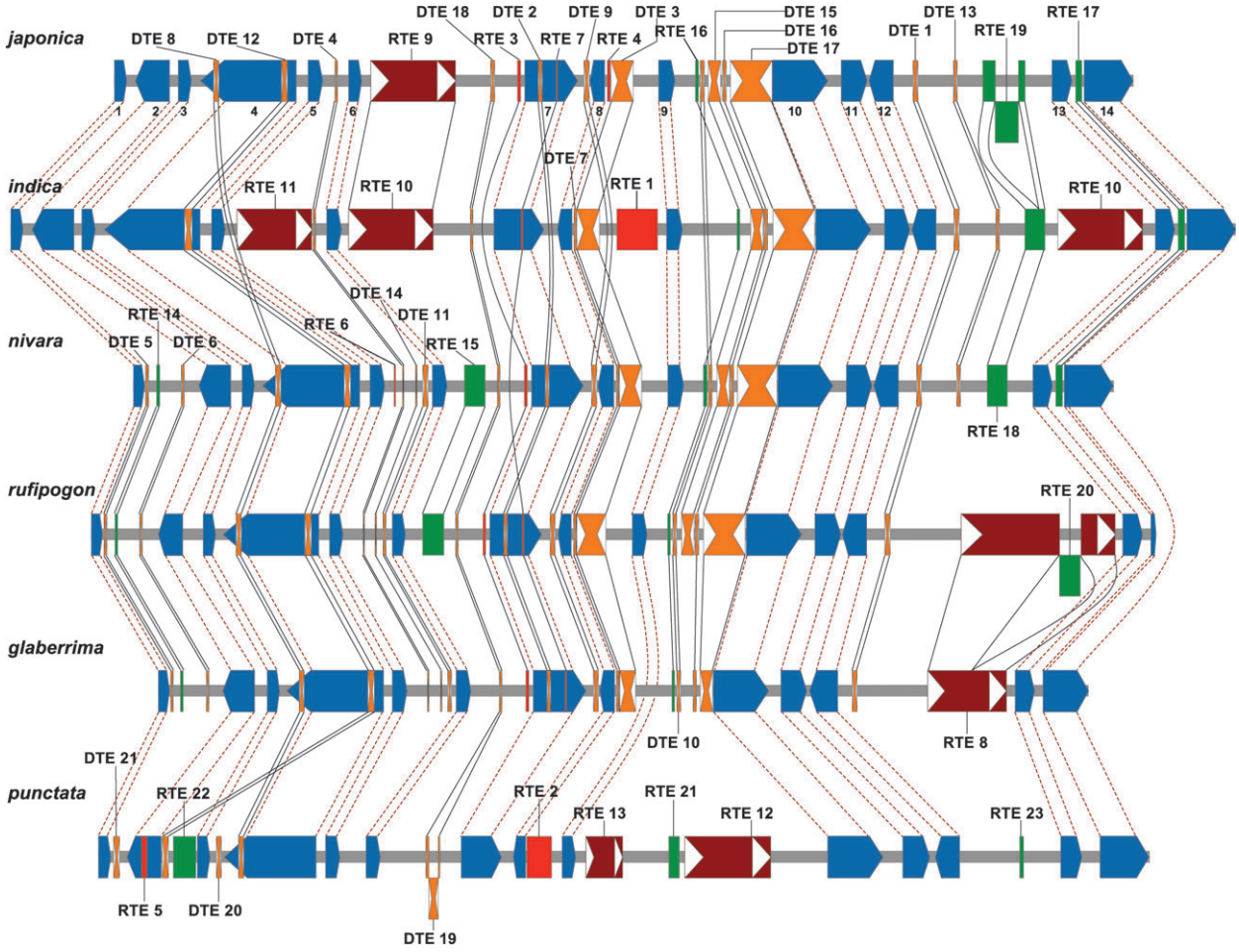

Fig. 2. (*a*) Gene and repeat sharing in the *Hd1* region of AA and BB genomes. Dashed red lines = orthologous genes; black lines = shared repeats. Box colors—dark red with arrows = LTR-RT; green = solo LTRs; red = non-LTR RT; orange = DNA TEs. Gene 9 = *Hd1*. In *rufipogon*, gene 14 is truncated due to end of BAC and only its first exon can be identified. (*b*) Gene and repeat organization in the *Hd1* region of CC–GG genomes and sorghum. Box colors in addition to those in (*a*)—pale blue = *Helitrons*; dark red box no arrows = truncated LTR-RT; hashed box = unknown. Gene 9 = *Hd1*. Repeats less than 600 bp are not represented.

average length of the protein-coding exons and introns that could have had an impact on the size variation. Gene densities in the Hd1 region did, however, vary among the genome types as a result of differences in the length of intergenic regions and were correlated with TE content ($r^2 = 0.93$). The gene density for the AA, BB, and CC genomes was not significantly different from *japonica*, average of one gene every 12.1 kb (Gojobori 2007). However, species with larger genome sizes than *japonica*, such as *australiensis* (EE), *granulata* (GG), and sorghum, had significantly lower gene densities ($P < 0.05$) (table 2). *Oryza brachyantha* (FF) had the highest gene density (1 gene/9.4 kb).

For the rest of the comparative study, we focused on an approximately 155-kb core segment from the *japonica Hd1* reference sequence between genes *lsp* and *kanadaptin* as the complete orthologous sequence was available in at least five of the diploids. The core segment was compared across the *Oryza* genomes and sorghum in terms of the different TE classes identified, their number, and their relative content (table 3). On average, 36% of the *Hd1* region was repetitive in the AA genome species. Outside the AA lineage, the repeat content varied between 26% in *brachyantha* and 80% in *australiensis*. With the exception of *japonica*, *nivara*, *rufipogon*, *glaberrima*, and *brachyantha*, RNA TEs contributed to the bulk of the repeat sequences. However, DNA transposons outnumbered the RNA elements by approximately 1.2- to 8-fold in all species except *australiensis* and sorghum, where the numbers were comparable. Previous studies have shown that DNA transposons occupy primarily euchromatic regions, whereas RNA transposons are more concentrated in heterochromatic regions of the rice genome (International Rice Genome Sequencing Project 2005; Bennetzen 2007).

The single largest TE class contributing to size variation in the *Hd1* region core segment was the LTR-RT. The *gypsy* subclass of this TE was more abundant than the *copia* subclass in all species, both in number and nucleotide contribution. Surprisingly, *copia* elements in both subspecies of *O. sativa* and *gypsy* elements in *nivara* were not found. In *nivara* and *brachyantha*, intact LTR-RT elements were absent, whereas solo LTRs were not found in *glaberrima* causing these species to have less total RNA TE content.

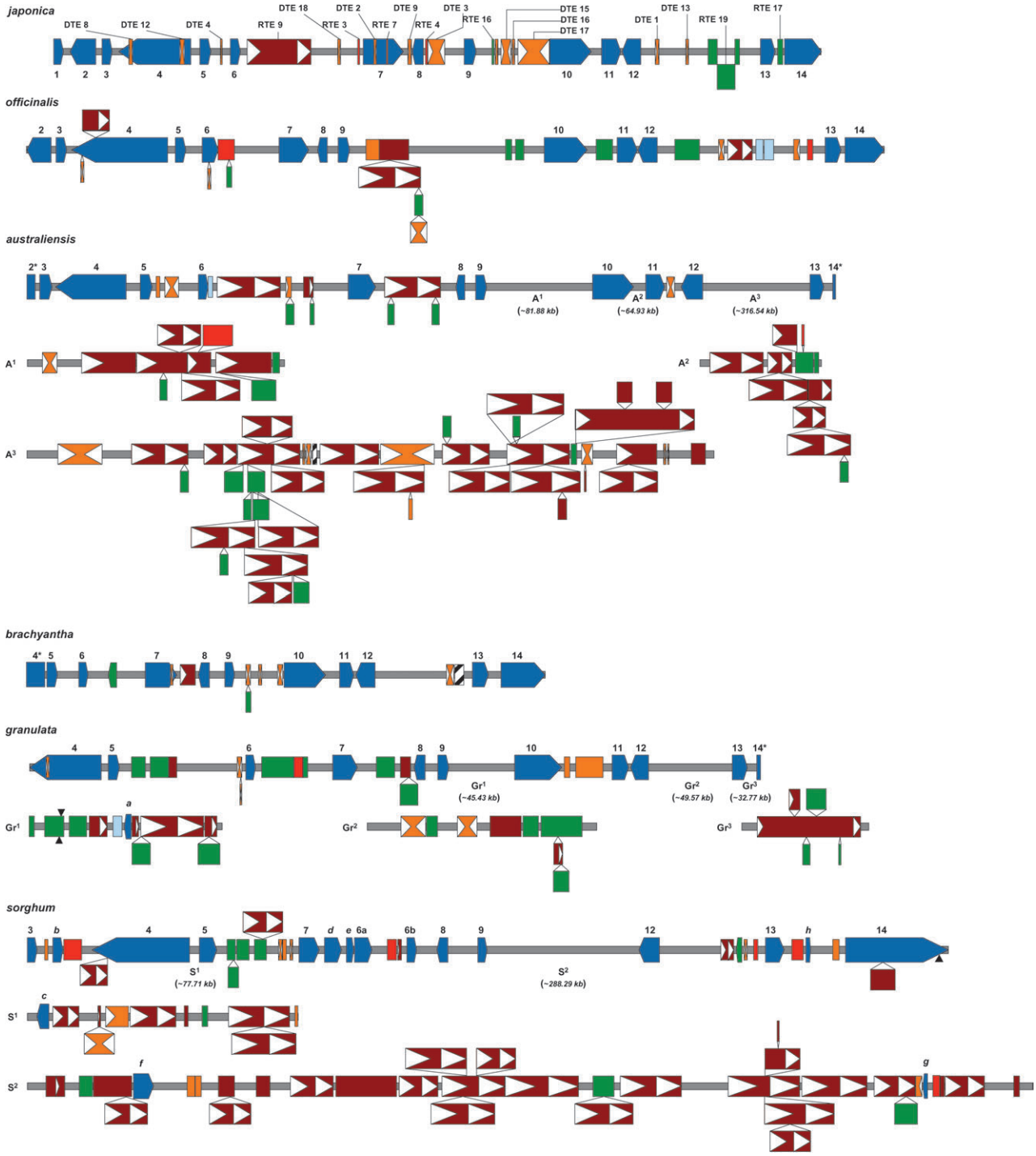

FIG. 2. (Continued)

Japonica and *indica* had a comparable number of RNA transposons; however, *indica* had twice the amount of RT-related sequence as *japonica*, primarily due to three intact LTR-RTs and an intact LINE element in this region. Nested RTs (SanMiguel et al. 1996) were most abundant in *australiensis* and sorghum.

All DNA TEs classes (*Mutator*, *MULEs*, *CACTA*, *MuDR*, *Helitrons*, and so on) discovered in the *Hd1* region core segment were nonautonomous. A few autonomous members were detected in *officinalis*, *australiensis*, *granulata*, and sorghum, but their sequences were interrupted by other elements and therefore do not appear to be mobile. In terms of copy number, miniature inverted-repeat transposable elements (MITEs) outnumbered all other DNA transposon classes, concurring with previous findings (reviewed in Feschotte et al. 2002; Jiang et al. 2004; see also Oki et al. 2008). The MITE density in the AA (one MITE per 3.3 kb) and FF genomes (one MITE per 2.9 kb) was higher than the *japonica* whole-genome average of one per 6.2 kb (Turcotte et al. 2001) and was comparable with the MITE

Table 2. General Features of the *Hd1* Locus Raw Sequence Data Set from *Oryza* and Sorghum.

Sequence Features	*JA* AA	*IN* AA	*NV* AA	*RU* AA	*GL* AA	*PU* BB	*OF* CC	*AU* EE	*BR* FF	*GR* GG	*SB*
Genome size (Mb)	389	466	448	439	368	425	651	965	362	882	735
Sequence length (kb)	192	205	324	225	241	167	218	577	113	243	534
No. of intact genes	16[a]	15	19	22[a]	16	14	13	13[a]	12[a]	12[a]	20[a]
Gene density (kb/gene)	12.0	13.7	17.1	10.2	15.1	12.0	16.8	44.4	9.4	20.2	26.7
GC content (%)	45	45	42	43	42	42	43	46	41	45	45
Genic region (%)[b]	36.6	33.8	24.1	33.8	29.5	38.3	30.7	8.7	40.3	19.7	15.1
TE region (%)[c]	40.3	44.3	39.4	38.7	37.6	29.8	50.7	80.6	27.7	57.4	60.0

[a] Genes at the end of the BAC sequence are counted as intact.
[b] Includes introns and untranslated regions.
[c] Simple sequence repeats and low-complexity regions not included.

density observed at the *monoculm1* locus (Lu et al. 2009). The TE families in each genome were identified based on homology to known *O. sativa* repeats; thus, outside the AA lineage, the divergence of the TE sequences progressively increased leading to greater difficulty in identifying complete elements.

Loss of Sequence Conservation beyond the AA Genomes

Although gene order and orientation were highly conserved in the *Hd1* core segment, the intergenic regions were peppered with TEs of different classes, leading to increased variability outside the AA lineage in *Oryza*. The unmasked *Hd1* core from all species was compared with the *japonica* reference sequence using global pairwise sequence alignments (see Materials and Methods) to determine the amount of sequence divergence in intergenic regions. mVISTA (Frazer et al. 2004) was used to generate percent identity plots (data not shown) indicating an average of 18–40% sequence divergence among the AA genomes and a 72–95% sequence divergence from *punctata* (BB) to *granulata* (GG) in the intergenic regions (supplementary table S17, Supplementary Material online). The intergenic sequence of *japonica* and sorghum are 99% different, as there is about 50 My of divergence between them. These results are in agreement with recent studies in the *Adh1* region (Ammiraju et al. 2008) and also with previous studies in the AA lineage (Ma and Bennetzen 2004; Ma et al. 2004) and cross-genera (Tikhonov et al. 1999; Ramakrishna, Dubcovsky, et al. 2002) that demonstrated the effect of counter-evolutionary forces such as LTR-RT amplification, unequal and illegitimate homologous recombination, small indels, and numerous spontaneous substitutions in defining the highly divergent sequence composition of the intergenic regions over a 2- to 14-My interval.

We found 30 (18 intronic and 12 intergenic) significant ($P < 0.05$) CNS regions ranging in size from 60 to 1,634 bp in the *Hd1* region, using a distant species comparison between *japonica* and *granulata*. Of these, 11 were at orthologous positions in sorghum (3 intronic and 8 intergenic), whereas the rest were *Oryza* specific. To further investigate the potential regulatory nature of the CNSs, we queried PLACE (Higo et al. 1999), a database of *cis*-acting DNA elements, and found on average three binding targets for 5–89 different regulatory factors per CNS region (supplementary table S18, Supplementary Material online). The mean length of each potential site was about 6 bp resulting in the high frequency of regulatory sites.

The sequence of the *Hd1* core segment was highly conserved in the genic regions, with 99% of the exons, 55–75% of the introns, and 60–94% of the untranslated regions being conserved with at least 80% identity. The average percent protein similarity was greater than percent nucleotide identity, suggesting conservation of putative protein function (data not shown).

Rapid *Hd1* Gene Dynamics in *Oryza* and Disruptions of Microsynteny in Sorghum

Despite high gene colinearity and sequence conservation in the *Hd1* core region, the *Hd1* gene of *glaberrima*, *rufipogon*, and *punctata* had major structural changes. The *Hd1* gene (gene 9) was potentially deleted in *glaberrima* (fig. 2*a* and *b*), based on the following. First, after hybridization to the *glaberrima* BAC library, no single or overlapping BAC clones positive for all probes (*Nhp2*, *Hd1*, *NAA* [*neutral aminoacid transporter* gene], and *kanadaptin*) were found (fig. 1). Second, on the *glaberrima* BAC Fingerprint Contig map (www.omap.org/cgi-bin/status/status.cgi), the *Nhp2–NAA*-positive clones formed a contig anchored to chromosome 6, whereas 21 of the 24 *Hd1*-positive BACs were on contig 626 anchored to chromosome 9. Third, the comparative synteny map between *glaberrima* and *japonica* (http://www.omap.org/cgi-bin/status/status.cgi) shows no synteny between chromosomes 6 and 9 in *glaberrima*. Fourth, homology searches against the *japonica* genome, with the *Hd1* probe sequence, identified with low significance (*E* value $= 4 \times 10^{-8}$), a *zinc finger protein* gene (GI: 50726044) on chromosome 9. The above results indicate that the *Hd1* hits to the *glaberrima* library were false positives and that the gene is probably deleted in *glaberrima*. Sequence annotation confirmed the absence of the *Hd1* gene from the *Lsp–Hd1–kanadaptin* region of *glaberrima*.

In *rufipogon*, a double-thymidine deletion in the second exon of *Hd1* results in a frame-shift mutation that possibly inactivates the gene or the open reading frame (ORF) might code for a 342–amino acid protein. Interestingly, this putative *Hd1* ORF is an exact match to an allele (GI: 11094207) from *O. sativa* ssp. *indica* cv. Kasalath, suggesting that the *rufipogon* gene may still be protein coding. In *punctata*, a single-base transversion (C → A) in the first exon of

Table 3. Content and Classification of TEs in the Core Segment (*a*) across the AA Genomes of *Oryza* and (*b*) across the Wild *Oryza* Genomes BB through GG and Sorghum.

	a										*b*											
Species	JA		IN		NV		RU		GL		PU		OF		AU		BR		GR		SB	
Genomes	AA		AA		AA		AA		AA		BB		CC		EE		FF		GG			
TE classes	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%	NE	%
Class I: RTE																						
LTR																						
Ty1-*copia*		NF		NF	1	0.20	1	1.17	1	0.21	1	3.55	1	3.10	5	6.20	1	0.07	4	11.35	13	8.74
Ty3-*gypsy*	1	8.29	4	20.53		NF	2	12.11	2	8.67	1	8.15	4	13.49	39	56.99	2	3.08	6	9.86	29	36.09
TRIM	1	0.15	1	0.10	1	0.08	1	0.11	1	0.13	1	0.13		NF		NF		NF		NF		NF
Solo LTR	4	5.06	3	2.33	5	5.28	3	3.82		NF	3	3.42	7	8.05	21	6.97	3	1.32	18	23.40	9	3.96
Non-LTR																						
LINE	4	0.91	5	6.10	1	0.14	4	0.93	5	1.62	3	2.43	2	2.56	2	0.99	1	0.14	1	0.69	11	1.91
SINE	3	0.55	1	0.10	2	0.37	2	0.32	2	0.34	1	0.45	1	0.19	2	0.08		NF		NF		NF
Total RTE	13	14.96	14	29.16	10	6.06	13	17.45	11	10.96	10	18.13	15	27.38	69	71.23	7	4.60	29	45.30	62	50.70
Class II: DTE																						
Mutator	3	1.30		NF	5	1.78	4	1.45	4	1.78	1	0.31	2	0.85	4	0.41	1	0.49	2	0.58		NF
MULE	12	7.76	22	5.27	10	8.44	13	4.87	17	4.25	20	3.90	8	1.78	10	0.54	11	4.80	4	2.70	2	0.24
En-Spm		NF		NF		NF		NF		NF	1	0.90		NF	4	3.58		NF		NF	13	3.18
CACTA		NF		NF		NF		NF		NF	1	0.35	5	2.77	4	1.01		NF		NF		NF
hAT	2	0.92	3	0.66	2	0.96	2	0.85	2	1.00	2	0.62	1	1.90	2	0.09	8	2.63	4	3.69	4	0.28
MITE-*Stowaway*	22	3.09	21	2.43	28	4.09	29	3.86	27	4.33	14	2.04	14	1.82	13	0.56	14	2.60	7	0.54	7	0.32
MITE-*Tourist*	26	3.42	27	2.54	25	3.66	20	2.69	19	2.91	11	1.36	16	2.36	13	0.54	21	5.84	15	1.54	26	1.56
Helitrons	2	1.76	5	1.72	5	3.13	6	3.45	6	2.65	4	0.64	3	2.12	2	0.16		NF	1	0.73	2	0.02
Unclass		NF	3	0.67		NF	1	0.53	2	0.30	1	0.14	2	0.19		NF		NF	1	0.13		NF
Total DTE	67	18.25	81	13.29	75	22.06	75	17.72	77	17.23	55	10.26	51	13.77	52	6.88	55	16.36	34	9.91	54	5.60
Unknown	1	1.04	2	0.99	18	2.49	18	3.26	15	1.36	17	1.95	10	1.08	54	2.03	26	4.11	11	0.65	4	4.98
SSR	19	0.54	25	0.50	13	0.50	14	0.37	9	0.35	11	0.53	10	0.20	18	1.19	13	0.47	15	0.34	33	0.45
Total repeats	100	34.79	122	43.94	116	31.12	120	38.79	112	29.90	93	30.86	86	42.44	193	80.33	101	25.54	89	56.21	153	61.72
LC	16	0.55	13	0.24	22	0.71	20	0.64	24	0.82	20	0.69	15	0.44	38	0.25	22	1.06	22	0.46	56	0.49
Core length (bp)		154,662		186,063		148,778		168,269		141,300		159,427		167,133		576,861		101,112		241,588		523,305

NOTE.—Unknown refers to de novo repeats identified using RECON except in sorghum where it corresponds to unknown LTR-RT. Percent values represent the absolute abundance of the TE classes in the core segment of each species. RTE, RNA TE; DTE, DNA TE; NE, number of TEs (intact and incomplete); TRIM, terminal repeat in miniature; NF, not found; SSR, simple sequence repeats; LC, low-complexity regions.

Hd1 inserts a stop codon that may inactivate the *Hd1* protein. However, other outcomes such as a generation of two peptide fragments may also be possible. Other noticeable differences in the *Hd1* core region included an unreported alternative splice variant with FL-cDNA evidence for *importin 9* (gene 4) and *TA1* (*Transcriptional Activator protein*; gene 11) found in all *Oryza* species but sorghum. A non–transposon-related hypothetical gene (fig. 2*b*, green arrow) and *AdIT* (*adenylate isopentenyl transferase* protein-coding gene; gene *a*) were identified in *brachyantha* and *granulata*, respectively, but were not found in the rest of the species.

The *Hd1* syntenic region in sorghum was more dynamic than *Oryza* in the number of structural changes disrupting microcolinearity. Sorghum had seven additional genes (*b–h*) and one hypothetical in the *Hd1* region. Of these, six (*b–f* and *h*) had *japonica* homologues, whereas gene *g* was found to have a homologue in maize (*Zea mays*) only. The gene *polycomb protein EZ1* (gene 10) and *TA1* were found to be absent from the *Hd1* region of sorghum. *EZ1* was present about 8.05 Mb upstream of *HSBP* (*Heat Shock Factor Binding Protein*; gene 3) on chromosome 10 of sorghum, whereas *TA1* was found on chromosome 2. We did not find any significant TE sequence overlap between the flanking regions of these genes (~10 kb on each side) and the *Hd1* region of sorghum that would suggest TE involvement in gene translocation, nor were the genes embedded within transposons.

A multiple alignment of the genes *SEPK1* (*Somatic Embryogenesis Protein Kinase 1*; gene 7) and *SERK1* (*Somatic Embryogenesis Receptor Kinase 1*; gene 6) in each species suggested that *SERK1* is derived from *SEPK1* via gene duplication followed by a stop codon in the seventh exon of *SERK1*. *SERK1* protein differs from *SEPK1* in that it is shorter with no kinase domain and no known function as opposed to *SEPK1*, which is expressed in developing ovules, the embryo sac, and anthers (Hecht et al. 2001; Albrecht et al. 2005). We found two copies of *SERK1* (*6a* and *6b*) in sorghum, and its order with *SEPK1* was reversed, but transcriptional orientation was same as in the rest of the species. *SERK1-6a* and *SERK1-6b* were 26% divergent from each other and 17% and 28% divergent from *SEPK1*, respectively, indicating that *SERK1-6a* was evolutionarily more closely related to *SEPK1* than *SERK1-6b*.

Genus Phylogeny and the Timing of Speciation Events

To better understand gene and TE dynamics in *Oryza* and sorghum and to relate them to speciation events, we determined the phylogenetic relationships and estimated time of species radiation using a set of 13 orthologous genes (genes 5–8 and 10–13 were fully represented in all species, whereas partial CDS sequences were available for genes 2, 3, 4, 9, and 14) and two noncoding alternative transcripts from genes 4 and 11 (fig. 3). Gene *6a* was used, as it was the least diverged from the *SERK1* found in the other species. The coding and noncoding sequences were used to build a consensus tree using maximum-likelihood search methods. The tree topology (fig. 4) and the timing of species radiation largely agreed with previous reports (Ge et al. 1999; Guo and Ge 2005; Zou et al. 2008). However, there were some incongruities such as *brachyantha* being the most basal species and *glaberrima* segregating in the same clade as *rufipogon*. Such types of incongruence are commonly observed when the data set is limited and only one type of gene data is used (organelle genes were not investigated) to build a phylogeny (Holder and Lewis 2003). Also, the rate of evolution in different chromosomal regions may vary causing noticeable differences in phylogeny among the same group of species (Duan et al. 2007).

The estimation of sequence divergence and calculation of time of speciation within *Oryza* were done by computing the Ks and Ka rates for each of the core genes (supplementary table S15, Supplementary Material online). All eight fully represented genes appeared to be under strong purifying selection as evidenced by Ka/Ks values less than 1, except *Hd1* in *nivara* and *EZ1* in *glaberrima*. Previous studies have shown that averaging of the synonymous distances over a number of genes can help to reduce inaccuracies in estimates of evolutionary rate variation (Ma and Bennetzen 2004). Therefore, the Ks values were averaged for the combined core gene data set of eight genes and used for the estimates of divergence times among the species. Using the *Adh1* synonymous substitution rate of 6.5×10^{-9} (Gaut et al. 1996), we estimate the divergence time of genus *Oryza* from the common ancestor to be about 16 My (supplementary table S16, Supplementary Material online). The Ks values of *SEPK1* and *TA1* genes of sorghum were unusually higher than the rest of the genes probably due to low codon-usage biases (Sharp and Li 1987; Fitch and Strausbaugh 1993) and were not included in the calculation of divergence time between *Oryza* and sorghum, which was found to be about 51 My in agreement with previous fossil estimates (Stebbins 1981; Wolfe et al. 1987). The AA genomes diverged from each other approximately 1.5 My, whereas the divergence of the two *O. sativa* subspecies occurred about 300,000 years ago. The *punctata*, *officinalis*, and *australiensis* (BB, CC, and EE) genomes last shared a common ancestor between approximately 7 and 10 My. In *punctata*, seven core genes had a higher number of synonymous substitutions than *officinalis*, causing an overestimation of its divergence time from the AA genome clade. *Oryza brachyantha* is known to evolve at a faster rate than the other *Oryza* species (Zou et al. 2008). The time of speciation within *Oryza* is largely concordant with other studies (Ammiraju et al. 2008; Lu et al. 2009).

TE-Driven Complexities in the Genomes of *Oryza* and Sorghum

Even though LTR-RTs were the most dominant class of TE, the high numbers of DNA TEs made them interesting. Most of the shared DNA transposons were among the *Oryza* AA genomes as it was more difficult to identify shared transposons in the more divergent genomes (fig. 2*a* and supplementary table S13, Supplementary Material online). Out of 21 intact (both TIRs present) DNA TEs identified in the AA and BB genomes, 18 were shared across the AA lineage,

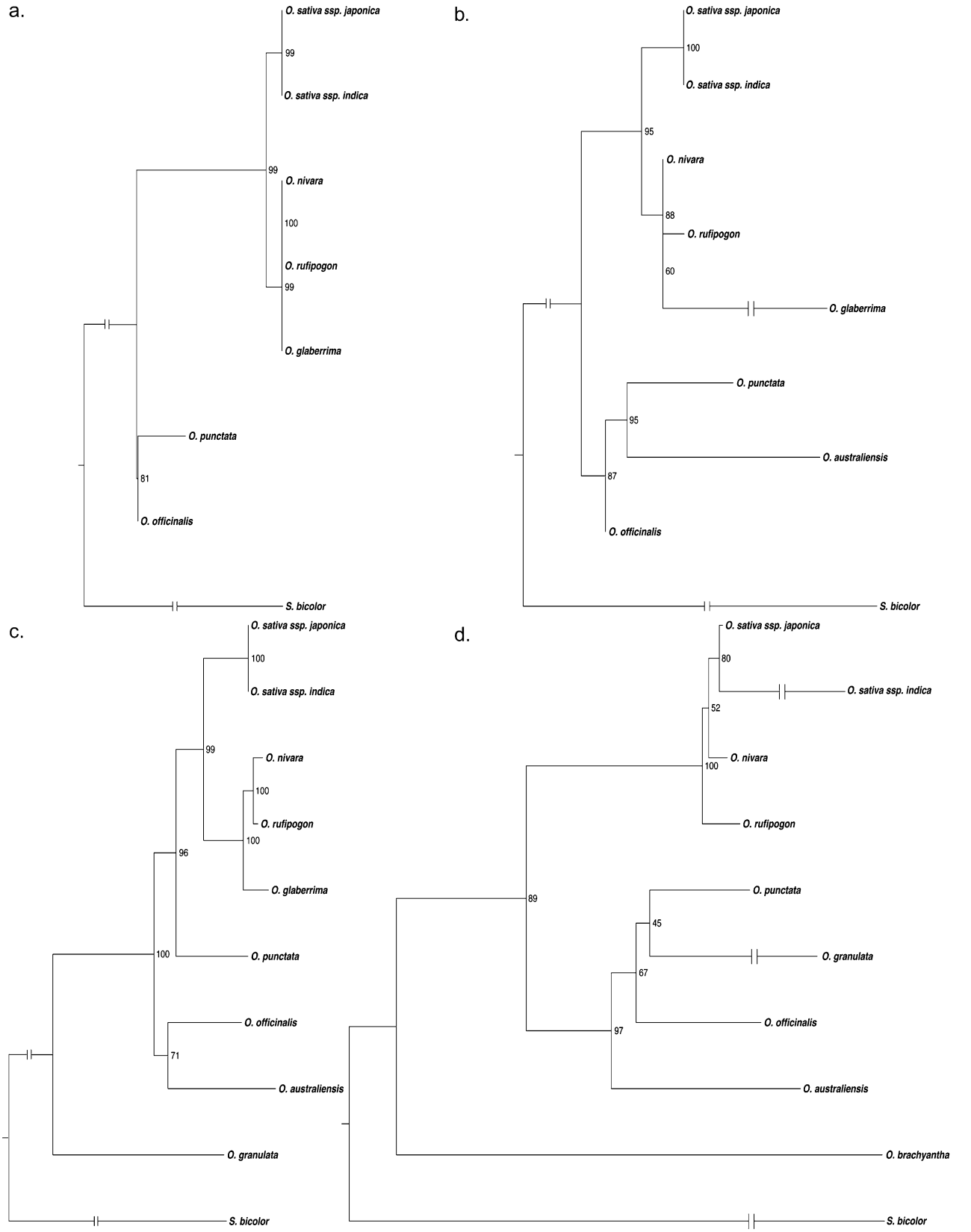

Fig. 3. Phylogenetic relationship from a combined data set of 13 genes across *Oryza* and sorghum. All trees were built using maximum-likelihood method. Bootstrap values are at the nodes. (*a–d*) Individual gene trees of four partially represented genes (2, 3, 4, and 9). (*e–h*) Individual phylogenies of one partially represented gene (14), two noncoding transcripts (genes 4 and 11), and a single phylogeny built using a concatenated sequence data set from eight fully represented genes (5–8 and 10–13).

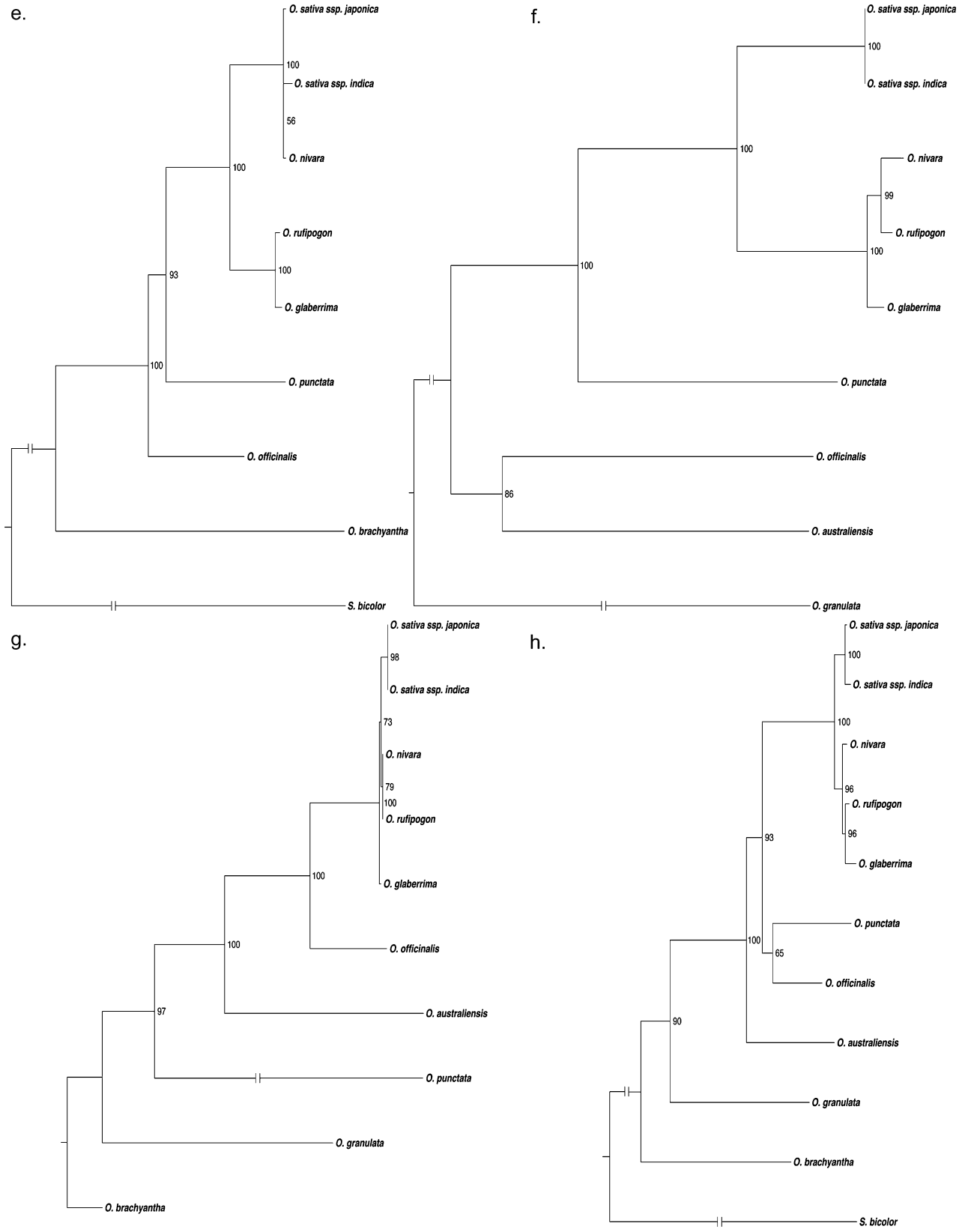

FIG. 3. (Continued)

including three with the BB genome (DTE8, DTE12, and DTE18). Interestingly DTE8 was in the intron of gene 4 and in all *Oryza* lineages except *australiensis*. DTE12 showed movement out of the intergenic region between genes 2 and 3 and into the intron of gene 4 after the AA–BB genome split. We identified a nonautonomous

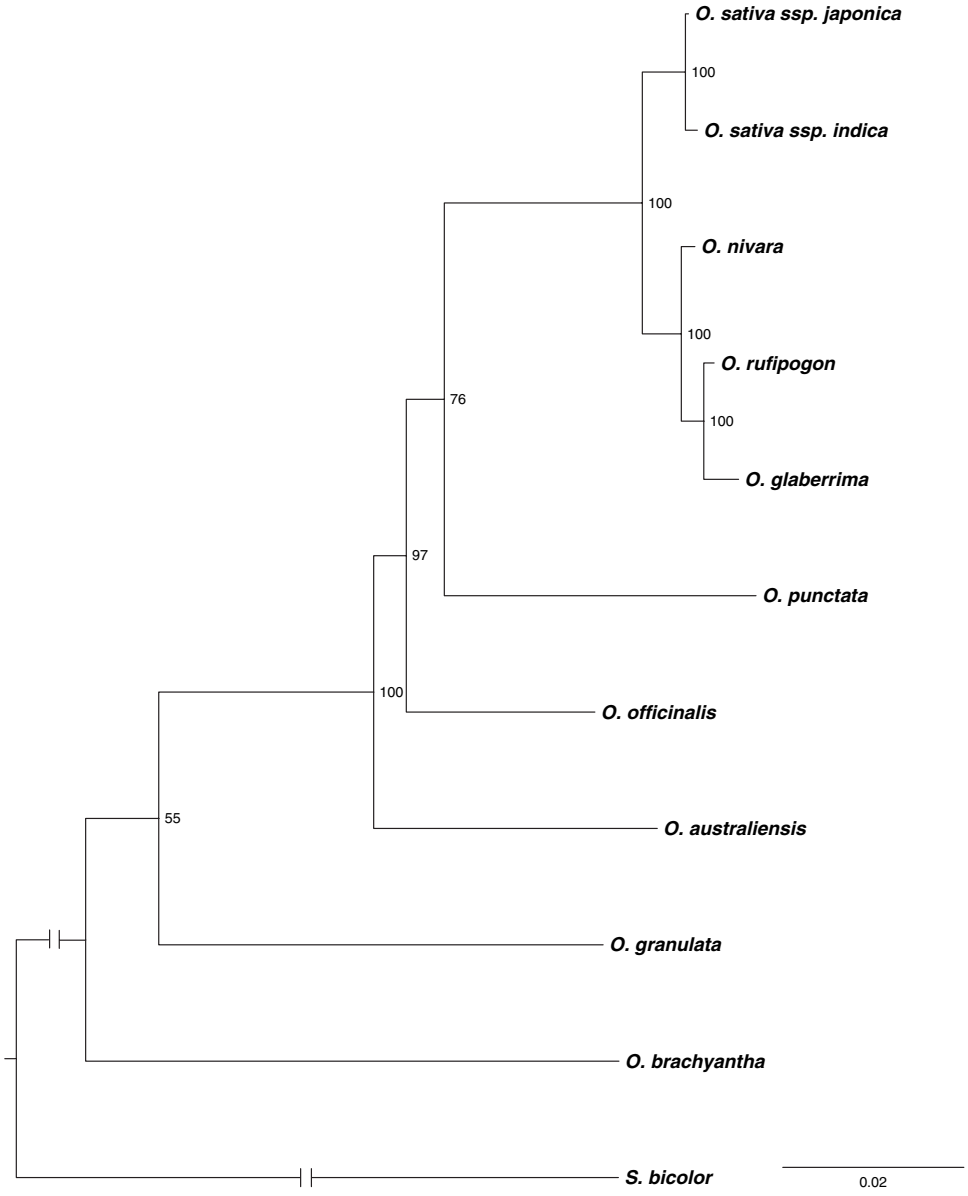

FIG. 4. Consensus *Oryza* phylogeny. Sorghum is the outgroup and bootstrap values are at the nodes.

2.4-kb *Helitron* element (DTE3) shared in all AA genomes, with a lineage-specific 883-bp *LINE-6* indel in *glaberrima* and a 1.2-kb unknown insertion in *rufipogon*. Each of the genomes had 25–50 MITE elements, the majority of which were shared in the AA genomes. In total, 108 of 141 *Stowaway* MITEs and 82 of 128 *Tourist* MITEs in the *Hd1* region of the AA species were shared (supplementary table S14, Supplementary Material online).

Across all genomes, we identified 61 intact LTR-RT, with 24 present in each of *australiensis* and sorghum. In all species, the chromosomal region 3′ to the *Hd1* gene was found to be more prone to TE insertions. This is most evident in the CC, EE, and GG genomes of *Oryza* and sorghum, where the region beyond *Hd1* expanded five to seven times over *japonica*, primarily due to nested LTR-RTs. We also identified 67 solo LTRs in all the species, the bulk being present in *australiensis* and *granulata*. *Oryza granulata* had the highest density of solo LTRs indicating a higher rate of unequal recombination, the primary method of solo LTR formation (Ma et al. 2004; Bennetzen et al. 2005). Sharing of RNA TEs in *Oryza* was not as common as for DNA elements, even among the AA genomes. *Indica* had three intact *gypsy* elements, one of which (RTE9) was shared with *japonica*. A *gypsy* element (RTE8) was shared between *glaberrima* and *rufipogon*, although the *rufipogon* copy had numerous small indels, a large number of substitutions in its LTRs, and a nested solo LTR insertion. The estimated insertion times of the LTR-RTs calculated using the nucleotide substitution rate of 1.3×10^{-8} (Ma and Bennetzen 2004) indicate that the expansion of the *Hd1* region in sorghum occurred exclusively in the past 1–2 My, whereas in *australiensis* the expansion occurred in at least two stages; 3–5 My and less than 2 My. In the rest of the species, 85% of the LTR-RT insertions were in the past 2 My (table 4).

Table 4. Intact LTR-RT and Solo LTRs of the *Hd1* Core Region.

Species	Intact LTR-RT	Solo LTRs	Range (My)	Average (My)	Intact : Solo LTR	Repeat Sharing
japonica	1	3	0.04	0.04	0.33	RTE9 shared with *indica*
indica	3	2	0.07–0.18	0.12	1.50	RTE10 shared with *japonica*
nivara	NP	4	NA	—	0.00	
rufipogon	1	2	1.94	1.71	0.50	RTE8 shared with *glaberrima*
glaberrima	1	NP	1.71	NA	RTE8 shared with *rufipogon*	
punctata	2	3	0.82–2.01	1.42	0.67	
officinalis	2	7	1.35–6.35	3.85	0.29	
australiensis	24	20	0.08–9.05	2.06	1.20	
brachyantha	NP	2	NA	—	0.00	
granulata	3	17	0.04–4.31	1.91	0.18	
sorghum	24	7	0–2.56	0.84	3.43	

NOTE.—NP, not present; NA, not applicable.

LTR-RT insertions in *japonica*, *indica*, and *punctata* occurred after speciation in their respective lineages, whereas the shared LTR-RT between *glaberrima* and *rufipogon* (RTE8) was inserted before their divergence. In *indica*, RTE10 has two independent insertions in the *Hd1* region (fig. 2*a*) occurring about 70,000 years ago; and in *japonica*, the RTE9 element was inserted about 30,000 years ago in the same segment of the *Hd1* region as its counterpart in *indica* (~1,000-bp segments flanking RTE9 in both subspecies are 99% identical). Considering that the two subspecies diverged 0.32 My (Ma and Bennetzen 2004; this study), this suggested either a possible mixing of genealogies leading to errors in relating LTR-RT insertion times with speciation similar to previous studies (Zou et al. 2008) or a horizontal transfer of TEs, which are rare in plants (Fortune et al. 2008). The data indicated that RTE8 inserted into the progenitor of the *nivara*, *rufipogon*, and *glaberrima* lineages about 1.94 My, and was deleted in *nivara* and retained in *rufipogon*, but with more sequence changes than its homologue in *glaberrima*.

Discussion

TEs Have Shaped the Architecture of the *Hd1* Region in *Oryza* and Sorghum

Most angiosperms have large complex genomes with a large proportion being composed of repetitive DNA (Kellogg and Bennetzen 2004). The TE-driven size variation of gene-rich euchromatin in closely related (Ma and Bennetzen 2004; Ammiraju et al. 2008; Lu et al. 2009) and divergent plant genomes (reviewed in Bennetzen 2002) causes species with larger genomes to have lower gene densities. This is also demonstrated in the *Hd1* region where genome size, TE content, and gene density variation are all correlated with each other ($r^2 = 0.84–0.93$) mirroring previous investigations within *Oryza* (Piegu et al. 2006; Zuccolo et al. 2007; Ammiraju et al. 2008; Lu et al. 2009) and between rice and sorghum (Ramakrishna, Dubcovsky, et al. 2002; Ma et al. 2005). The only anomaly to this rule was *officinalis*, which showed increased LTR-RT removal in this region and despite having a large genome size had a gene density comparable with the AA and BB genome species. This is in contrast to the *Adh1–Adh2* locus, which shows marked expansion (Ammiraju et al. 2008) in *officinalis*. This reflects local variation within genomes indicating that regional expansion and contraction forces do not act uniformly in genomes.

Our analyses also indicate that the intergenic regions have experienced rapid and mostly independent lineage-specific changes over an approximately 15 My time period within the *Oryza* during which almost 95% of the sequence progressively diverged relative to *japonica*. This is due largely to DNA transposon indel activity among the AA genomes and LTR-RT dynamics in BB to GG genomes. Since the AA–BB genome split about 6–9 My (Ammiraju et al. 2008; Lu et al. 2009; this study), there have been 20–26 new DNA transposon insertions in the *Hd1* region of each of the AA genomes, and an equal number (20–24) are in a fragmented state suggesting active excision and/or deletional decay of transposon sequences. In conjunction with the *Adh1* study (Ammiraju et al. 2008), this suggests that the rate of DNA transposon turnover has remained constant in the AA genome species keeping the average whole-genome DNA TE content at about 18% (Ammiraju et al. 2006).

On a whole-genome level, independent LTR-RT amplifications cause much of the size variation observed in the grasses (SanMiguel et al. 1996; Chen et al. 1998; Tikhonov et al. 1999; Tarchini et al. 2000). Assuming that the *Hd1* region is an indicator of genome-level size variation, we found expansion in *australiensis* (EE), *granulata* (GG), *indica* (AA), and sorghum and contraction in *glaberrima* (AA) and *brachyantha* (FF), relative to *japonica* and due primarily to LTR-RT dynamics. In *brachyantha*, no intact LTR-RTs were found and nearly 9% of the *Hd1* region was covered in MITEs. This suggests 1) no recent insertions of LTR-RT in the *Hd1* region, 2) genome-level suppression of LTR-RT amplification, and 3) MITE amplification, especially of one or more members of the *Tourist* superfamily. Previous sequencing of genomic BACs and cytogenetic evidence indicates that LTR-RT amplification is indeed suppressed in *brachyantha* (Uozu et al. 1997; Zhang et al. 2007; Zuccolo et al. 2007). Jiang et al. (2003) found dramatic amplification of a *Tourist* superfamily MITE (*miniature Ping*) in *O. sativa* that inserted copies into gene-rich, low-copy regions of the genome in cultivars adapted to environmental extremes (Jiang et al. 2003). It is possible that a similar burst in amplification of *Tourist*-related MITEs may have occurred in

brachyantha that grows on harsh lateritic soils of Africa (Vaughan et al. 2008).

Intact LTR-RT to solo LTR ratio is an indicator of the amount of retroelement recombination in a genome and can lead to regional contraction (*nivara* [AA]) or marginal expansion (*rufipogon* [AA], *punctata* [BB], and *officinalis* [CC]) in otherwise large genome species relative to *japonica*. Approximately 88% of the solo LTRs identified in this analysis had TSDs, of which the majority (69%) was perfect duplications indicating intra-element intra-strand unequal homologous recombination as the primary mechanism for LTR-RT removal. The LTR-RT insertion times and their percent sequence contribution (~63%) to the *Hd1* region of *australiensis* agreed with the postspeciation timing of three retrotranspositional bursts in this species (~0.75 My [*RIRE1–copia*], ~1.8 My [*Wallabi–gypsy*], and ~3 My [*Kangourou–gypsy*]) and the collective contribution (~60%) of all three LTR-RTs toward doubling the genome size (Piegu et al. 2006).

An equally high incidence of solo LTR and fragmented LTR-RTs (24 intact LTR-RT, 20 fragmented LTR-RTs, and 20 solo LTRs) indicates that the *Hd1* region and perhaps the whole genome may be in a state of size equilibrium, at least for the past 3 My. Similarly, *granulata* also has undergone retrotranspositional bursts of three LTR-RTs, namely *Atlantys* (Zuccolo et al. 2008), *Gran3* (Ammiraju et al. 2007), and *Wallabi* (Piegu et al. 2006) in the past 3–6 My after speciation leading to its genome size increase. However, compared with *australiensis*, there is a higher incidence of solo LTR formation and fragmented LTR-RTs in the *Hd1* region of *granulata* (3 intact LTR-RT, 7 fragmented LTR-RTs, and 17 solo LTRs) leading to contraction. These observations lend support to the predicted half-life of less than 6 My for LTR-RTs in rice during which 75% of LTR-RTs are lost (Ma et al. 2004). Ten of 17 solo LTRs had imperfect TSDs and 3 of 10 LTR-RTs were intact suggesting an inter-element unequal recombination mechanism at work for LTR-RT removal in *granulata*. However, as proposed by earlier authors (Devos et al. 2002; Ma and Bennetzen 2004; Bennetzen et al. 2005), both forms of recombination only cause maintenance or attenuation of the rate of genome growth but not a reversal of genome "obesity." Coupled with studies of the *Adh1* region (Ammiraju et al. 2008), it appears that genome size evolution within *Oryza* is RT driven and lineage specific.

The repeat composition and the insertion times of the LTR-RTs in the *Hd1* region of sorghum mirrored the whole-genome data (Paterson et al. 2009). However, the sorghum *Hd1* region was rich in nested LTR-RT insertions contrary to previous studies (Tikhonov et al. 1999; Ramakrishna, Emberton, et al. 2002; Song et al. 2002; Bennetzen et al. 2005; Paterson et al. 2009) that found RTs to be preferentially eliminated from gene-rich regions of this species. Given the fact that LTR-RT distributions are random and the genome assembly is robust in the gene-rich regions of sorghum, where about 98% of the genes map correctly to their respective chromosomes and are validated using genetic, physical, and syntenic information (Paterson et al. 2009), this suggests that the region between *Hd1* (gene 9) and *Armadillo* (gene 12) is highly labile to RT insertions. Nonuniform gene distribution (tracks of high gene density flanking RT-rich environments) has previously been observed at the *FER* and *Mi* locus of tomato located in the pericentromeric heterochromatin (van Daelen et al. 1993; Guyot et al. 2005). Although the *Hd1* region is euchromatic, it may behave as facultative heterochromatin in TE-rich environments of *australiensis*, *granulata*, and sorghum (Dimitri et al. 2005). Future experiments in these species can focus on investigating the cytological organization of the *Hd1* region, the transcriptional activity of the genes in such TE-rich regions, and association of heterochromatic landmarks such as DNA methylation.

Gene Irregularities in the *Hd1* Region

The *Hd1* in *japonica* is a dual-function gene: It is a floral inducer under short-day (SD) conditions and a floral repressor under long-day (LD) conditions (Yano et al. 2000; Izawa et al. 2002). *Hd1* has been deleted from the *Nhp2–Hd1–kanadaptin* region on chromosome 6 in *glaberrima*. An RFLP mapping study of heading date QTL in a *glaberrima–japonica* backcross identified three significant QTL on chromosomes 1, 6, and 10 (Doi et al. 1998). The QTL on chromosome 10 was later identified to be a major gene *Ehd1*. The *glaberrima* allele of this gene was dominant, conferred early flowering (Doi and Yoshimura 1998; Doi et al. 1998), and functioned independently of *Hd1* (Doi et al. 2004). We propose that in the absence of *Hd1*, *Ehd1* may function to promote flowering time in *glaberrima*. To our knowledge, *Hd1* sequences and their role in other *glaberrima* accessions have not been investigated, without which we cannot confirm this hypothesis.

Interspersed repeats flanking genes often contribute to gene deletion through unequal crossovers between nonallelic interspersed repeats, on sister or non-sister chromatids, followed by breakage and rejoining of the chromatid fragments (Strachan and Read 1999). Shared interspersed repeats are found flanking the *Hd1* gene in the AA genome species indicating that unequal double crossover event may have led to the deletion of *Hd1* in *glaberrima*.

A double-thymidine deletion in *rufipogon* and a single-base transversion (C → A) in *punctata* resulted in a frame-shift and a nonsense mutation, respectively, in the *Hd1* coding sequence. Small deletions (1 bp to a few kb) and single-base substitutions are quite common in grass genomes caused by illegitimate recombination during double-stranded break repair (reviewed in Bennetzen 2007). Although frame-shift deletions and nonsense mutations can produce premature termination codons that generally lead to loss of gene expression, other outcomes such as unstable mRNA that is rapidly degraded by way of nonsense-mediated mRNA decay and production of truncated polypeptide are also possible (Strachan and Read 1999). Splice site conservation is extensive among the *Oryza* genomes as revealed by studies at the *monoculm1* locus (Lu et al. 2009), and plants in general exhibit low levels of

alternative splicing because of widespread whole-genome and segmental duplication events (Kim, Goren, et al. 2008). Therefore, it is quite likely that the *Hd1* gene in *rufipogon* and *punctata* is nonfunctional and other flowering QTLs such as *Hd2–Hd14* (Yano et al. 1997; Lin et al. 1998; Yamamoto et al. 2000; Monna et al. 2002) play roles in determining flowering time in these species. This might be the case in *rufipogon* where the *Hd1* coding sequence was 100% identical to the *Hd1* mRNA from the Kasalath cultivar of *indica*, which has a double-thymidine deletion at an identical site as *rufipogon*. Kasalath has functional *Ehd1* (Doi et al. 2004), and *Hd3a* and *3b* (Monna et al. 2002) genes, all of which promote flowering under SD and LD conditions.

The *Hd1* region in sorghum had seven noncolinear genes (*b–h*) and at least two incidences of gene movement (*EZ1* and *TA1*). Without additional data to cover the phylogenetic gap between rice and sorghum, it is impossible to determine if the noncolinear genes are true orthologs of their rice counterparts and if these genes were inserted in sorghum or deleted from the *Oryza* lineage after the rice–sorghum split. Four noncolinear genes are found in the *orp* region of sorghum and present a similar problem (Ma et al. 2005). The mechanism of single gene movement as observed in the case of *EZ1* and *TA1* is unknown. Nonhomologous illegitimate recombination following gene amplification has been suggested to be the cause (Song et al. 2002) but the results are still unsubstantiated (Devos 2005).

The *SERK1-6a/SEPK1* organization as it appears in *Oryza* and sorghum maybe another case of gene movement. After dating the two rounds of gene duplication of *SEPK1* and comparing them to the rice–sorghum divergence dates, we suggest the following chronology of events. *SEPK1* was present in the rice–sorghum ancestor about 78 My. It underwent duplication events at about 55 and 53 My giving rise to *SERK1-6a* and *SERK1-6b*, respectively. After the rice–sorghum split about 50 My (Stebbins 1981; Wolfe et al. 1987; Gaut et al. 1996), the younger *6b* gene was lost from the *Oryza* lineage and a possible gene movement may have occurred to reverse the order of *SERK1-6a* relative to *SEPK1* in *Oryza* without changing the transcriptional orientation. Neofunctionalization is a process whereby duplicated genes acquire novel functions (Force et al. 1999; Lynch and Conery 2000). *SERK1*, which appears to be under strong purifying selection, may have been retained in both *Oryza* and sorghum as neofunctionalized genes. EST evidence for *SERK1* and FL-cDNA evidence for *SEPK1* are found suggesting that they are actively expressed; however, further work needed at this level is tissue specificity to understand if *SERK1* has truly neofunctionalized.

Oryza Speciation and Regional Differences in Genome Evolution

The Ks and Ka values of the eight genes used to build the *Oryza* phylogeny demonstrated that after the divergence of *punctata*, very few amino acids changed before the radiation within the AA genome species. By applying the fossil-calibrated synonymous rate of *Adh1* divergence (Gaut et al. 1996), the AA divergence time was estimated to be about 2 My, similar to previous reports (Zhu and Ge 2005). Using a concatenated set of 62 genes sampled from the *Oryza* genome, Zou et al. (2008) found that the substitution rate of the BB genome is higher than the CC genome. In our data set, five of the eight genes in *punctata* had higher synonymous changes than *officinalis*, which may have led to the skewed estimates of the time of species radiation suggesting that *officinalis* diverged later than *punctata*. This is not the case, however, as seen from earlier studies (Ge et al. 1999; Ammiraju et al. 2008; Zou et al. 2008). Our results indicate that the divergence date of the *O. sativa* subspecies is about 0.3 My, much earlier than the domestication of crops about 10,000 years ago (reviewed in Kovach et al. 2007; Sang and Ge 2007), and support the multiple independent domestication events hypothesis proposed by earlier investigators studying hybrid sterility (Kato et al. 1928; Zhu et al. 2005), isozyme polymorphisms (Second 1982), indel events in chloroplast genome types (A–H) of rice (Kawakami et al. 2007), and intron sequence divergence of nuclear genes (Zhu and Ge 2005).

The *O. sativa* subspecies have a mean sequence divergence of 18% from *nivara* and 40% from *rufipogon* in the intergenic regions. In conjunction with the phylogenies of shared transposons (fig. 5) and previous reports on the genomic paleontology of cultivated rice (Vitte et al. 2004; Zhu and Ge 2005; Kawakami et al. 2007), our data suggest that *nivara* maybe the possible progenitor of both subspecies of cultivated Asian rice. Although both morphologically and physiologically, *nivara* is more similar to cultivated rice, *rufipogon* is more similar to rice cultivars adapted for growing in deepwater conditions or which have strong photoperiod sensitivity (Sang and Ge 2007). Admixed *nivara* and *rufipogon* populations have been found in Thailand and are thought to be the ancestors of domesticated rice (Sano et al. 1980). Although there is strong evidence that the *O. sativa* subspecies diverged before their domestication, the identity of the progenitor of modern cultivated rice has been confounded by continuous gene flow among rice cultivars and wild populations of *nivara* and *rufipogon*. A comparative sequence study of more cultivated genotypes and different populations of wild *O. sativa* relatives such as made possible by recent advances in high-throughput resequencing strategies (Huang et al. 2009), in addition to removing taxonomical inconsistencies and misidentification in germplasm collections (Aggarwal et al. 1999; Ge et al. 2001; Bao et al. 2005), may help to answer a wide range of biological questions dealing with *O. sativa* ancestry and domestication.

Our conclusions about the *Oryza* genome evolution were similar in many respects to earlier comparative studies at the *Adh1* (Ammiraju et al. 2008) and *monoculm1* (Lu et al. 2009) loci. There were, however, noticeable regional differences. First, the re-annotation of the *japonica* reference sequence using stringent criteria improved the *japonica* annotation quality at all loci with the removal of TE-related gene structures. As opposed to the *Adh1* locus, both the *Hd1* and the *MOC1* loci in the diploid *Oryza* had no

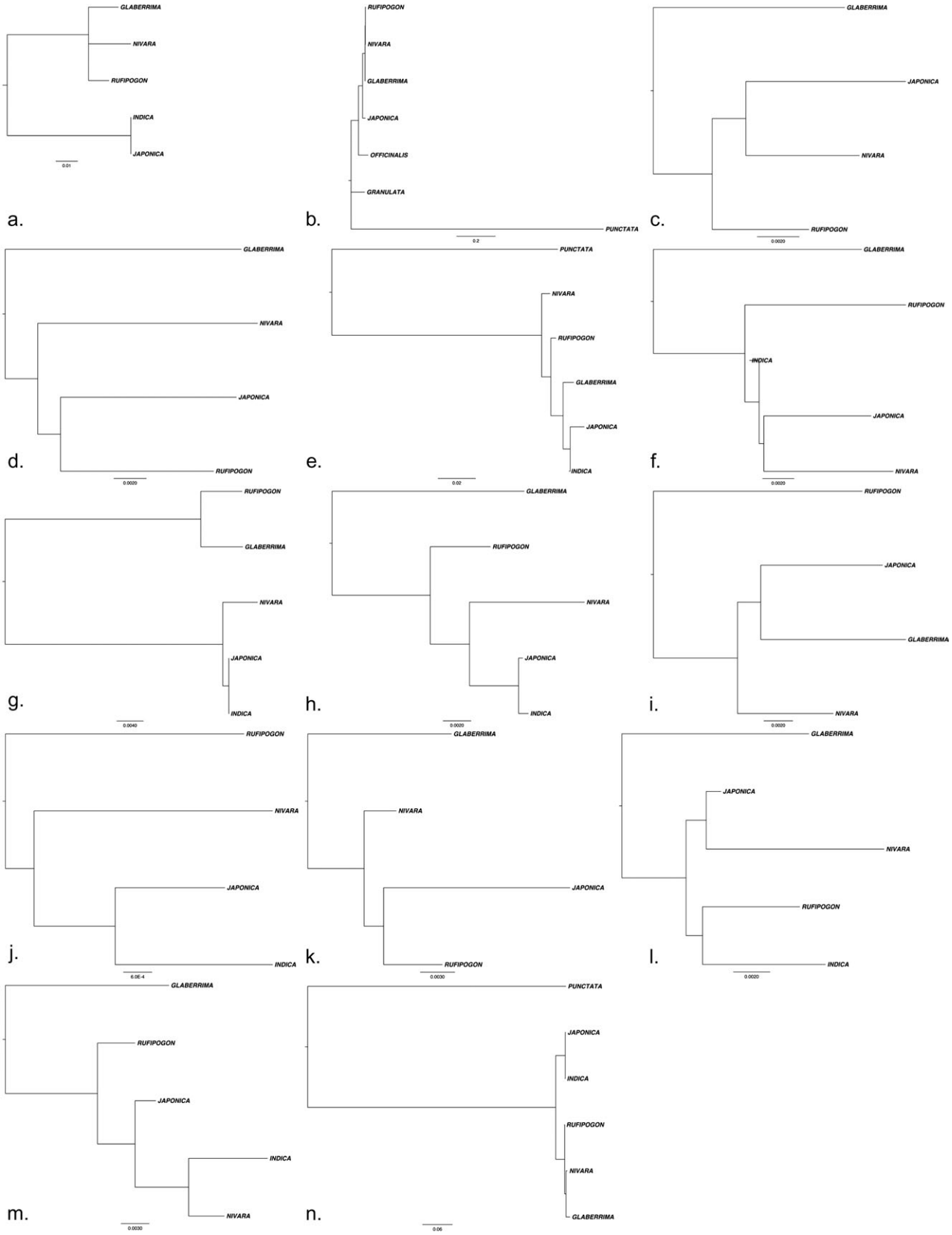

Fig. 5. Distance-based Neighbor-Joining tree phylogenies of selected DNA TEs shared among rice species. (*a*) DTE4, (*b*) DTE8, (*c*) DTE10, (*d*) DTE9, (*e*) DTE18, (*f*) RTE16, (*g*) DTE1, (*h*) DTE3, (*i*) DTE2, (*j*) DTE15, (*k*) RTE3, (*l*) DTE16, (*m*) DTE17, and (*n*) DTE12.

pseudogenes. Pseudogenization is most prevalent in polyploid genomes or in regions of diploid genomes showing extensive tandem gene duplication (Lynch and Conery 2000; Wendel 2000). The *Adh1* locus has eight gene families and 22 apparent pseudogenes across the *Oryza* genomes, whereas at the *MOC1* locus only polyploids showed the presence of pseudogenes. Second, TEs were a central factor in genome size, gene density, and region-specific variation at all loci. At all loci, DNA TEs outnumbered RNA TEs, whereas LTR-RTs dominated the TE classes in terms of size contribution. Locus-specific variation in TE content is evidenced from the fact that on average both *Adh1* and *Hd1* loci had higher total TE content (~46%) than the *MOC1* locus (~30%) across *Oryza*, but the *MOC1* locus had higher gene density (7.8 kb/gene) than the *Adh1* and *Hd1* loci (12.1 kb/gene) among the AA genomes. Third, both the *Hd1* and *MOC1* loci in the wild rice species were highly stable in gene order and transcriptional orientation, with 98% of the genes having *japonica* orthologs at colinear positions. The *Adh1* locus in contrast was very unstable demonstrating lineage-specific variation in gene family copy number and several rearrangement polymorphisms, which may be the consequences of nonhomologous and illegitimate recombination occurring at duplication-rich segments of the genome. Fourth, the *Hd1* gene in *glaberrima*, *rufipogon*, and *punctata* was either deleted or carried disruptive mutations as opposed to the *MOC1* locus where the nonduplicated underlying gene remained intact. Finally, all three regions appeared to be under purifying selection as evidenced by Ka/Ks less than 1. Although the timing of *Oryza* speciation events and the phylogenetic relationships among the species was largely similar, there were topological inconsistencies also confirming lineage-specific and regional variation.

Supplementary Material

Supplementary tables S1–S18 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the National Science Foundation (grants DBI-0321678 and DBI-0227414) for providing funds for this project. We thank the members of the Arizona Genomics Institute and the Purdue Genomics Facility for generating and processing the high-quality sequence data used in this analysis. We also acknowledge the contributions of other members of the OMAP research team and the Jackson Lab at Purdue. We thank Dr Ning Jiang of Michigan State University, Ann Arbor, for providing the comprehensive rice repeat database, which helped to extensively annotate the rice BACs for repeats.

References

Aggarwal RK, Brar DS, Khush GS. 1997. Two new genomes in the Oryza complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Mol Gen Genet.* 254:1–12.

Aggarwal RK, Brar DS, Nandi S, Huang N, Khush GS. 1999. Phylogenetic relationships among Oryza species revealed by AFLP markers. *Theor Appl Genet.* 98:1320–1328.

Albrecht C, Russinova E, Hecht V, Baaijens E, de Vries S. 2005. The Arabidopsis thaliana SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASES1 and 2 control male sporogenesis. *Plant Cell.* 17:3337–3349.

Ammiraju JSS, Lu F, Sanyal A, et al. (19 co-authors). 2008. Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell.* 20:3191–3209.

Ammiraju JSS, Luo M, Goicoechea JL, et al. (27 co-authors). 2006. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res.* 16:140–147.

Ammiraju JSS, Zuccolo A, Yu Y, et al. (15 co-authors). 2007. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. *Plant J.* 52:342–351.

Bao Y, Lu BR, Ge S. 2005. Identification of genomic constitutions of Oryza species with the B and C genomes by the PCR-RFLP method. *Genet Resour Crop Evol.* 52:69–76.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12:1269–1276.

Bennetzen JL. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115:29–36.

Bennetzen JL. 2007. Patterns in grass genome evolution. *Curr Opin Plant Biol.* 10:176–181.

Bennetzen JL, Ma J. 2003. The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr Opin Plant Biol.* 6:128–133.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot.* 95:127–132.

Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33:W451–W454.

Brar DS, Khush GS. 1997. Alien introgression in rice. *Plant Mol Biol.* 35:35–47.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13:721–731.

Chen M, Presting G, Barbazuk WB, et al. (39 co-authors). 2002. An integrated physical and genetic map of the rice genome. *Plant Cell.* 14:537–545.

Chen M, SanMiguel P, Bennetzen JL. 1998. Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* 148:435–443.

Cheng Z, Buell CR, Wing RA, Gu M, Jiang J. 2001. Toward a cytological characterization of the rice genome. *Genome Res.* 11:2133–2141.

Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J. 2001. High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. *Genetics* 157:1749–1757.

Devos KM. 2005. Updating the 'Crop Circle'. *Curr Opin Plant Biol.* 8:155–162.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* 12:1075–1079.

Dimitri P, Corradini N, Rossi F, Verni F. 2005. The paradox of functional heterochromatin. *BioEssays* 27:29–41.

Doi K, Izawa T, Fuse T, Yamanouchi U, Kubo T, Shimatani Z, Yano M, Yoshimura A. 2004. Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1. *Genes Dev.* 18:926–936.

Doi K, Yoshimura A. 1998. RFLP mapping of a gene for heading date in an African rice. *Rice Genet Newslett.* 15:148–149.

Doi K, Yoshimura A, Iwata N. 1998. RFLP mapping and QTL analysis of heading date and pollen sterility using back-cross populations between Oryza sativa L. and Oryza glaberrima Steud. *Breeding Sci.* 48:395–399.

Duan S, Lu B, Li Z, Tong J, Kong J, Yao W, Li S, Zhu Y. 2007. Phylogenetic analysis of AA-genome Oryza species (Poaceae) based on chloroplast, mitochondrial, and nuclear DNA sequences. *Biochem Genet.* 45:113–129.

Felsenstein J. 1985. Confidence-limits on phylogenies—an approach using the bootstrap. *Evolution* 39:783–791.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet.* 3:329–341.

Fitch DHA, Strausbaugh LD. 1993. Low codon bias and high-rates of synonymous substitution in Drosophila hydei and Drosophila melanogaster histone genes. *Mol Biol Evol.* 10:397–413.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Fortune PM, Roulin A, Panaud O. 2008. Horizontal transfer of transposable elements in plants. *Commun Integr Biol.* 1:74–77.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32:W273–W279.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci.* 12:543–548.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci U S A.* 93:10274–10279.

Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc Natl Acad Sci U S A.* 96:14400–14405.

Ge S, Sang T, Lu BR, Hong DY. 2001. Rapid and reliable identification of rice genomes by RFLP analysis of PCR-amplified Adh genes. *Genome* 44:1136–1142.

Gojobori T. 2007. Curated genome annotation of Oryza sativa ssp japonica and comparative genome analysis with Arabidopsis thaliana—the Rice Annotation Project. *Genome Res.* 17:175–183.

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF. 2007. Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (Gossypium). *Plant J.* 50:995–1006.

Grover CE, Yu Y, Wing RA, Paterson AH, Wendel JF. 2008. A phylogenetic analysis of indel dynamics in the cotton genus. *Mol Biol Evol.* 25:1415–1428.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Guo HN, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell.* 15:1143–1158.

Guo Y-L, Ge S. 2005. Molecular phylogeny of Oryzeae (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *Am J Botany.* 92:1548–1558.

Guyot R, Cheng XD, Su Y, Cheng ZK, Schlagenhauf E, Keller B, Ling HQ. 2005. Complex organization and evolution of the tomato pericentromeric region at the FER gene locus. *Plant Physiol.* 138:1205–1215.

Harismendy O, Ng PC, Strausberg RL, et al. (11 co-authors). 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10:R32.

Harushima Y, Yano M, Shomura A, et al. (17 co-authors). 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148:479–494.

Hass-Jacobus BL, Futrell-Griggs M, Abernathy B, et al. (22 co-authors). 2006. Integration of hybridization-based markers (overgos) into physical maps for comparative and evolutionary explorations in the genus Oryza and in Sorghum. *BMC Genomics.* 7:199.

Hecht V, Vielle-Calzada JP, Hartog MV, Schmidt ED, Boutilier K, Grossniklaus U, de Vries SC. 2001. The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR KINASE 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. *Plant Physiol.* 127:803–816.

Higo K, Ugawa Y, Iwamoto M, Korenaga T. 1999. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27:297–300.

Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet.* 4:275–284.

Huang XH, Feng Q, Qian Q, et al. (13 co-authors). 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19:1068–1076.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800.

Izawa T, Oikawa T, Sugiyama N, Tanisaka T, Yano M, Shimamoto K. 2002. Phytochrome mediates the external light signal to repress FT orthologs in photoperiodic flowering of rice. *Genes Dev.* 16:2006–2020.

Jamison DC, Thomas JW, Green ED. 2000. ComboScreen facilitates the multiplex hybridization-based screening of high-density clone arrays. *Bioinformatics* 16:678–684.

Jiang N, Bao ZR, Zhang XY, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* 421:163–167.

Jiang N, Feschotte C, Zhang XY, Wessler SR. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr Opin Plant Biol.* 7:115–119.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 110:462–467.

Kao FI, Cheng YY, Chow TY, Chen HH, Liu SM, Cheng CH, Chung MC. 2006. An integrated map of Oryza sativa L. chromosome 5. *Theor Appl Genet.* 112:891–902.

Kato S, Kosaka H, Hara S. 1928. On the affinity of rice varieties as shown by the fertility of rice plants. *Centr Agric Inst Kyushu Imp Univ.* 2:241–276.

Kawakami S, Ebana K, Nishikawa T, Sato Y, Vaughan DA, Kadowaki K. 2007. Genetic variation in the chloroplast genome suggests multiple domestication of cultivated Asian rice (Oryza sativa L.). *Genome* 50:180–187.

Kellogg EA, Bennetzen JL. 2004. The evolution of nuclear genome structure in seed plants. *Am J Botany.* 91:1709–1725.

Khush GS. 1997. Origin, dispersal, cultivation and variation of rice. *Plant Mol Biol.* 35:25–34.

Kim E, Goren A, Ast G. 2008. Alternative splicing: current perspectives. *BioEssays* 30:38–47.

Kim H, Hurwitz B, Yu Y, et al. (18 co-authors). 2008. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza. *Genome Biol.* 9:R45.

Kim H, San Miguel P, Nelson W, Collura K, Wissotski M, Walling JG, Kim JP, Jackson SA, Soderlund C, Wing RA. 2007. Comparative

physical mapping between Oryza sativa (AA genome type) and O. punctata (BB genome type). *Genetics* 176:379–390.
Kovach MJ, Sweeney MT, McCouch SR. 2007. New insights into the history of rice domestication. *Trends Genet.* 23:578–587.
Lin HX, Yamamoto T, Sasaki T, Yano M. 2000. Characterization and detection of epistatic interactions of 3 QTLs, Hd1, Hd2, and Hd3, controlling heading date in rice using nearly isogenic lines. *Theor Appl Genet.* 101:1021–1028.
Lin SY, Sasaki T, Yano M. 1998. Mapping quantitative trait loci controlling seed dormancy and heading date in rice, Oryza sativa L., using backcross inbred lines. *Theor Appl Genet.* 96:997–1003.
Lu F, Ammiraju JS, Sanyal A, et al. (15 co-authors). 2009. Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. *Proc Natl Acad Sci U S A.* 106:2071–2076.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A.* 101:12404–12410.
Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* 14:860–869.
Ma J, SanMiguel P, Lai J, Messing J, Bennetzen JL. 2005. DNA rearrangement in orthologous orp regions of the maize, rice and sorghum genomes. *Genetics* 170:1209–1220.
Ma J, Wing RA, Bennetzen JL, Jackson SA. 2007. Evolutionary history and positional shift of a rice centromere. *Genetics* 177:1217–1220.
Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133–141.
McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19:362–367.
Monna L, Lin X, Kojima S, Sasaki T, Yano M. 2002. Genetic dissection of a genomic region for a quantitative trait locus, Hd3, into two loci, Hd3a and Hd3b, controlling heading date in rice. *Theor Appl Genet.* 104:772–778.
Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp japonica. *Genes Genet Syst.* 83:321–329.
Paterson AH, Bowers JE, Bruggmann R, et al. (45 co-authors). 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* 457:551–556.
Piegu B, Guyot R, Picault N, et al. (11 co-authors). 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Res.* 16:1262–1269.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* 162:1389–1400.
Ramakrishna W, Emberton J, SanMiguel P, Ogden M, Llaca V, Messing J, Bennetzen JL. 2002. Comparative sequence analysis of the sorghum Rph region and the maize Rp1 resistance gene complex. *Plant Physiol.* 130:1728–1738.
Rensink WA, Buell CR. 2005. Microarray expression profiling resources for plant genomics. *Trends Plant Sci.* 10:603–609.
Rice Annotation Project Database (update). 2008. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36:D1028–D1033.
Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* 10:516–522.
Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.
Sang T, Ge S. 2007. The puzzle of rice domestication. *J Integr Plant Biol.* 49:760–768.
SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet.* 20:43–45.
SanMiguel P, Tikhonov A, Jin YK, et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–768.
Sano Y, Morishima H, Oka HI. 1980. Intermediate perennial-annual populations of Oryza-Perennis found in Thailand and their evolutionary significance. *Bot Mag Tokyo.* 93:291–305.
Sasaki T, Burr B. 2000. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr Opin Plant Biol.* 3:138–141.
Second G. 1982. Origin of the genic diversity of cultivated rice (Oryza spp.): study of the polymorphism scored at 40 isozyme loci. *Jpn J Genet.* 57:25–57.
Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol.* 4:222–230.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol.* 26:1135–1145.
Song R, Llaca V, Messing J. 2002. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* 12:1549–1555.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167:GC1–GC10.
Stebbins GL. 1981. Coevolution of grasses and herbivores. *Ann MO Bot Gard.* 68:75–86.
Strachan T, Read AP. 1999. Instability of the human genome: mutation and DNA repair. In: Kingston F, editor. Human molecular genetics. New York: Wiley-Liss. p. 209–240.
Swofford D. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A. 2000. The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell.* 12:381–391.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc Natl Acad Sci U S A.* 96:7409–7414.
Turcotte K, Srinivasan S, Bureau T. 2001. Survey of transposable elements from rice genomic sequences. *Plant J.* 25:169–179.
Uozu S, Ikehashi H, Ohmido N, Ohtsubo H, Ohtsubo E, Fukui K. 1997. Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus Oryza. *Plant Mol Biol.* 35:791–799.
Usuka J, Zhu W, Brendel V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* 16:203–211.
van Daelen RA, Gerbens F, van Ruissen F, Aarts J, Hontelez J, Zabel P. 1993. Long-range physical maps of two loci (Aps-1 and GP79) flanking the root-knot nematode resistance gene (Mi) near the centromere of tomato chromosome 6. *Plant Mol Biol.* 23:185–192.

Varshney RK, Nayak SN, May GD, Jackson SA. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27:522–530.

Vaughan DA, Ge S, Kaga A, Tomooka N. 2008. *Phylogeny and biogeography of the genus Oryza*. In: Hirano H-Y, Hirai A, Sano Y, Sasaki T, editors. Rice biology in the genomics era. Springer Berlin Heidelberg. p. 219–234.

Vaughan DA, Morishima H, Kadowaki K. 2003. Diversity in the Oryza genus. *Curr Opin Plant Biol.* 6:139–146.

Vitte C, Ishii T, Lamy F, Brar D, Panaud O. 2004. Genomic paleontology provides evidence for two distinct origins of Asian rice (Oryza sativa L.). *Mol Genet Genomics.* 272:504–511.

Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J, Siepel A, Tanksley SD. 2008. Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. *Genetics* 180:391–408.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol Biol.* 42:225–249.

Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31:3537–3539.

Wing RA, Ammiraju JS, Luo M, et al. (17 co-authors). 2005. The oryza map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol Biol.* 59:53–62.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci U S A.* 84:9054–9058.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35:W265–W268.

Yamamoto T, Kuboki Y, Lin SY, Sasaki T, Yano M. 1998. Fine mapping of quantitative trait loci Hd-1, Hd-2 and Hd-3, controlling heading date of rice, as single Mendelian factors. *Theor Appl Genet.* 97:37–44.

Yamamoto T, Lin H, Sasaki T, Yano M. 2000. Identification of heading date quantitative trait locus Hd6 and characterization of its epistatic interactions with Hd2 in rice using advanced backcross progeny. *Genetics* 154:885–891.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yano M, Harushima Y, Nagamura Y, Kurata N, Minobe Y, Sasaki T. 1997. Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *Theor Appl Genet.* 95:1025–1032.

Yano M, Katayose Y, Ashikari M, et al. (11 co-authors). 2000. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. *Plant Cell.* 12:2473–2484.

Zhang S, Gu YQ, Singh J, Coleman-Derr D, Brar DS, Jiang N, Lemaux PG. 2007. New insights into Oryza genome evolution: high gene colinearity and differential retrotransposon amplification. *Plant Mol Biol.* 64:589–600.

Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. *New Phytol.* 167:249–265.

Zhu S, Wang C, Zheng T, Zhao Z, Ikehashi H, Wan J. 2005. A new gene located on chromosome 2 causing hybrid sterility in a remote cross of rice. *Plant Breeding.* 124:440–445.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* 9:R49.

Zuccolo A, Ammiraju SSJ, HyeRan K, Sanyal A, Jackson S, Wing R. 2008. Rapid and differential proliferation of the Ty3-gypsy LTR retrotransposon Atlantys in the genus Oryza. *RICE.* 1:85–99.

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA. 2007. Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC Evol Biol.* 7:152.