

## Analysis of collinear regions of *Oryza* AA and CC genomes

Qi Feng<sup>a, b</sup>, Tao Huang<sup>a</sup>, Qiang Zhao<sup>a, b</sup>, Jingjie Zhu<sup>a, b</sup>, Zhixin Lin<sup>a</sup>, Bin Han<sup>b, \*</sup>

<sup>a</sup> College of Life Science & Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

<sup>b</sup> National Center for Gene Research & Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China

Received for publication 3 June 2009; revised 25 September 2009; accepted 12 October 2009

### Abstract

Comparative analyses of genome structure and sequence of closely related species have yielded insights into the evolution and function of plant genomes. A total of 103,844 BAC end sequences delegated ~73.8 Mb of *O. officinalis* that belongs to the CC genome type of the rice genus *Oryza* were obtained and compared with the genome sequences of rice cultivar, *O. sativa* ssp. *japonica* cv. Nipponbare. We found that more than 45% of *O. officinalis* genome consists of repeat sequences, which is higher than that of Nipponbare cultivar. To further investigate the evolutionary divergence of AA and CC genomes, two BAC-contigs of *O. officinalis* were compared with the collinear genomic regions of Nipponbare. Of 57 genes predicted in the AA genome orthologous regions, 39 had orthologs in the regions of the CC genome. Alignment of the orthologous regions indicated that the CC genome has undergone expansion in both genic and intergenic regions through primarily retroelement insertion. Particularly, the density of RNA transposable elements was 17.95% and 1.78% in *O. officinalis* and *O. sativa*, respectively. This explains why the orthologous region is about 100 kb longer in the CC genome in comparison to the AA genome.

**Keywords:** *Oryza sativa*; *Oryza officinalis*; retroelement; orthologous

### Introduction

Comparative analysis of plant genomes has yielded important insights into the evolution of genome size, organization, and synteny (Bennetzen, 2007; Tang et al., 2008). The rice genus, *Oryza*, has become a useful system for investigating genome evolution at the interspecific level. The genus *Oryza* consists of more than 20 wild and two cultivated species, which were recognized in ten types of genomes, AA, BB, BBCC, CC, CCDD, EE, FF, GG, HHJJ and HHHK (Vaughan, 1994; Ge et al., 1999). A wealth of genomic resources has been developed for species representing all genome types through the *Oryza* Map

Alignment Project (OMAP, [www.omap.org](http://www.omap.org); Wing et al., 2005).

In this study, we conducted a comparative genomic analysis between the AA and CC genomes. The AA and CC genomes are closely related and have species found in Asia (Vaughan, 1994; Zou et al., 2008). *Oryza officinalis* is an Asian CC genome species with the estimated genome size of ~651 Mb (Kim et al., 2008), which is larger than ~430 Mb of cultivated rice, *O. sativa* of the AA genome (Arumuganathan et al., 1991; Sasaki et al., 2000; Feng et al., 2002). From the previous analyses of the OMAP data, *Oryza* comparative genomics has provided insights into rice genome evolution (Kim et al., 2007; Ma et al., 2007; Zhang et al., 2007) and genome size variation (Piegu et al., 2006; Ammiraju et al., 2007; Lu et al., 2009). Here we generated sequences of the BAC ends and two contigs of four BACs from *O. officinalis* and compared them with the

\* Corresponding author. Fax: +86-21-6482 5775.

E-mail address: [bhan@negr.ac.cn](mailto:bhan@negr.ac.cn)

genome sequences of *O. sativa* ssp. *japonica* cv. Nipponbare (International Rice Genome Sequencing Project, 2005).

To address the question why the CC genome size is larger than the AA genome, we selected two collinear contigs which consist of four *O. officinalis* BAC clones according to *O. sativa* ssp. *japonica* reference genome to perform accurate sequencing. By sequencing the collinear regions, we gained clues as to the nature of genome expansion and divergence of these two *Oryza* species. The total length of two orthologous regions on *O. officinalis* and *O. sativa* were 429 kb and 323 kb, respectively. Comparative analyses reveal that gene content and orientation are moderately conserved between these two genome types. Presumably, the accumulated discrepancy of intron and exon size should be a conceivable reason for CC genome obese. Our results also suggest that retrotransposons play a major role in genomes size expansion. Furthermore, the collinearity between *O. officinalis* and *O. sativa* demonstrates the possibility to use comparative genomics to explore the genetic diversity of wild species for rice improvement (Zhang et al., 2007).

## Materials and methods

### *OO\_Ba\_officinalis* BAC end sequencing and analysis

*OO\_Ba\_officinalis* BAC library is one of the deep-coverage, large-insert BAC libraries developed from 12 sampled *Oryza* species for The *Oryza* Map Alignment Project (OMAP; Luo and Wing, 2003). The average insert size of the library was ~141 kb. We combined 101,091 BAC ends from OMAP community with ~50,000 BAC ends from our lab and generated 103,844 individual end sequence tags. The 5' and 3' end DNA sequences of *O. officinalis* BAC clones were determined using vector-specific universal primers (SP6 and T7) by ABI 3730xl capillary sequencers. Base calling of the ABI DNA analyzer was confirmed by analyzing the trace files with the Phred-Phrap software. Repeat sequences were identified with the RepeatMasker (<http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>) program. The end sequences were searched against those of *Oryza sativa* ssp. *japonica* cv. Nipponbare using BLASTN with the filter parameter of 1e-10.

### Contig identification

Based on the comparative mapping by annotation and

sequence similarity, we picked ~20 *O. officinalis* BAC clones whose ends were highly similar (> 95%) to the *O. sativa* ssp. *japonica* cv. Nipponbare chromosome 4 reference genome sequence. These BACs were digested with *Hind* III enzyme and loaded on 1% agarose gel at 30 voltages for overnight. DNA fragments were then transferred to nitrocellulose membrane for Southern blotting. We used *O. sativa* ssp. *japonica* cv. Nipponbare genomic DNA as probe for hybridization. The hybridization was conducted according to a standard ECL (Enhanced Chemiluminescence) DNA filter hybridization protocol at moderate stringency (60°C, 0.5 × SSC). Four BAC clones were identified according to strong signals, suggesting that these clones have high similarity between *O. sativa* ssp. *japonica* cv. Nipponbare and *O. officinalis* (data not shown). These 4 BACs were assembled into two contigs each containing two overlapped BAC clones, with a length of 192,320 bp and 226,644 bp.

### BAC sequencing, assembling, and annotation

We used conventional BAC shot-gun sequencing strategy to construct sub-clone libraries, sequence, assemble, and gap-fill. For annotation, the assembled CC-genome BACs were uploaded to RiceGAAS (Rice Genome Automated Annotation System, <http://ricegaas.dna.affrc.go.jp/>) annotation system. This system integrated GENSCAN, RiceHMM, FGENESH, MZEF programs for prediction and analysis of protein-coding gene structure, SplicePredictor for splice site prediction, Blast, HMMER, ProfileScan, MOTIF for homology search, tRNAscan-SE for tRNA gene prediction, RepeatMasker and Printrepeats for repetitive DNA analysis, Signal Scan for signal scan search, PSORT for protein localization site prediction, SOSUI for classification and secondary structure prediction of membrane proteins, and Blast against full-length cDNA sequences of *indica* and *japonica* cultivars. All gene predictions were checked manually. Regions collinear that those of *O. sativa* ssp. *japonica* cv. Nipponbare reference genomes were annotated with the RiceGAAS. Synonymous and nonsynonymous substitutions were calculated using tools available at <http://services.cbu.uib.no/tools/kaks>.

## Results and discussion

### BAC end sequence analysis

When the 103,844 BAC end sequences delegated ~73.8

Mb of *O. officinalis* genome sequences were analyzed by the RepeatMasker program, more than 33.5 Mb were masked as repeat sequences. This suggests that perhaps more than 45% of *O. officinalis* genome consists of repeat sequences. This proportion of repeat sequences is higher than that of *O. sativa* ssp. *japonica* cv. Nipponbare calculated at 38.87%. Among the repeat elements of *O. officinalis*, interspersed repeats are most abundant, accounting for nearly 99% of all repeat sequences. Of the BAC end sequences, *O. officinalis* has 37.06% of retroelements and 7.49% of DNA transposons. Among the retroelements, the most abundant class is long terminal repeat (LTR) elements, 36.37%; followed by LINE, 0.59% and SINE, 0.10% (Table 1). In comparison, the Nipponbare genome contains 20.43% of retroelements and 11.38% of DNA transposons (Mao et al., 2000; Bruggmann et al., 2006). The accumulation of retroelements could have been a major reason for the larger genome size of *O. officinalis* than *O. sativa* (Bruggmann et al., 2006).

When sequences of 103,844 *O. officinalis* BAC ends were compared with the genome sequence of *O. sativa* ssp. *japonica* cv. Nipponbare, 90,037 were matched to the genome sequence. Of these, 20,796 (or 10,398 pairs) ends were optimally mate-paired with the genome sequence. Among them, 3,958 BAC clones had both ends matched the Nipponbare genome sequence with 50–200 kb distance. The average distance of each pair of both ends was 114.6 kb, i.e., each *O. officinalis* BAC clone spanned an average of 114.6 kb in the Nipponbare genome. This is shorter than the average BAC insert length of 141 kb estimated by Wing et al. (2005), which might be due to a larger genome size of *O. officinalis* or shorter than average length of the 3,958 BACs. Testing these alternatives may provide insights into size and structure differences between the AA and CC genomes of *Oryza*. Thus we sequenced two contigs of *O. officinalis* BACs and compared them with the collinear regions of the Nipponbare genome.

Table 1  
Repeated sequence summary of 103,844 *O. officinalis* BAC ends

Classification	Number of elements	Length occupied (bp)	Sequence (%)
Retroelements	61,999	27,346,063	37.06
SINEs	570	73,321	0.1
LINEs	1,113	434,783	0.59
L1/CIN4	1,113	434,783	0.59
LTR elements	60,316	26,837,959	36.37
Ty1/Copia	7,637	3,633,248	4.92
Gypsy/DIRS1	48,791	21,710,272	29.42
DNA transposons	20,519	5,526,796	7.49
hobo-Activator	1,317	315,098	0.43
Tc1-IS630-Pogo	2,934	490,394	0.66
En-Spm	6,201	2,611,632	3.54
MuDR-IS905	2,854	872,217	1.18
Tourist/Harbinger	3,593	630,676	0.86
Unclassified	645	205,205	0.28
Total interspersed repeats		33,078,064	44.83
Small RNA	203	41,316	0.06
Satellites	74	13,952	0.02
Simple repeats	3,370	180,447	0.25
Low complexity	5,634	244,401	0.33

### Comparison of repeat sequences between *O. officinalis* contigs with the collinear region of Nipponbare

To compare genome structures between *O. officinalis* and *O. sativa*, we assembled two contigs each containing two BACs of *O. officinalis*, named as contig 1 spanning clones, OO\_Ba0013j05 and OO\_Ba0033a15 (EMBL accession no. FP565616), and contig 2 spanning clones, OO\_Ba0005110 and OO\_Ba0081k17 (EMBL accession no. FP565615). Contigs 1 and 2 are 192,320 and 226,644 bp in length, respectively. Based on sequence similarity search, the total of ~419 kb sequences of both contigs are potentially orthologous to 323 kb of Nipponbare genome sequences. A number of features were compared between the contigs and their collinear regions of Nipponbare.

A comparison of repeat elements in these two regions reveals that more than 25.92% of *O. officinalis* sequences are repeat sequences whereas 14.88% of the Nipponbare sequences are repeats (Table 2). The density of DNA TEs in *O. sativa* and *O. officinalis* was 9.66% and 5.99%, respectively. However, the density of RNA TEs was 1.78% in *O. sativa* and 17.95% in *O. officinalis* (Table 3). The genome fraction of RNA TEs in *O. officinalis* was 10 folds

larger than in *O. sativa*. This finding from the comparison of the two collinear regions indicates that RNA TEs explains most of genome size difference between *O. sativa* and *O. officinalis*.

### Comparison of gene composition and collinearity

In both *O. officinalis* contigs of ~419 kb in total, 83 open reading frames were identified (data not shown), resulting in a gene density of one gene per 4.9 kb (Table 4). The majority of annotated genes were supported by *O. sativa* ESTs or full-length cDNAs.

To be consistent for comparison, we re-annotated the orthologous Nipponbare sequences using the same procedures as we adopted for *O. officinalis*. Of 57 predicted genes of Nipponbare, 39 were collinear with those of *O. officinalis*, indicating moderately conserved gene collinearity between AA and CC genomes of *Oryza* (Fig. 1). In addition, the lower than 1 *Ka/Ks* ratios estimated for the vast majority of the orthologous pairs indicate that purifying selection was dominant following the divergence between CC and AA genomes (Yang et al., 2000; Zhang et al., 2006).

Table 2  
Comparison of repeat elements between *O. sativa* and *O. officinalis* orthologous region

Classification	Number of elements		Percentage of sequence (%)	
	<i>O. sativa</i>	<i>O. officinalis</i>	<i>O. sativa</i>	<i>O. officinalis</i>
Retroelements	14	77	1.78	17.95
SINEs	4	1	0.23	0.06
LINEs	5	3	0.35	0.52
L1/CIN4	5	3	0.35	0.52
LTR elements	5	73	1.20	17.37
Ty1/Copia	1	13	0.25	3.57
Gypsy/DIRS1	4	57	0.94	13.20
DNA transposons	154	116	9.66	5.99
hobo-Activator	7	7	0.37	0.29
Tc1-IS630-Pogo	39	16	2.00	0.62
En-Spm	13	11	0.73	0.44
MuDR-IS905	11	12	1.23	1.22
Tourist/Harbinger	52	41	3.41	2.12
Unclassified	10	5	1.97	0.62
Total interspersed repeats			13.41	24.56
Satellites	0	1	0.00	0.09
Simple repeats	47	49	0.90	0.72
Low complexity	46	54	0.68	0.55

Table 3  
TE composition in two orthologous regions

<i>Oryza</i> species	TE type	Genome fraction (%)	Copy no.	Mean size of TEs (bp)	TE density (TEs/10 kb)
<i>O. officinalis</i>	DNA	6.0	116	216	2.8
	RNA	17.9	87	864	2.1
	Total	23.9	203	494	4.8
<i>O. sativa</i>	DNA	9.7	154	203	4.8
	RNA	1.8	14	411	0.4
	Total	11.5	168	220	5.2

Table 4  
Statistics of gene models for *O. sativa* and *O. officinalis*

Features	<i>O. sativa</i>	<i>O. officinalis</i>
Collinear region (bp)	323,436	418,964
Number of predicted genes	57	70*
Number of predicted exons	295	358
Average number of exons per gene	5.3	4.3
Average intron size (bp)	377	416
Average exon size (bp)	273	274
Average gene size (kb)	3.1	2.6
Average gene density (kb/gene)	5.8	4.9
Intergenic percent in sequenced region (%)	47.2	47.5
G+C content of predicted genes (%)	55.19	59.07

\*TE related genes are not included.

Of the collinear genes, two sets of them do not have the one-to-one orthologous relationship. In contig 1, *Gene27* and *Gene28* of *O. sativa* are related to the same gene, *Gene36*, of *O. officinalis* (Fig. 1A). From the detailed gene predicted information, we observed that the gene content showed little differentiation. The gene encodes a Putative LanC-like protein3. In contig 2, *Gene22* of *O. sativa* has two orthologous genes, *Gene31* and *Gene36*, of *O. officinalis* (Fig. 1B). This gene encodes a Putative F-box domain containing protein. The F-box gene family constitutes one of the largest families in the rice genome, with nearly 687 members distributed across all 12 chromosomes. The F-box family expanded quite recently in the rice genome, predominantly by localized tandem duplications (Jain et al., 2007; Ammiraju et al., 2008).

#### Conserved gene coding region and diverged intergenic and CC genome specific region

Statistics of gene models for *O. sativa* and *O. officinalis* (Table 4) indicated that there was not significant difference

in gene content between them. The average gene size of *O. sativa* is larger than that of *O. officinalis*, whereas its gene density is smaller. In addition, *O. sativa* has larger average number of exons per gene. On average, the mean gene intron size shows more than 10% difference between the sequenced collinear regions of *O. officinalis* and *O. sativa*, while their average exon sizes are nearly the same (Table 4). The accumulated discrepancy of intron sizes may be another reason for the larger size of the CC genome than the AA genome.

Based on sequence alignment analysis by using VISTA (Mayor et al., 2000), we observed that sequence conservation was largely maintained in orthologous genic regions, more diverged in intergenic regions and *O. officinalis* specific gene regions (Fig. 2). It suggested that the major reason for CC genome being much larger than AA genome should come from the intergenic region and *O. officinalis* specific genes. We also found that 18 of *O. officinalis* genes had no hits to the previously predicted genes of *O. sativa* genome sequences, nearly half of which encoded retrotransposon protein or reverse transcriptase (Table 5).

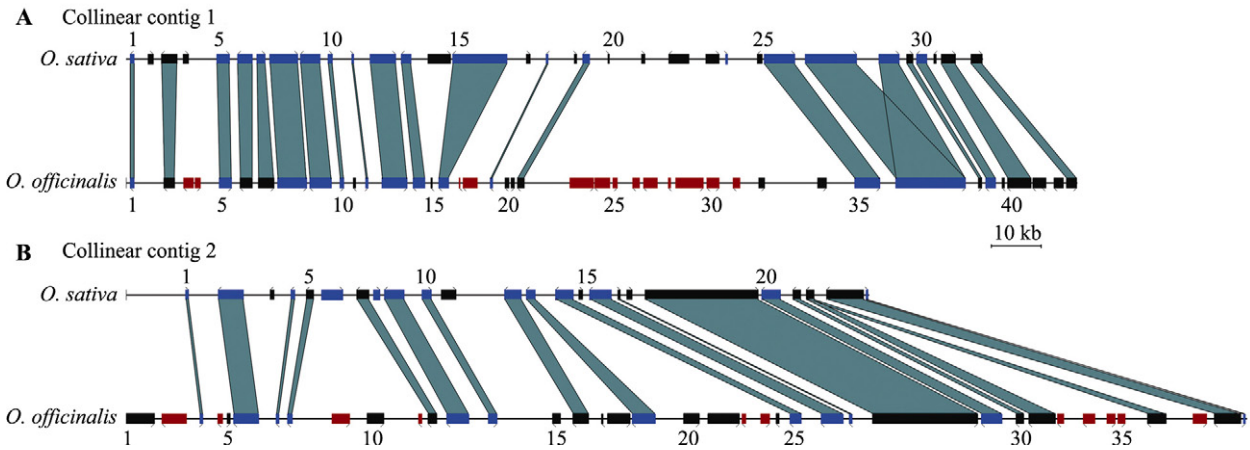


Fig. 1. Gene collinearity in orthologous regions of *O. officinalis* and *O. sativa*. Two sequenced regions of *O. officinalis* are in length of 192 kb and 226 kb. Their orthologous regions in *O. sativa* ssp. *japonica* are in length of 173 kb and 150 kb, respectively. Genes are indicated as filled boxes, slashes on the corner of filled boxes show the orientation of predicted genes, left slashes indicate the gene orientation is minus and right slashes indicate the gene orientation is plus; bold lines between filled boxes represent intergenic regions. Genes with EST and/or full-length cDNA matched are represented by blue boxes. *O. officinalis* extra genes are shown with red boxes; most of these genes are coded retrotransposon protein or reverse transcriptase. The black color boxes of genes on *O. sativa* represented no hits on *O. officinalis* and the black boxes of genes on *O. officinalis* presented no hits on *O. sativa* chromosome 4 or hit on *O. sativa* other chromosomes. Gene number is indicated at every five genes. Orthologous genes are connected by dotted lines and filled with dark-green. **A:** *Gene27* and *Gene28* of *O. sativa* in contig1 related to the same *Gene36* of *O. officinalis*. **B:** *Gene22* of *O. sativa* in contig 2 related to two orthologous genes, *Gene31* and *Gene36*, on *O. officinalis*. The bar represents a 10 kb segment.

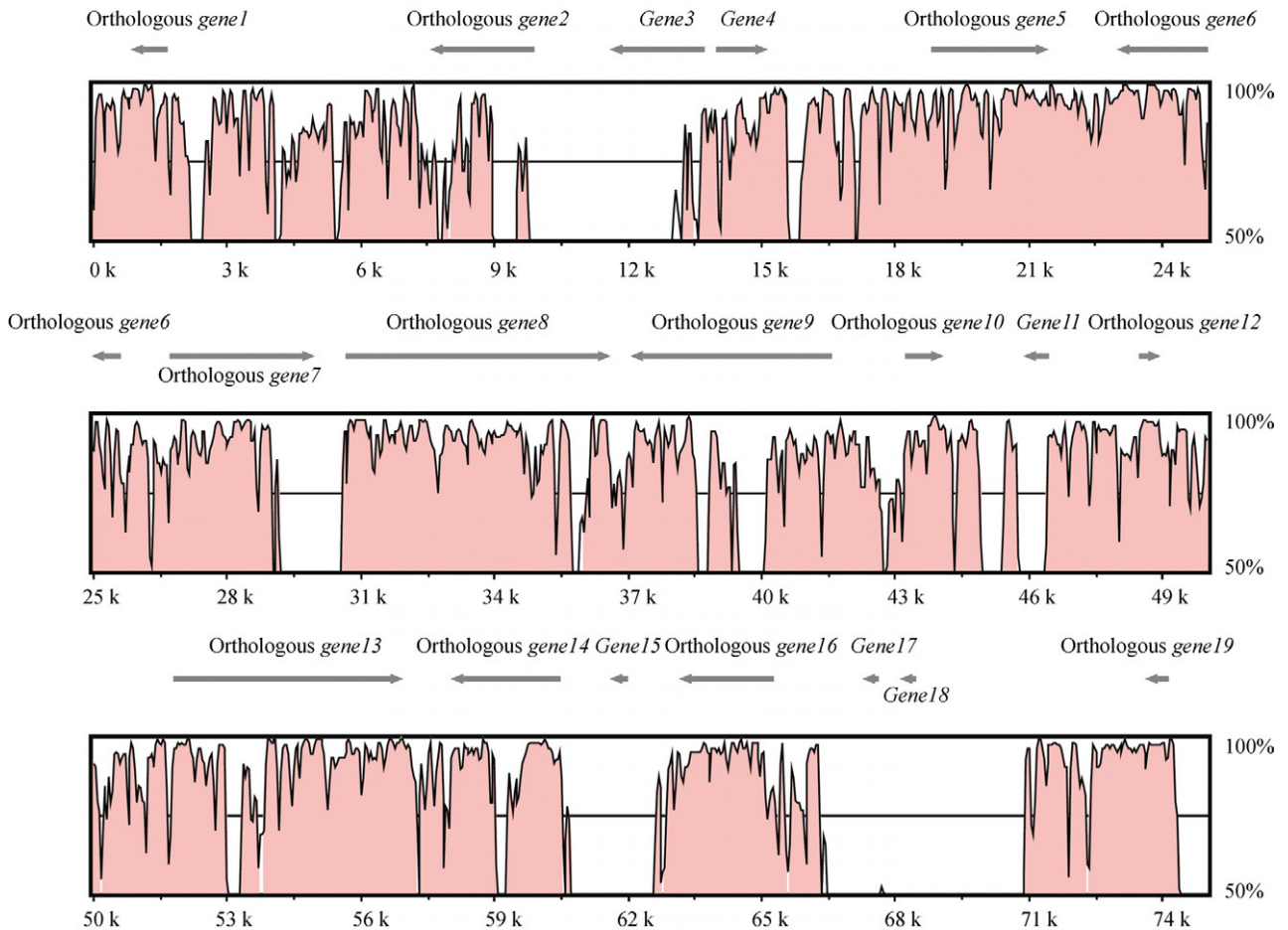


Fig. 2. VISTA analysis of orthologous regions of *O. officinalis* and *O. sativa*. Genes identified in two *Oryza* genera genomes are denoted by numbers, and orthologous genes are designated with signalment. Arrows indicate transcription orientation. X-axis is *O. officinalis* genome sequence and Y-axis shows the similarity of *O. officinalis* (192 kb) with *O. sativa* (173 kb). Sequence conservation was largely maintained in orthologous genic regions, more diverged in intergenic regions and *O. officinalis* extra gene regions.

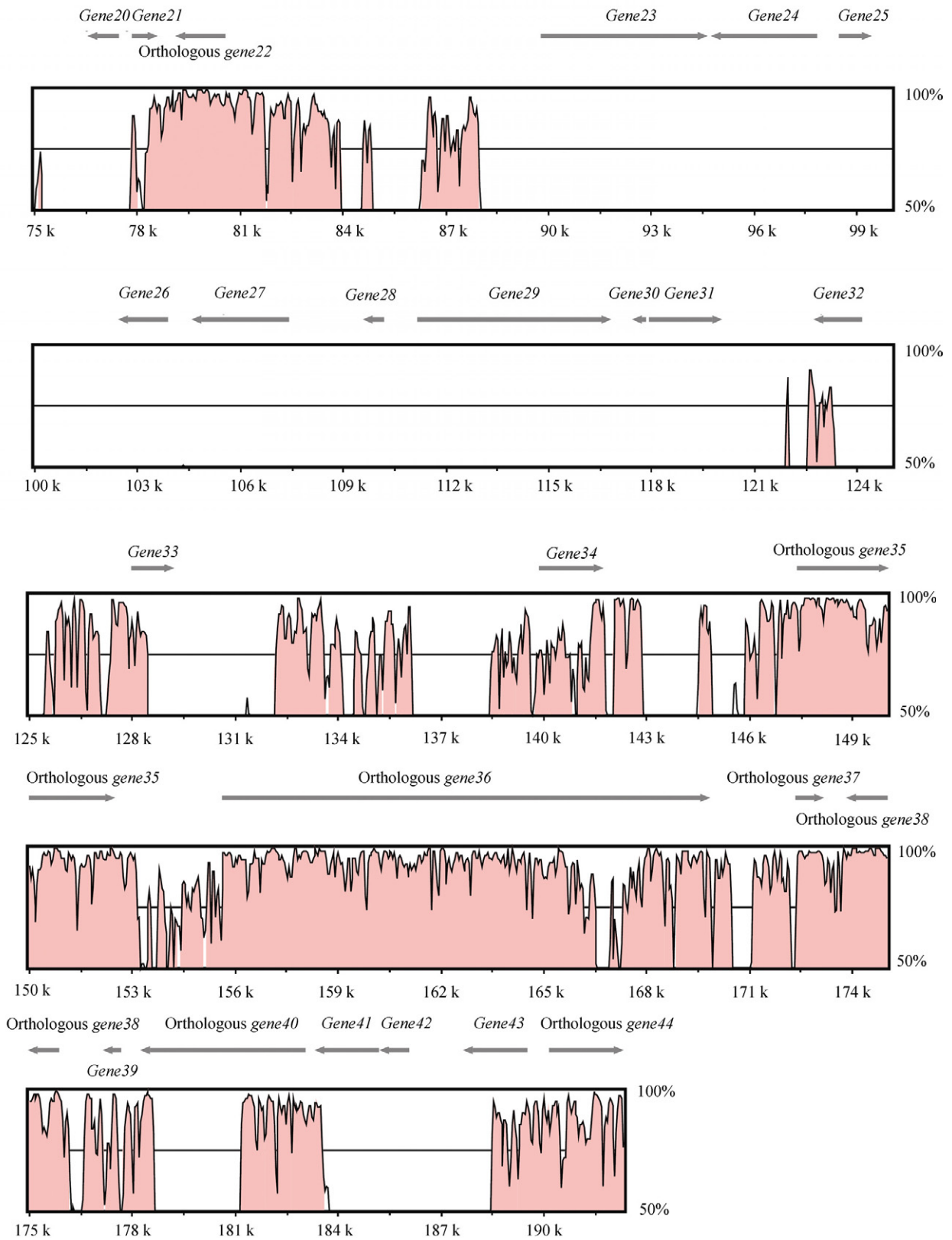


Fig. 2. Continued.

Table 5  
Gene characterization for AA and CC genome

Gene type	Contig	CC gene no.	AA gene no.	Predicted gene function	EST	Ka	Ks	Ka/Ks		
Orthologue gene in Chr. 4 collinear region	Contig 1	1	1	Putative GEM-like protein 6	Yes	0.012	0.041	0.293		
		2	3	Putative cyanophycin synthetase	No	0.211	0.223	0.948		
		5	5	Putative mitochondrial intermembrane space import and assembly protein 40	Yes	0.009	0.035	0.256		
		6	6	Putative mitochondrial substrate carrier family protein	Yes	0.029	0.053	0.556		
		7	7	Mitochondrial import inner membrane translocase	Yes	0.007	0.033	0.217		
		8	8	Putative membrane protein MP28	Yes	0.020	0.050	0.404		
		9	9	Putative protein involved in mRNA turnover and stability (ISS)	Yes	0.018	0.035	0.525		
		10	10	Putative beta-ketothiolase	Yes	0.018	0.042	0.418		
		12	11	Unknown protein similar to Os04g0528200	Yes	0.006	0.028	0.229		
		13	12	Putative white-brown complex homolog protein 3	Yes	0.001	0.039	0.017		
		14	13	Putative checkpoint protein HUS1 (mHUS1)	Yes	0.005	0.023	0.200		
		16	15	Putative cell wall protein AWA1 precursor	Yes	0.015	0.052	0.293		
		19	17	Putative ferredoxin-thioredoxin reductase	Yes	0.010	0.018	0.530		
		22	19	Putative cadmium-induced protein	Yes	0.007	0.029	0.239		
		35	26	Putative WD repeat-containing protein 44	Yes	0.039	0.069	0.567		
		36	27(28)	Putative conjugal transfer protein	Yes	0.016	0.025	0.635		
		37	29	Putative lactosylceramide 4-alpha-galactosyltransferase	No	0.003	0.017	0.177		
		38	30	Putative arabinose-proton symporter	Yes	0.002	0.027	0.092		
		40	32	GRP Glycine rich protein family	No	0.107	0.152	0.706		
		44	33	Putative expansin-B17 precursor	No	0.008	0.062	0.121		
		Contig 2	Contig 2	3	1	Pseudo-response regulator 5	Yes	0.008	0.066	0.123
				6	2	Putative optimisation of the surface electrostatics	Yes	0.021	0.035	0.606
				7	4	Putative crystal structure of mavicyanin	Yes	0.020	0.043	0.471
				8	5	Unknown protein similar to Os04g0674000	Yes	0.014	0.030	0.464
				12	7	Putative binding	Yes	0.014	0.036	0.387
				13	9	Putative root phototropism protein 3	Yes	0.012	0.052	0.226
				14	10	Putative protein DRE2	Yes	0.008	0.023	0.345
				16	12	Putative YSL transporter 1	Yes	0.007	0.047	0.147
				19	13	Putative AMP-binding protein	Yes	0.012	0.050	0.244
				25	14	Putative endoglucanase 13	Yes	0.013	0.036	0.363
				26	16	Putative protein involved in mRNA turnover and stability	Yes	0.019	0.032	0.588
				27	17	Unknown protein	Yes	0.320	0.486	0.658
				28	19	Putative RecName: full=sacsin	Yes	0.008	0.026	0.303
29	20			Oligosaccharyl transferase subunit STT3B	Yes	0.004	0.049	0.080		
30	21			PPR repeat/Tetratricopeptide repeat	No	0.010	0.047	0.214		
31(36)	22			Putative F-box domain containing protein	No	0.084	0.201	0.421		
38	23	Putative ARC6H	No	0.010	0.040	0.249				
39	24	Putative thioredoxin	Yes	0.041	0.032	1.311				

(to be continued on the next page)

Table 5 (Continued)

Gene type	Contig	CC gene no.	AA gene no.	Predicted gene function	EST	Ka	Ks	Ka/Ks	
Orthologue gene in noncollinear region	Contig 1	15	Chr. 6	Hypothetical protein	Yes	0.374	0.450	0.833	
		20	Chr. 4	Hypothetical protein	No	0.495	0.530	0.933	
		23	Chr. 11	Hypothetical protein	No	0.053	0.238	0.223	
		24	Chr. 11	Hypothetical protein similar to aminotransferase-like protein	No	0.064	0.268	0.240	
		25	Chr. 11	Hypothetical protein	No	Null	Null	Null	
		27	Chr. 11	Hypothetical protein	No	Null	Null	Null	
		28	Chr. 2	Repeat-seq (multiple-match with low similarity)/putative gag-pol protein	No	Null	Null	Null	
		42	Chr. 6	Unknown protein	No	0.030	0.092	0.323	
		43	Chr. 4	Putative MutA	No	0.153	0.358	0.427	
		Contig 2	18	Chr. 3	Putative transposon protein Pong sub-class	No	0.155	0.437	0.356
			21	Chr. 11	Putative reverse transcriptase domain containing protein	Yes	0.024	0.167	0.147
			22	Chr. 11	Putative microtubule binding protein D-CLIP-190	No	Null	Null	Null
			23	Chr. 11	Putative peptidase	No	0.084	0.105	0.801
			24	Chr. 11	Putative nuclear envelope protein	No	0.014	0.144	0.099
CC genome specific gene	Contig 1	11	No hit	Uroporphyrinogen decarboxylase					
		17	No hit	Hypothetical protein similar to retrotransposon protein					
		18	No hit	Reverse transcriptase					
		21	No hit	Putative two component LuxR family transcriptional regulator					
		26	No hit	Sodium- and chloride-dependent betaine transporter					
		29	No hit	Putative reverse transcriptase					
		30	No hit	Unknown protein similar to retrotransposon protein					
	Contig 2	31	No hit	Unknown protein similar to polyprotein					
		4	No hit	Putative cytochrome P450 93A3					
		9	No hit	Reverse transcriptase					
		10	No hit	Putative mutator-like transposase					
		11	No hit	Putative lysR family transcriptional regulator					
		17	No hit	Putative ST8SIA2 protein					
		20	No hit	Reverse transcriptase					
		33	No hit	Hypothetical protein					
		34	No hit	Putative p16INK4a					
		35	No hit	Unknown protein					
		37	No hit	Putative heavy metal efflux pump CzcA					

Long terminal repeat (LTR) retrotransposons constituted the single largest TE class (class I TEs) in terms of size contribution (Ammiraju et al., 2008). And different from DNA transposons (class II TEs), which can be deleted precisely at a relatively low frequency, the vast majority of retrotransposon insertions (class I TEs) are irreversible, rarely undergoing precise excision (Huang et al., 2008).

When *O. officinalis* genome accumulate large number of retrotransposon elements, and occurring with their transposition through reverse transcription of an RNA intermediate, the genome size changed simultaneously. Therefore, it appears that *O. officinalis* has expanded major in both intergenic regions and the genes themselves relative to *O. sativa*.

There is another type of genes of *O. officinalis* that had

no orthologs in the orthologous regions of *O. sativa*, but had orthologs to *O. sativa* genes located in other regions of the genome. Most of these orthologs are on chromosome 11 of *O. sativa*. This is probably a result of early genome duplication or chromosome segmental duplication. When sequence divergence of orthologous pairs are compared between the CC and AA genomes, genes that maintained collinearity in the corresponding chromosome regions have lower sequence divergence for Ks (*T*-test,  $P = 0.0017$ ) than those orthologous pairs in different chromosome regions (Table 5), suggesting that the genome or segmental duplication occurred prior to the divergence of AA and CC genomes.

### Acknowledgements

This work was supported by the grants from the Ministry of Science and Technology of China (the China Rice Functional Genomics Programs, No. 2006AA10A102 and 2005CB120805), the Chinese Academy of Sciences (No. KSCX2-YWN-024) and the Shanghai Municipal Commission of Science and Technology. We thank all colleagues in sequencing department for generating the sequence and Hao Hu (CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, CAS, China) and Kai Ying (Schnable lab, Carver co lab, ISU, Ames, IA, USA) for assistance of bioinformatics, Tao Sang (Department of Plant Biology, Michigan State University, USA) for comments on the manuscript, Rod A. Wing (Arizona Genomics Institute, The University of Arizona, Tucson, AZ, USA) and Dave Kudrna (Arizona Genomics Institute, The University of Arizona, Tucson, AZ, USA) for providing *O. officinalis* BAC library, Xuehui Huang (National Center for Gene Research, CAS, China) and Yan Zhou (National Center for Gene Research, CAS, China) for helpful discussion.

### References

- Ammiraju, J.-S.S., Lu, F., Sanyal, A., Yu, Y., Song, X., Jiang, N., Pontaroli, A.C., Rambo, T., Currie, J., Collura, K., Talag, J., Fan, C.Z., Goicoechea, J.T., Zuccolo, A., Chen, J.F., Bennetzen, J.L., Chen, M.S., Jackson, S.A., and Wing, R.A. (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, J.-S.S., Zuccolo, A., Yu, Y., Song, X., Piegu, B., Chevalier, F., Walling, J.G., Ma, J.X., Talag, J., Brar, D.S., SanMiguel, P., Jiang, N., Jackson, S.A., Panaud, O., and Wing, R.A. (2007). Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J.* **52**: 342–351.
- Arumuganathan, K., and Earle, E.D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**: 208–218.
- Bennetzen, J.L. (2007). Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**: 176–181.
- Bruggmann, R., Bharti, A.K., Gundlach, H., Lai, J.S., Young, S., Pontaroli, A.C., Wei, F.S., Haberer, G., Fuks, G., Du, C.G., Raymond, C., Estep, M.C., Liu, R.Y., Bennetzen, J.L., Chan, A.P., Rabinowicz, P.D., Quackenbush, J., Barbazuk, W.B., Wing, R.A., Birren, B., Nusbaum, C., Rounsley, S., Mayer, K.F.-X., and Messing, J. (2006). Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* **16**: 1241–1251.
- Feng, Q., Zhang, Y.J., Hao, P., Wang, S.Y., Fu, G., Huang, Y.C., Li, Y., Zhu, J.J., Liu, Y.L., Hu, X., Jia, P.X., Zhang, Y., Zhao, Q., Ying, K., Yu, S.L., Tang, Y.S., Weng, Q.J., Zhang, L., Lu, Y., Mu, J., Ku, Y.Q., Zhang, L.-S., Yu, Z., Fan, D.L., Liu, X.H., Lu, T.T., Li, C., Wu, Y.R., Sun, T.G., Lei, H.Y., Li, T., Hu, H., Guan, J.P., Wu, M., Zhang, R.Q., Zhou, B., Chen, Z.H., Chen, L., Jin, Z.Q., Wang, R., Yin, H.F., Cai, Z., Ren, S.X., Lv, G., Gu, W.Y., Zhu, G.F., Tu, Y.F., Jia, J., Zhang, Y., Chen, J., Kang, H., Chen, X.Y., Shao, C.Y., Sun, Y., Hu, Q.P., Zhang, X.L., Zhang, W., Wang, L.J., Ding, C.W., Sheng, H.H., Gu, J.L., Chen, S.T., Ni, L., Zhu, F.H., Chen, W., Lan, L.F., Lai, Y., Cheng, Z.K., Gu, M.H., Jiang, J.M., Li, J.Y., Hong, G.F., Xue, Y.B., and Han, B. (2002). Sequence and analysis of rice chromosome 4. *Nature* **420**: 316–320.
- Ge, S., Sang, T., Lu, B.R., and Hong, D.Y. (1999). Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Natl. Acad. Sci. USA* **96**: 14400–14405.
- Huang, X.H., Lu, G.J., Zhao, Q., Liu, X.H., and Han, B. (2008). Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.* **148**: 25–40.
- International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature* **436**: 793–800.
- Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A.K., and Khurana, J.P. (2007). F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol.* **143**: 1467–1483.
- Kim, H., Hurwitz, B., Yu, Y., Collura, K., Gill, N., SanMiguel, P., Mullikin, J.C., Maher, C., Nelson, W., Wissotski, M., Braidotti, M., Kudrna, D., Goicoechea, J.T., Stein, L., Ware, D., Jackson, S.A., Soderlund, C., and Wing, R.A. (2008). Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. *Genome Biol.* **9**: R45.1–15.
- Kim, H., SanMiguel, P., Nelson, W., Collura, K., Wissotski, M., Walling, J.G., Kim, J.P., Jackson, S.A., Soderlund, C., and Wing, R.A. (2007). Comparative physical mapping between *O. sativa* (AA genome type) and *O. punctata* (BB genome type). *Genetics* **176**: 379–390.
- Lu, F., Ammiraju, J.-S.S., Sanyal, A., Zhang, S.L., Song, R.T., Chen, J.F., Li, G.S., Sui, Y., Song, X., Cheng, Z.K., de Oliveira, A.C., Bennetzen, J.L., Jackson, S.A., Wing, R.A., and Chen, M.S. (2009). Comparative sequence analysis of *MONOCULMI*-orthologous regions in 14 *Oryza* genomes. *Proc. Natl. Acad. Sci. USA* **106**: 2071–2076.
- Luo, M.Z., and Wing, R.A. (2003). An improved method for plant BAC library construction. *Methods Mol. Biol.* **236**: 3–20.
- Ma, J.X., Wing, R.A., Bennetzen, J.L., and Jackson, S.A. (2007). Evolutionary history and positional shift of a rice centromere. *Genetics* **177**: 1217–1220.
- Mao L., Wood T.C., Yu Y., Budiman M.A., Tomkins J., Woo S.-S.,

- Sasinowski M., Presting G., Frisch D., Goff S., Dean R.A., and Wing R.A. (2000). Rice transposable elements: A survey of 73,000 Sequence-Tagged-Connectors. *Genome Res.* **10**: 982–990.
- Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**: 1046–1047.
- Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S.A., Wing, R.A., and Panaud, O. (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.
- Sasaki, T., and Burr, B. (2000). International rice genome sequencing project: The effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3**: 138–141.
- Tang, H.B., Bowers, J.E., Wang, X.Y., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. *Science* **320**: 486–488.
- Vaughan, D.A. (1994). Wild relatives of rice: Genetic resources handbook. (International Rice Research Institute, Manila, Philippines).
- Wing, R.A., Ammiraju, J-S.S., Luo, M.Z., Kim, H., Yu, Y., Kudrna, D., Goicoechea, J.L., Wang, W.M., Nelson, W., Rao, K., Brar, D., Mackill, D.J., Han, B., Soderlund, C., Stein, L., SanMiguel, P., and Jackson, S.A. (2005). The *Oryza* Map Alignment Project: The golden path to unlocking the genetic potential of wild rice species. *Plant Mol. Biol.* **59**: 53–62.
- Yang, Z.H., and Nielsen, R. (2000). Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**: 32–43.
- Zhang, S.B., Gu, Y.Q., Singh, J., Coleman-Derr, D., Brar, D.S., Jiang, N., and Lemaux, P.G. (2007). New insights into *Oryza* genome evolution: High gene colinearity and differential retrotransposon amplification. *Plant Mol. Biol.* **64**: 589–600.
- Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K-S., and Yu, J. (2006). KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics.* **4**: 259–263.
- Zou, X.H., Zhang, F.M., Zhang, J.G., Zang, L.L., Tang, L., Wang, J., Sang, T., and Ge, S. (2008). Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* **9**: R49.1–13.