10-2-2010

Dynamic Oryza Genomes: Repetitive DNA Sequences as Genome Modeling Agents

Navdeep Gill

Phillip SanMiguel

Braham Deep Singh Dhillon

Brian Abernathy

HyeRan Kim

See next page for additional authors

Authors

Navdeep Gill, Phillip SanMiguel, Braham Deep Singh Dhillon, Brian Abernathy, HyeRan Kim, Lincoln Stein, Doreen Ware, Rod Wing, and Scott A. Jackson

Rice (2010) 3:251–269
DOI 10.1007/s12284-010-9054-7

Dynamic *Oryza* Genomes: Repetitive DNA Sequences as Genome Modeling Agents

Navdeep Gill · Phillip SanMiguel · Braham Deep Singh Dhillon · Brian Abernathy · HyeRan Kim · Lincoln Stein · Doreen Ware · Rod Wing · Scott A. Jackson

Received: 2 November 2009 / Accepted: 5 September 2010 / Published online: 2 October 2010

Abstract Repetitive sequences, primarily transposable elements form an indispensable part of eukaryotic genomes. However, little is known about how these sequences originate, evolve and function in context of a genome. In an attempt to address this question, we performed a comparative analysis of repetitive DNA sequences in the genus *Oryza*, representing ~15 million years of evolution. Both Class I and Class II transposable elements, through their expansion, loss and movement in the genome, were found to influence genome size variation in this genus. We identified 38 LTRretrotransposon families that are present in 1,500 or more copies throughout *Oryza*, and many are preferentially amplified in specific lineages. The data presented here, besides furthering our understanding of genome organization in the genus *Oryza*, will aid in the assembly, annotation and analysis of genomic data, as part of the future genome sequencing projects of *O. sativa* wild relatives.

Keywords Repetitive sequences · Transposable elements · LTR-retrotransposons · *Oryza* · Genome size variation

Introduction

The genus *Oryza*, to which cultivated rice belongs, is composed of 23 species (Vaughan et al. 2003), including 21 wild and two cultivated species. Based on interspecific crossing (Tateoka 1963, 1964), chromosome pairing (Nayar 1973; Li et al. 2001) and total genomic DNA hybridization (Aggarwal et al. 1997), these species have been divided into ten distinct genome types: six diploid ($2n=24$) and four

Electronic supplementary material The online version of this article (doi:10.1007/s12284-010-9054-7) contains supplementary material, which is available to authorized users.

N. Gill · B. Abernathy · S. A. Jackson (✉)
Department of Agronomy, Purdue University,
West Lafayette, IN 47907, USA
e-mail: sjackson@purdue.edu

P. SanMiguel
Department of Horticulture and Landscape Architecture,
Purdue University,
West Lafayette, IN 47907, USA

P. SanMiguel
Purdue Genomics Facility, Purdue University,
West Lafayette, IN 47907, USA

B. D. S. Dhillon
Department of Botany and Plant Pathology, Purdue University,
West Lafayette, IN 47907, USA

L. Stein · D. Ware
Cold Spring Harbor Laboratory,
Cold Spring Harbor, NY 11724, USA

D. Ware
Soil and Nutrition Laboratory Research Unit,
USDA-ARS NAA Plant,
Ithaca, NY 14853, USA

H. Kim · R. Wing
Department of Plant Sciences, Arizona Genomics Institute,
University of Arizona,
Tucson, AZ 85721, USA

allotetraploid (2*n*=48). Owing to its economic importance (Khush 1997), small genome size (Arumuganathan and Earle 1991) and evolutionary relationship with other cereals (Moore et al. 1995), rice was the first crop to be sequenced (IRGSP 2005).

Analysis of the rice genome has shown that ~40% of the genome consists of known repetitive deoxyribonucleic acid (DNA; unpublished data), of which, at least 35% are transposons (IRGSP 2005). Repetitive sequences form a crucial component of many eukaryotic genomes, so much so that certain features of eukaryotic genome organization have been implicated as consequences of evolutionary forces acting on repetitive sequences (Charlesworth et al. 1994). Both tandem arrays and transposable elements (TEs) have been found to be associated with non-recombining heterochromatic regions, which may be due to their differential accumulation in genomic regions where recombination is suppressed (Charlesworth et al. 1986; Charlesworth and Langley 1989; Charlesworth 1991). Repetitive sequences, primarily TEs, can be a major force driving gene/genome evolution due to their tendency to insert either near/within genes (Yang et al. 2005, 2007), or intergenic regions (San Miguel et al. 1996; San Miguel and Bennetzen 1998). For instance, LTR retrotransposons (LTR-RTs) have been shown to determine fruit shape in tomato, whereby a retrotransposon-mediated gene duplication event resulted in elongated fruit shape (Xiao et al. 2008).

Other repetitive elements, such as Pack-MULEs, have been shown to carry fragments of cellular genes from multiple chromosomal loci, some of which can be fused together to form novel open-reading frames that are expressed as chimeric transcripts (Jiang et al. 2004). Similarly, special types of Class II DNA transposons, called helitrons have been reported to capture complete or incomplete copies of host genes as they transpose (Morgante et al. 2005; Kapitonov and Jurka 2007). Such instances of genes/gene fragment acquisition by TEs represent a mechanism for the formation of new genes.

Besides their role in driving gene and genome evolution (Bennetzen 2000; Jiang et al. 2004; Shapiro and Sternberg 2005; Kapitonov and Jurka 2007), gene regulation (Lippman et al. 2004; Feschotte 2008; Okamura and Nakai 2008), and other important developmental and evolutionary beneficial effects, TE activity can also result in a fitness loss to the host (Mackay 1986). Their activity in terms of insertions and/or chromosomal rearrangements can cause deleterious mutations (Crow and Simmons 1983; Mackay 1986) including human genetic diseases (Wallace et al. 1991; Holmes et al. 1994). The dynamic nature of repetitive sequences thus has long-term evolutionary as well as functional significance for the host genome.

The DNA sequence of a repeat and its copy number can evolve rapidly, leading to specificity within a particular species/genome or even a chromosome (Galasso et al. 1995; Wang et al. 1995; Matyasek et al. 1997). During the course of evolution, the loss or gain of sequences at the corresponding orthologous locations can lead to variations in the quantity of genome-specific repetitive sequences. In rice, preferential amplification of specific repetitive sequences has been shown to have an influence on genome differentiation, irrespective of the genome size (Uozu et al. 1997), and may be involved in domestication and/or speciation events. Some recently amplified retrotransposons have been proposed to be the source of genomic differentiation in *Oryza* (Panaud et al. 2002).

The genus *Oryza* is an excellent system for intraspecific comparative genomics because the ten different genome types (both diploids and polyploids) diverged from each other ~15 MYA and from a common ancestor with sorghum and maize about 50–70 MYA (Wolfe et al. 1989). In addition, the amount of diversity contained within the genus *Oryza* is immense, in terms of variation in genome size, ploidy level, morphological traits, and ecological adaptations. Comparative analysis of repetitive sequences across these ten genome types will thus help to improve our understanding of the role of repetitive DNA sequences in shaping *Oryza* genomes, domestication, speciation, polyploidy, size variation, etc.

Toward this end, the availability of finished genomic sequence of *Oryza sativa* (IRGSP 2005) is an invaluable tool. The genome sequence can be used for comparative analyses with the wild relatives, for which Bacterial Artificial Chromosome (BAC) libraries, BAC-end sequences (BESs), and integrated physical maps are available (Wing et al. 2005; Ammiraju et al. 2006; Kim et al. 2008). Using these resources, we investigated the repetitive sequences within the genus *Oryza* and found association of these elements, particularly, the Class I LTR-RTs and Class II miniature inverted TEs (MITEs), with genome size variation. Preferential amplification of different types of repetitive sequences was seen in different genomes, illustrating the role of such sequences in genome expansion and contraction.

Results

BESs of 13 *Oryza* species representing 8–17% (Kim et al. 2008) of each of the ten *Oryza* genome types were analyzed for their repetitive DNA content. Both homology-based (RepeatMasker, Blast) and de novo (Tallymer, RECON) methods were used. A detailed analysis of all TE classes was done to determine their relative abundance and distribution across *Oryza*. A significant portion of each of

the genomes was found to consist of repetitive DNA sequences, with LTR-RTs being a major component and hence one of the factors contributing to genome size variation in the genus *Oryza*.

Cataloging high, mid, and low repetitive BAC clones

Tallymer (Kurtz et al. 2008), a de novo approach, was used to identify mathematically defined repeats in the BESs of all *Oryza* species. The method consists of digesting a query sequence into overlapping 20-mers in the 5′ to 3′ direction, and copy numbers of these fragments are computed relative to an independent, unbiased sequence using algorithms based on vmatch (www.vmatch.de) that employ suffix arrays to index the unbiased sequence libraries necessary for statistical annotation (Kurtz et al. 2008; http://www.zbh.uni-hamburg.de/Tallymer). Using this approach, we computed and analyzed 20-mer frequencies for the entire *Oryza* BES dataset. BES pairs were assembled with Ns in between the forward and reverse read to obtain one sequence/BAC clone (referred to as a BAC clone for Tallymer analysis). Based on the frequencies of overlapping 20-mers for each BAC clone, the clones were categorized into low repetitive (0–40% repetitive), mid repetitive (40–70% repetitive), and high repetitive (70–100% repetitive) clones (Fig. 1a). When plotted logarithmically on a genomic scale, these frequencies form a repeat landscape, whereby high copy repetitive regions are easily distinguished from low copy, putatively genic

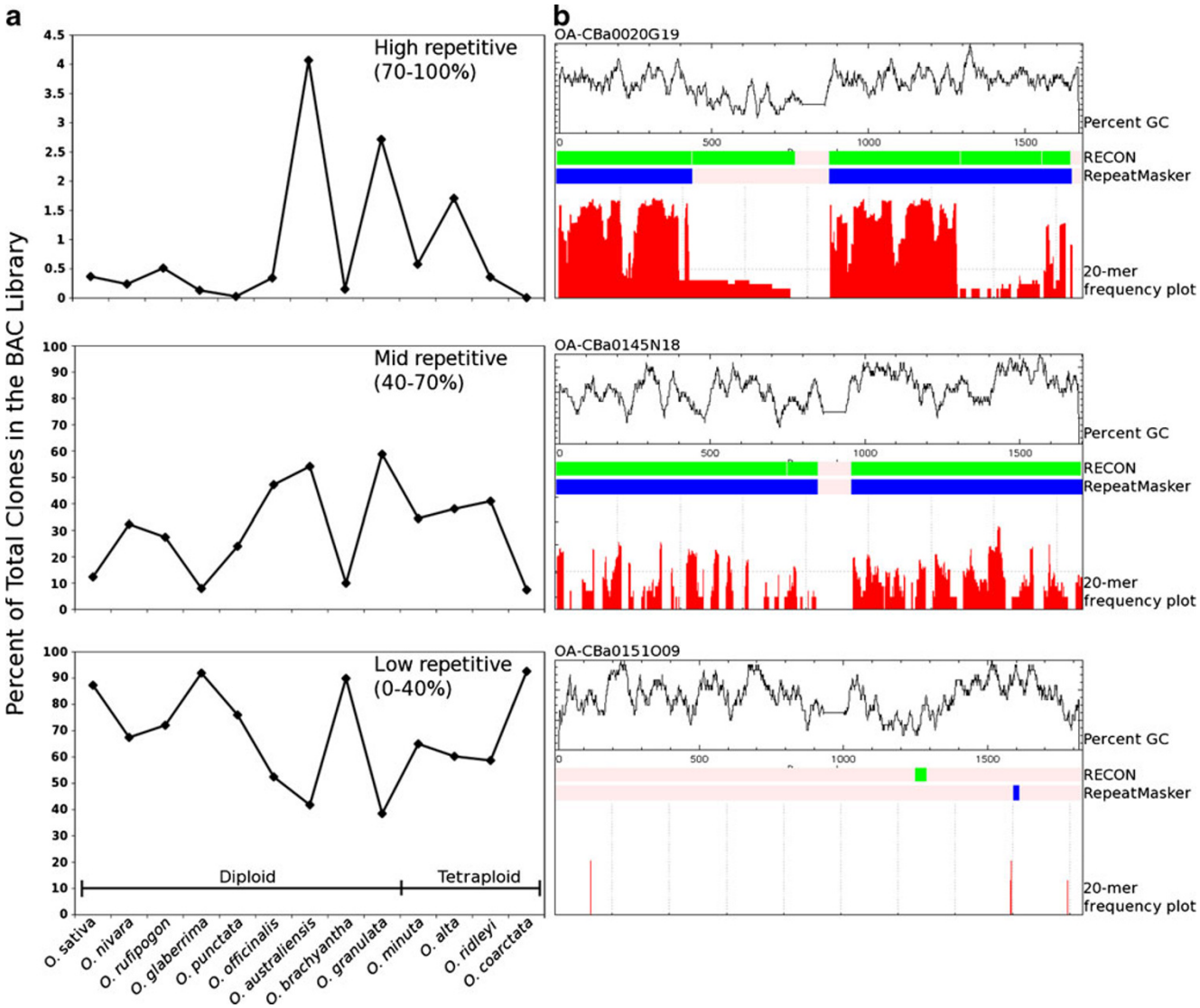


Fig. 1 **a** Percent of total clones in the BAC library of each species classified into low, mid and high repetitive based on the frequencies of the overlapping 20-mers for each clone. The species are arranged according to the ploidy level. **b** An illustration of each type: low (OA_CBa0151O09), mid (OA_CBa 0145N18), and high (OA_CBa0020G19) repetitive clone from *O. australiensis* along with the percent GC, RECON, and RepeatMasker annotations for each clone.

regions. An example of each type: low (OA_CBa0151O09), mid (OA_CBa0145N18), and high (OA_CBa0020G19) repetitive clones from *Oryza australiensis* are shown (Fig. 1b) with comparison of two other methods, de novo (RECON) and similarity-based (RepeatMasker) annotations.

Based on the K-mer analyses, 67.5–91.9% of the clones in all the diploid species are low repetitive except *Oryza officinalis* [CC], *O. australiensis* [EE], and *Oryza granulata* [GG], which have 52.4%, 41.8%, and 38.5% low repetitive clones, respectively. Interestingly, these three species have the highest percentage of mid repetitive clones (47.2–58.8%) among all the diploids. Approximately 4.1% of all *O. australiensis* and 2.7% of all *O. granulata* clones are 70–100% repetitive, whereas for all other diploids, only 0% (*Oryza punctata*) to 0.5% (*Oryza rufipogon*) of clones fall into this category. Most of the clones in *O. officinalis* therefore are either low or mid repetitive with only 0.3% high repetitive clones. *O. australiensis* and *O. granulata* (the two largest and most repetitive genomes in *Oryza*), on the other hand, have mostly mid to high repetitive clones, suggesting the presence of more high copy sequences as compared to other diploids.

Oryza brachyantha [FF], the smallest diploid genome, has 89.9%, 10%, and 0.2% of the clones in the low, mid, and high repetitive category, respectively, suggesting the prevalence of low copy sequences in its genome. In contrast, *O. australiensis* with the biggest diploid genome, and repetitive content highest among all *Oryza* species, has a bulk of its clones that are mid to high copy. Alternatively, individual BAC end reads from the two genomes were plotted against their repetitive content as determined by RepeatMasker for an overview of their distribution pattern at the whole genome level (Fig. S1). Of the total reads, 86.3% and 66.1% are repetitive in *O. australiensis* and *O. brachyantha*, respectively. Again, the preponderance of high copy repeats in *O. australiensis* is inferred from the distribution pattern of individual reads as more number of sequences are clustered in the 70–100% repetitive range in *O. australiensis* (~73% of the total reads), higher than *O. brachyantha* (~27% of the total reads).

Tetraploids, with the exception of *Oryza coarctata*, have 58.7–65%, 34.4–41%, and 0.4–1.7% of the clones that are low, mid, and high repetitive, respectively. Another exception is *Oryza alta* that has the highest percentage of clones in the 70–100% repetitive category (1.7%). This is consistent with an earlier report where *O. alta* has been shown to contain a Ty3-gypsy type of retrotransposon amplified to significant portions of its genome (Zuccolo et al. 2007). *O. coarctata* [HHKK] is the only tetraploid species which has 92.6% of its clones that are low repetitive, 7.4% mid repetitive, and 0% high repetitive, indicating an overall low repetitive content in terms of total number of repetitive bases in the genome.

A list of all the clones belonging in each of these repetitive categories is provided (Supplemental Files 1–3).

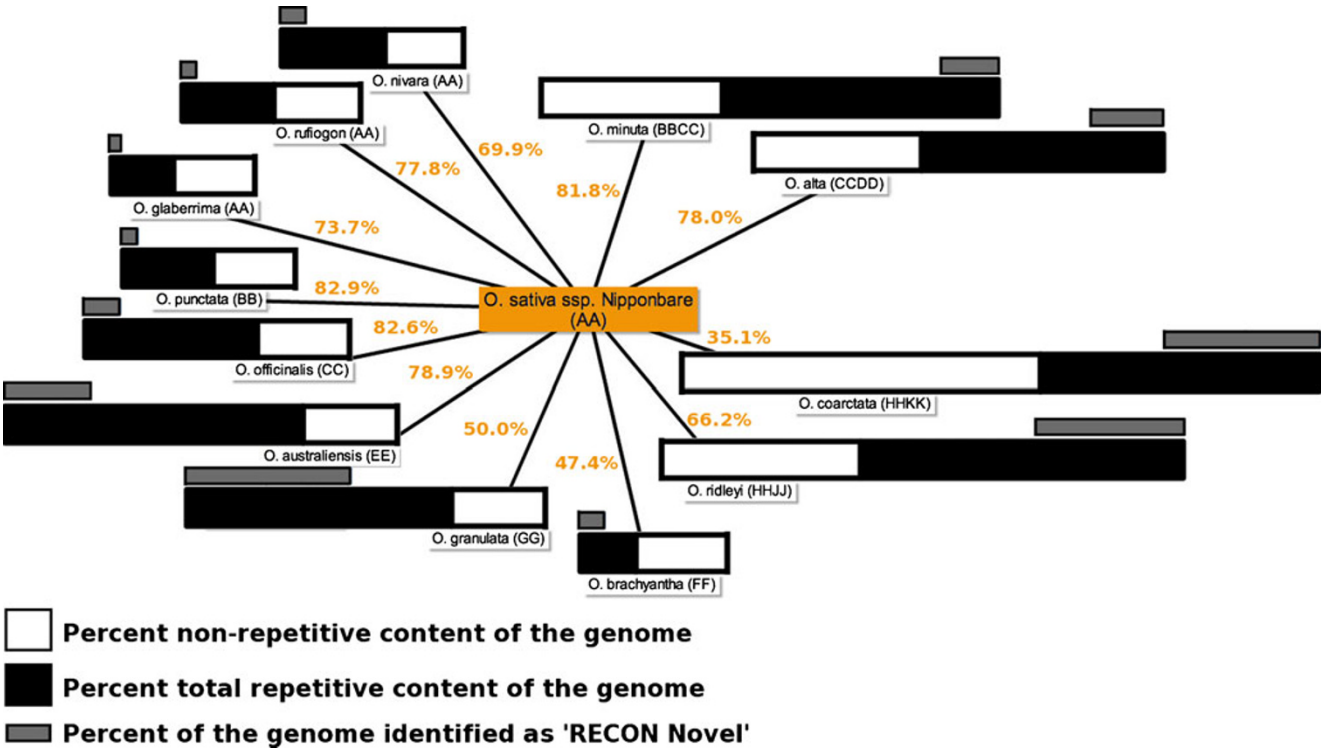


Fig. 2 Repeat analysis in *Oryza* using RECON and RepeatMasker in comparison to *O. sativa* ssp. Nipponbare. The numbers indicate the percentage of de novo repeats shared with *O. sativa* custom repeat library.

Table 1 Identification and classification of repetitive elements in all *Oryza* species using RepeatMasker, in conjunction with individual species-specific repeat databases

Repeat class	Classs	Repeat type	Percent repetitive content of the genome (percent of total repetitive)												
			Os	*On*	*Or*	*Og*	*Op*	*Oo*	*Oa*	*Ob*	*Ogr*	*Om*	*Oal*	*Ori*	*Oc*
Class I TEs		Ty1-copia	3.34 (5.65)	3 (4.51)	3.09 (5.02)	3.55 (6.65)	4.16 (6.31)	4.27 (5.54)	7.14 (8.33)	3.98 (8.84)	4.28 (5.21)	4.43 (6.17)	5.55 (7.99)	5.14 (6.94)	5.73 (11.65)
		Ty3-gypsy	11.54 (19.51)	14.28 (21.45)	13.34 (21.65)	9.1 (17.04)	17.56 (26.63)	17.95 (23.28)	32.3 (37.69)	2.84 (6.31)	13.68 (16.64)	18.98 (26.45)	12.78 (18.39)	14.06 (18.98)	3.86 (7.85)
		CRRs	2.73 (4.61)	3.23 (4.85)	3.27 (5.31)	2.01 (3.77)	2.75 (4.18)	2.03 (2.63)	2.39 (2.79)	0.48 (1.07)	2.51 (3.05)	1.88 (2.62)	1.66 (2.39)	2.44 (3.29)	1.05 (2.13)
		TRIM	0.04 (0.08)	0.05 (0.08)	0.05 (0.09)	0.05 (0.1)	0.05 (0.08)	0.04 (0.05)	0.02 (0.02)	0.06 (0.14)	0.05 (0.06)	0.04 (0.05)	0.04 (0.06)	0.02 (0.02)	0.04 (0.08)
		Solo LTRs	0.05 (0.08)	0.04 (0.05)	0.04 (0.06)	0.05 (0.1)	0.04 (0.06)	0.06 (0.08)	0.03 (0.04)	0.18 (0.4)	0.42 (0.51)	0.06 (0.08)	0.04 (0.06)	0.18 (0.24)	0.05 (0.09)
		Unique(LTR_STRUC)	8.99 (15.19)	10.14 (15.23)	10.27 (16.67)	7.75 (14.51)	8.86 (13.44)	18.94 (24.57)	7.74 (9.03)	4.41 (9.79)	5.75 (6.99)	13.57 (18.91)	8.94 (12.87)	6.63 (8.96)	3.38 (6.87)
		LINEs	0.32 (0.55)	0.34 (0.51)	0.37 (0.6)	0.36 (0.67)	0.29 (0.45)	0.26 (0.34)	0.43 (0.5)	0.17 (0.38)	0.16 (0.2)	0.26 (0.36)	0.38 (0.54)	0.19 (0.25)	0.21 (0.42)
		SINEs	0.05 (0.09)	0.05 (0.08)	0.05 (0.09)	0.07 (0.13)	0.04 (0.06)	0.03 (0.04)	0.01 (0.01)	0.06 (0.14)	0.01 (0.01)	0.03 (0.04)	0.03 (0.04)	0.02 (0.02)	0.01 (0.03)
		Evirus/ERTBV	0.21 (0.36)	0.05 (0.07)	0.09 (0.14)	0.03 (0.06)	0.09 (0.14)	0.19 (0.25)	0.14 (0.17)	0.01 (0.02)	0.16 (0.19)	0.13 (0.18)	0.13 (0.18)	0.2 (0.28)	0.09 (0.17)
		Unknown	0.56 (0.95)	0.41 (0.62)	0.44 (0.72)	0.45 (0.84)	0.45 (0.69)	0.73 (0.95)	0.51 (0.59)	0.22 (0.49)	0.85 (1.03)	0.52 (0.72)	0.64 (0.92)	0.43 (0.58)	0.88 (1.79)
		Total	27.85 (47.08)	31.59 (47.45)	31.01 (50.34)	23.42 (43.87)	34.31 (52.02)	44.49 (57.71)	50.73 (59.18)	12.42 (27.59)	27.87 (33.91)	39.88 (55.58)	30.18 (43.44)	29.3 (39.57)	15.29 (31.09)
Class II TEs	(Non MITEs)	CACTA/EnSpm	2.67 (4.51)	2.26 (3.4)	2.36 (3.83)	2 (3.75)	3.48 (5.27)	3.6 (4.67)	2.62 (3.05)	0.92 (2.05)	0.43 (0.53)	3.54 (4.93)	1.76 (2.53)	2.47 (3.33)	0.99 (2.01)
		Mutator (MULEs)	2.65 (4.49)	2.73 (4.1)	2.88 (4.68)	2.97 (5.56)	2.74 (4.15)	2.28 (2.96)	1.62 (1.89)	1.82 (4.04)	1.23 (1.49)	2.52 (3.51)	2.08 (3)	2.96 (4)	1.3 (2.65)
		hAT	1.19 (2)	1.07 (1.6)	1.09 (1.76)	1.06 (1.99)	1.05 (1.6)	0.89 (1.15)	0.76 (0.89)	0.96 (2.13)	0.91 (1.1)	0.99 (1.38)	0.81 (1.16)	0.81 (1.09)	0.57 (1.15)
		Tc1	0.1 (0.17)	0.1 (0.15)	0.1 (0.16)	0.12 (0.22)	0.05 (0.08)	0.05 (0.07)	0.02 (0.03)	0.03 (0.06)	0.03 (0.04)	0.06 (0.09)	0.06 (0.08)	0.04 (0.06)	0.03 (0.07)
		Helitron	1.68 (2.84)	1.44 (2.17)	1.58 (2.57)	1.68 (3.15)	1.25 (1.89)	0.9 (1.17)	0.29 (0.34)	0.5 (1.11)	0.4 (0.49)	0.97 (1.35)	0.88 (1.27)	0.39 (0.53)	0.4 (0.81)
		Others[a]	0.48 (0.82)	0.42 (0.64)	0.45 (0.73)	0.43 (0.8)	0.44 (0.67)	0.38 (0.49)	0.28 (0.33)	0.4 (0.88)	0.4 (0.49)	0.26 (0.36)	0.32 (0.46)	0.35 (0.48)	0.19 (0.39)
		Unknown	0.21 (0.35)	0.21 (0.32)	0.23 (0.38)	0.23 (0.44)	0.2 (0.3)	0.15 (0.19)	0.08 (0.1)	0.26 (0.57)	0.08 (0.1)	0.14 (0.2)	0.14 (0.2)	0.14 (0.19)	0.14 (0.29)
	MITEs	Tourist	1.4 (2.37)	1.44 (2.16)	1.5 (2.44)	1.63 (3.04)	0.74 (1.12)	0.6 (0.78)	0.26 (0.3)	1.94 (4.32)	0.38 (0.47)	0.6 (0.83)	0.55 (0.79)	0.44 (0.6)	0.43 (0.88)
		Stowaway	0.96 (1.63)	0.92 (1.38)	1.02 (1.66)	1.12 (2.09)	0.68 (1.04)	0.56 (0.73)	0.24 (0.28)	0.82 (1.81)	0.32 (0.39)	0.53 (0.74)	0.52 (0.74)	0.35 (0.48)	0.34 (0.69)
		Explorer	0.08 (0.14)	0.08 (0.11)	0.08 (0.14)	0.08 (0.16)	0.04 (0.06)	0.03 (0.03)	0.01 (0.01)	1.11 (2.46)	0.02 (0.02)	0.03 (0.04)	0.03 (0.05)	0.07 (0.09)	0.04 (0.07)
		Others	0.15 (0.25)	0.13 (0.2)	0.16 (0.26)	0.18 (0.33)	1.48 (2.24)	0.11 (0.14)	0.06 (0.07)	0.14 (0.31)	0.05 (0.06)	0.1 (0.14)	0.09 (0.13)	0.13 (0.17)	0.07 (0.14)
		Total	11.58 (19.57)	10.8 (16.22)	11.47 (18.61)	11.5 (21.53)	12.15 (18.42)	9.56 (12.4)	6.26 (7.3)	8.89 (19.74)	4.25 (5.18)	9.75 (13.59)	7.24 (10.41)	8.16 (11.02)	4.51 (9.16)
Others		RECON novel	9.58 (16.2)	14.15 (21.25)	8.66 (14.06)	8.43 (15.8)	9.25 (14.02)	13.64 (17.7)	22.07 (25.74)	17.2 (38.2)	45.95 (55.91)	12.65 (17.63)	17.65 (25.4)	28.5 (38.48)	24.43 (49.65)
		Unclassified	0.17 (0.29)	0.15 (0.23)	0.15 (0.24)	0.18 (0.34)	0.12 (0.18)	0.13 (0.17)	0.08 (0.09)	0.06 (0.14)	0.04 (0.05)	0.14 (0.2)	7.32 (10.54)	0.13 (0.17)	0.06 (0.13)
Repetitive[b]			45.65	56.82	51.39	43.41	52.37	65.14	75.92	37.74	73.82	60.32	58.95	61.86	43.74

Os O. sativa, *On O. nivara*, *Or O. rufipogon*, *Og O. glaberrima*, *Op O. punctata*, *Oo O. officinalis*, *Oa O. australiensis*, *Ob O. brachyantha*, *Ogr O. granulata*, *Om O. minuta*, *Oal O. alta*, *Ori O. ridleyi*, *Oc O. coarctata*

[a] Others include Osmar/Mariner (MLE), Basho, PILE, POLE, and Micropon

[b] These values also include the following categories: other centromeric, telomeric, simple repeats, and low-complexity repeats

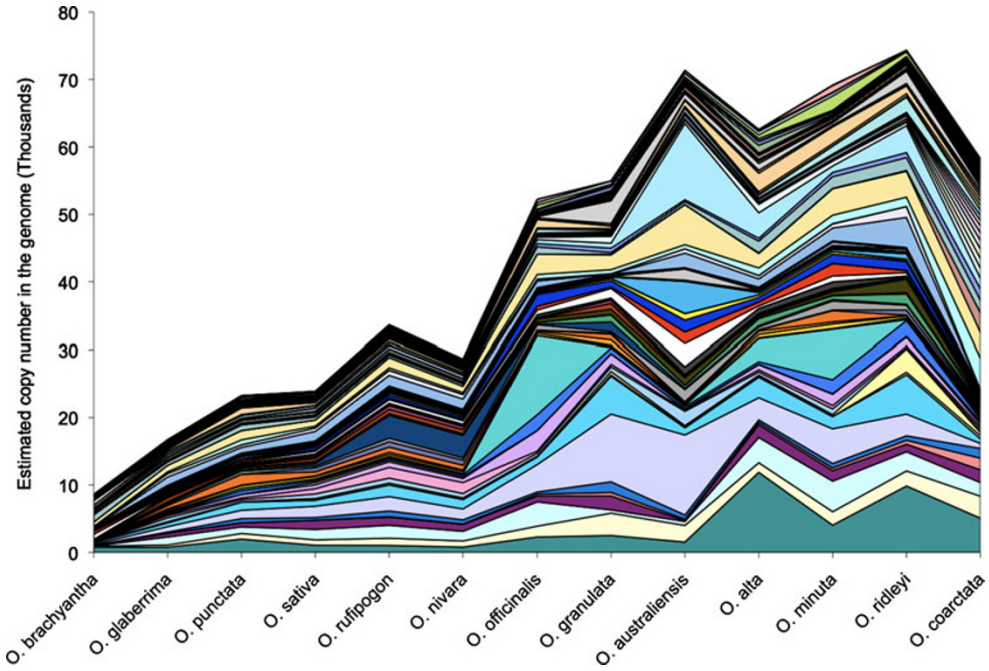


Fig. 3 Estimated copy numbers of 109 LTR retroransposon (LTR-RT) families (52 Ty1-copia and 57 Ty3-gypsy families) in the genomes of *Oryza* species. The species are arranged in order of their increasing genome sizes. Each line in the graph represents a particular LTR-RT family.

A practical application of this analysis will be for sequencing and/or assembly purposes. Clones that are 90–100% repetitive can be barred from a minimum tiling path for sequencing or during assembly of sequenced data. The low repetitive clones will be useful for accessing the genic portions of each genome.

Repertoire of repetitive sequences in different species

RECON was used to identify repeats de novo for each species. Overlaps between the de novo repeats and the curated rice (*O. sativa*) repeat database (described in the "Materials and methods" section) were determined using RepeatMasker. Redundant sequences were removed, forming a specific repeat database for each species. The percentage of the de novo repeats shared with *O. sativa* custom repeat library was calculated for all species (Fig. 2). Of the total de novo repeats identified in each species, *O. punctata* [BB], *O. officinalis* [CC], and their tetraploid *Oryza minuta* [BBCC] share the most repeats with *O. sativa*, even more than the other diploid A-genome species (*O. rufipogon*, *Oryza Nivara,* and *Oryza glaberrima*). *O. coarctata* [HHKK] has the least repeat similarity to *O. sativa* (35% of total de novo identified), followed by *O. brachyantha* [FF] (47.4%) and *O. granulata* [GG] (50%). This trend is expected as more distant genomes are expected to share fewer repeats with *O. sativa*. Of the total de novo repetitive element repertoire of *O. coarctata*, *O. brachyantha*, and *O. granulata*, ≥50% is unique with respect to *O. sativa*, implying the existence of repeats that are more diverged to *O. sativa* as compared to the rest of the *Oryza* species (17.1% in *O. australiensis* [EE] to 33.8% in *Oryza ridleyi* [HHJJ]).

After removing the shared repeats, the remaining sequences were annotated using the all-plant repeat database. Interestingly, repetitive sequences corresponding to 46% of *O. granulata*, 28.5% of *O. ridleyi*, 24.4% of *O. coarctata*, and 22% of *O. australiensis* genomes were classified as "RECON novel" (Table 1), as we were unable to assign any annotations to these repeats. In general, a greater evolutionary distance from *O. sativa* correlated with abundance of novel repeats. However, in *O. australiensis*, although 22% of the genome comprises novel repeats (equivalent to 25.7% of the total repetitive DNA), only 17.1% of the de novo repeats were unique to *O. australiensis* with respect to *O. sativa*. This indicates amplification of only a small portion of the de novo repeats (17.1%) to occupy approximately 22% of the genome, which is consistent with the observation of a rapid recent burst of retrotransposons (Wallabi, Kangourou, and a RIRE1 element) in this species (Piegu et al. 2006).

In order to identify and classify the repetitive elements in different species, RepeatMasker, in conjunction with each species-specific repeat database, was used. The amount of repetitive content of a genome (Table 1) was found to be correlated to its genome size (Pearson's correlation coefficient of 0.9). *O. australiensis* [EE], the largest diploid genome, and *O. brachyantha* [FF], the smallest diploid genome, had the highest (76%) and lowest (38%) amount of repetitive DNA, respectively, supporting the role of repetitive sequences in genome size expansion in *Oryza*. There were dramatic differences in the repeat profiles of *O. australiensis* and *O. brachyantha* with respect to Class I and Class II TEs. Approximately 59% of the total repetitive DNA in *O. australiensis* was Class I retrotransposons and

~7% was Class II DNA transposons, whereas in *O. brachyantha*, it was 27% and 20%, respectively.

Among the tetraploid species, *O. coarctata* [HHKK] had the lowest repetitive content (44%) compared to others. Interestingly, if only the similarity to the *O. sativa* repeat database is considered, *O. coarctata* has the lowest repeat content (19.3%) in entire *Oryza*, which is lower than *O. brachyantha* (20.5%). Among the diploids, *O. officinalis* [CC], *O. australiensis* [EE], and *O. granulata* [GG] have an unusually high repetitive content of 65%, 76%, and 74%, respectively, which are higher than the tetraploid genomes (*O. minuta* 60%, *O. alta* 59%, *O. ridleyi* 62%, and *O. coarctata* 44%).

Not surprisingly, Class I retrotransposons (both LTR and non-LTR) were identified as the largest class of repetitive sequences, followed by Class II DNA Transposons (both MITE and non-MITE; Table 1). Among Class I retrotransposons, centromeric retrotransposons of rice (CRRs) ranged from 1% of total repetitive in *O. brachyantha* [FF] to 5% in *O. rufipogon* [AA] suggesting either fewer copies or diverged CRRs from *O. sativa* or entirely different types of CRRs in *O. brachyantha* as previously suggested (Gao et al. 2009). Among Class II DNA transposons, helitrons were most abundant in the four A-genome species (2.1% of total repetitive in *O. nivara* to 3.1% in *O. glaberrima*) and decreases thereafter as the evolutionary distance increases. Excluding the A-genomes, the amount of helitrons range from 1.9% of total repetitive in *O. punctata* to 0.3% in *O. australiensis*.

Simple sequence and low complexity repeats were also identified using RepeatMasker. Their relative abundance and density [number of simple sequence repeats (SSRs)/Mbp of the genome] and the most frequent type of SSR motif within each di-, tri-, and tetranucleotide repeats were determined (Tables S1 and S2). Owing to their polymorphic nature and frequent associations with genes, SSRs have an advantage to be used as genetic markers for breeding as well as for intraspecific mapping populations for functional studies (Kim et al. 2008).

Retrotranspositional success of different LTR-RT families in different species

LTR-RTs have an important role in genome size expansion by virtue of their copy–paste mechanism of transposition (Kumar and Bennetzen 1999). There is a tremendous variation in the amount of a genome occupied by LTR-RTs, ranging from 49.8% in *O. australiensis* to 12% in *O. brachyantha*. We estimated the copy number of 111 LTR-RT families (53 Ty1-copia and 58 Ty3-gypsy families) in the genomes of all *Oryza* species (Fig. 3). The total number of LTR-RT copies in the genome is correlated with the repetitive content of the genome as well genome size (Pearson's correlation coefficient of 0.9 and 0.8, respectively). *O. brachyantha*, with the smallest genome and the least repetitive content had the lowest number of LTR-RT copies (13,602) as compared to *O. australiensis* (163,145), the largest diploid genome with the highest repetitive content. A general trend of rapid amplification of LTR-RTs in terms of increase in copy number was seen in all the polyploids with the exception of three diploid species- *O. officinalis*, *O. australiensis*, and *O. granulata*. These three species have an exceptionally high copy number of LTR-RTs, resulting in an increase in genome size as compared to other diploids. Ploidy alone therefore is not responsible for increased genome size in the four tetraploid genomes as preferential amplification of LTR-RTs in both diploid and polyploid genomes can also contribute to increase in genome size.

Estimated copy numbers for 58 Ty3-gypsy and 53 Ty1-copia families were compared throughout *Oryza* (Table S3). Copy number distribution of 52 Ty1-copia and 57 Ty3-gypsy families was plotted against the genome size of every species to compare the transpositional rate of these elements (Fig. 4). The uncharacterized families in both copia and gypsy class were excluded. Throughout *Oryza*, higher amplification of Ty3-gypsy, as compared to Ty1-copia-type elements, was observed, as seen by the differences in the number of individual families that have amplified in the genome in each of the three categories (500–1,000 copies, 1,000–1,500 copies, and >1,500 copies). Very few families have amplified to reach >1,500 copies in any genome of either Ty1-copia or Ty3-gypsy types. The polyploids had more high copy LTR-RT families than diploids except the three high repetitive diploid species: *O. officinalis*, *O. australiensis*, and *O. granulata* (Table 2). *O. brachyantha*, however, lacked any family that amplified to 1,500 copies or more. The RIRE1 Ty1-copia family has amplified to ~30,000 copies in the *O. australiensis* genome (Piegu et al. 2006). is also found in >1,500 copies in all tetraploids except *O. minuta* [BBCC], and is present in only 55 copies in *O. brachyantha*. Similarly, *O. coarctata* [HHKK] is the only tetraploid lacking significant amplification of RIRE2 gypsy-type family of retrotransposons. RIRE2 has been shown previously to account for a significant portion of the genome size variation in *Oryza* (Zuccolo et al. 2007).

RETROSAT2, a CRR, is highly amplified only in the *O. australiensis* genome. RC1067, a Ty3-gypsy type of retrotransposon family, is found in high copy in all the *Oryza* species except two diploids, *O. brachyantha* and *O. granulata*, and two tetraploids, *O. ridleyi* and *O. coarctata*. Among the four A-genome species, *O. rufipogon* is the only species where SZ5, a Ty1-copia-type family, has been amplified in large numbers. No other copia family is present in >1, 500 copies in the other A-genome species.

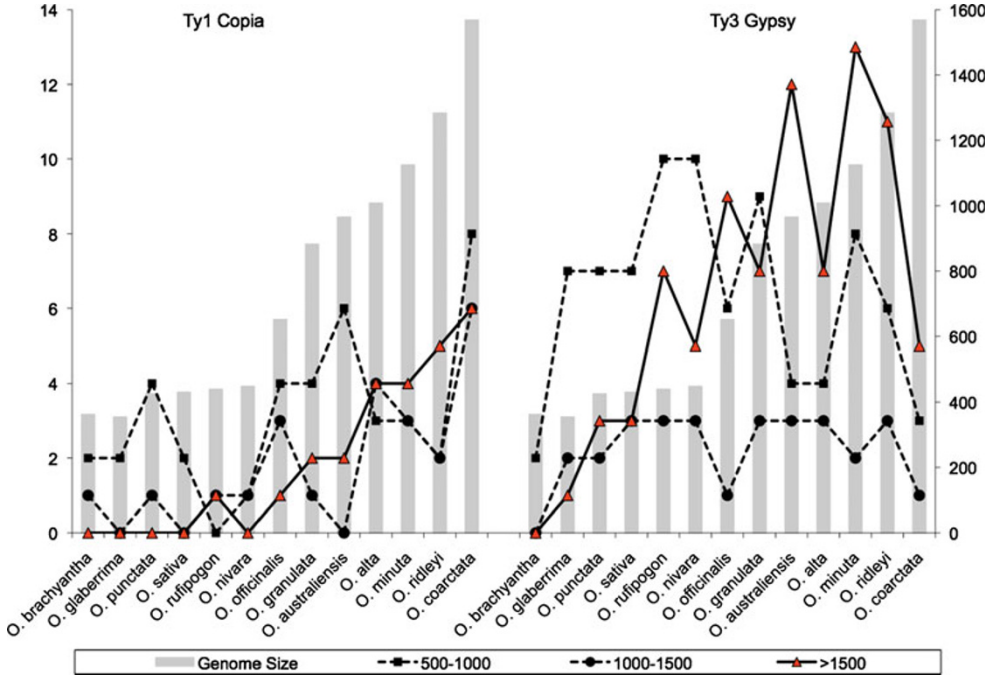


Fig. 4 Comparison of the transpositional rate of 52 Ty1-copia and 57 Ty3-gypsy families based on copy number distribution, plotted against the genome size of each *Oryza* species. Axis on the *left* corresponds to the copy number distribution and the axis on the *right* corresponds to the genome size. The species are arranged in order of their increasing genome sizes.

Similarly, SZ7 and SZ42 are two Ty3-gypsy types present in >1,500 copies only in *O. rufipogon* and absent in the other A-genome species. We identified 28 Ty3-gypsy and ten Ty1-copia families as possible candidates for preferential amplification in one species as compared to the others.

Preferential amplification of specific LTR-RTs in different species

In addition to LTR-RT families, we also analyzed the copy numbers of specific LTR-RT elements for varying rates of transposition throughout the genus *Oryza*. We found nine specific retrotransposons that appeared to have been preferentially amplified. Excluding Kangourou, Wallabi, and Gran3 which were previously shown to be amplified in large numbers in *O. australiensis* and *O. granulata* (Piegu et al. 2006; Ammiraju et al. 2007), we identified Atlantys, Dasheng, Dagul, Hopi, Houba, and Koala as preferentially amplified. These six retrotransposons (of which Koala and Houba are copia-type) are estimated to occupy 2.2–23% of the genome in *O. brachyantha* and *O. alta*, respectively [range, mean ± SD20.8, 11.5±6.4] (Fig. 5a). Based on an in-depth analysis of these individual elements, our results indicate a wide variation in the copy numbers, percent of total repetitive content, and percent of the genome occupied by these elements, indicating their favored amplification in one genome as compared to the others.

Variation in the estimated copy numbers of these six LTR-RTs was calculated for all the *Oryza* species (Fig. 5b; Kangourou, Wallabi, and Gran3 are also included). Copy number of Atlantys, a Ty3-gypsy type of retrotransposon is higher in BB, CC, and EE genomes and their corresponding tetraploids (*O. minuta* and *O. alta*) as compared to other species, with a maximum copy number in *O. alta* (14,727) and minimum in *O. brachyantha* (215). Atlantys has been shown previously to be abundant in the species from the *Officinalis* complex (Zuccolo et al. 2008). On the other hand, Koala, a copia type of RT, has increased in copy number only in the *O. coarctata* genome (1,462 copies), which is higher than all the tetraploids and also than the two most repetitive genomes *O. australiensis* and *O. granulata* (Fig. 5b).

Among all the *Oryza* species, *O. coarctata* has the highest number of copia elements (48,073 copies) and *O. minuta* has the highest number of gypsy-type elements (145,071). Among the diploids, however, *O. australiensis* has the highest number of both copia and gypsy, 30,993 and 132,151, respectively (Table S3), and excluding Kangourou, Wallabi, and Gran3, preferential amplification was seen for Dagul (*O. officinalis*, *O. granulata*), Dasheng (*O. australiensis*), and Houba (*O. granulata* and *O. australiensis*) LTR-RTs. In the teteraploids, Koala (*O. coarctata*), Hopi (*O. ridleyi*), Dasheng (*O. minuta*), Atlantys (*O. alta* and *O. minuta*), and Houba (all tetraploids) were the highest copy number elements. Of

Table 2 List of Ty1-copia and Ty3-gypsy families that have amplified to greater than 1,500 copies in the genus *Oryza*

Species	Repeat	Count	Families
O. brachyantha	Ty1-copia	0	–
	Ty3-gypsy	0	–
O. glaberrima	Ty1-copia	0	–
	Ty3-gypsy	1	RC1067
O. punctata	Ty1-copia	0	–
	Ty3-gypsy	3	RC1067, SZ21, SZ50
O. sativa	Ty1-copia	0	–
	Ty3-gypsy	3	RC1067, RIRE2, SZ27
O. rufipogon	Ty1-copia	1	SZ5
	Ty3-gypsy	7	RC1067, RIRE2, SZ7, SZ12, RIRE3, RIRE8, SZ42
O. nivara	Ty1-copia	0	–
	Ty3-gypsy	5	RC1067, RIRE2, SZ12, RIRE3, RIRE8
O. officinalis	Ty1-copia	1	SZ5
	Ty3-gypsy	9	RC1067, RIRE2, SZ7, SZ21, SZ, RCS1, SZ36, SZ35, SZ45
O. granulate	Ty1-copia	2	SZ5, SC22
	Ty3-gypsy	7	RIRE2, SZ21, SZ112, RCS1, SZ42, SZ35, Osr31
O. australiensis	Ty1-copia	2	SZ5, RIRE1
	Ty3-gypsy	12	RC1067, RIRE2, SZ21, SZ12, RCS1, GypsyA, SZ36, SZ107, SZ62, GypsyB, SZ101, RETROSAT2
O. alta	Ty1-copia	4	SZ5, RIRE1, SZ13, SZ27
	Ty3-gypsy	7	RC1067, RIRE2, SZ7, SZ21, SZ112, SZ42, SZ45
O. minuta	Ty1-copia	4	SZ5, SZ13, SZ27, SZ3
	Ty3-gypsy	13	RC1067, RIRE2, SZ7, SZ21, SZ112, SZ12, SZ, RCS1, SZ42, SZ45, SZ35, GypsyA, SZ50
O. ridleyi	Ty1-copia	5	SZ5, SC22, RIRE1, SZ13, SZ37
	Ty3-gypsy	11	RIRE2, SZ7, SZ21, SZ12, SZ, RCS1, SZ42, SZ106, RIRE7, RC1174, SZ104
O. coarctata	Ty1-copia	6	SZ5, RIRE1, SZ13, SZ6, SZ61, SZ30
	Ty3-gypsy	5	SZ7, SZ21, SZ112, RCS1, SZ110

the A-genome species, *O. glaberrima* seems to be an outlier with a deficiency of LTR-RTs (both Ty1-copia and Ty3-gypsy) especially the Dagul and Dasheng elements (Fig. 5b).

Ty1-copia outnumber Ty3-gypsy in *O. brachyantha* and *O. coarctata*

Throughout *Oryza*, the estimated copy number of gypsy LTR-RTs is higher than the copia types (Table S3). The ratio of gypsy:copia ranges from 2.97 in *O. glaberrima* to 6.33 in *O. minuta*, with the exception of *O. brachyantha* (0.89) among the diploids and *O. coarctata* (0.90) among the tetraploids. The amount of gypsy-type LTR-RTs in the genome is more correlated with the total repetitive content of the genome than copia types with correlation coefficients of 0.75 and 0.30, respectively. Interestingly, both species with more copia-type LTR-RTs than gypsy are also the least repetitive among the diploids and tetraploids. Of the total 53 families of Ty1-copia analyzed, we identified three families (SC13, SC22, and an uncharacterized copia family present in 1,079, 685, and 2,270 copies, respectively) in the in *O. brachyantha* genome (Table S3). Elements belonging to these three families form 56% of the total copia LTR-RTs present in *O. brachyantha*. Similarly, in *O. coarctata*, we identified seven copia families (SZ6, SZ57, SZ55, SZ30, SZ17, SZ18, and SZ13 present in 2,886, 1,435, 1,403, 1,682, 1,030, 1,496, and 1,869 copies, respectively) (Table S3) that account for 24.5% of the total copia elements in its genome. This suggests that majority of the copia families in *O. brachyantha* (50 out of 53) are relatively low copy forming 44% of the total copia LTR-RTs, whereas, in *O. coarctata*, a majority of the copia LTR-RTs are mid to high copy with 46 families out of 53 forming 75.5% of the total copia elements in the *O. coarctata* genome.

An unusual burst of LTR-RTs in *O. brachyantha*

We observed that *O. brachyantha* experienced an atypical increase in the copy number of certain retrotranposons (both copia and gypsy). The [range, mean±SD] for copy

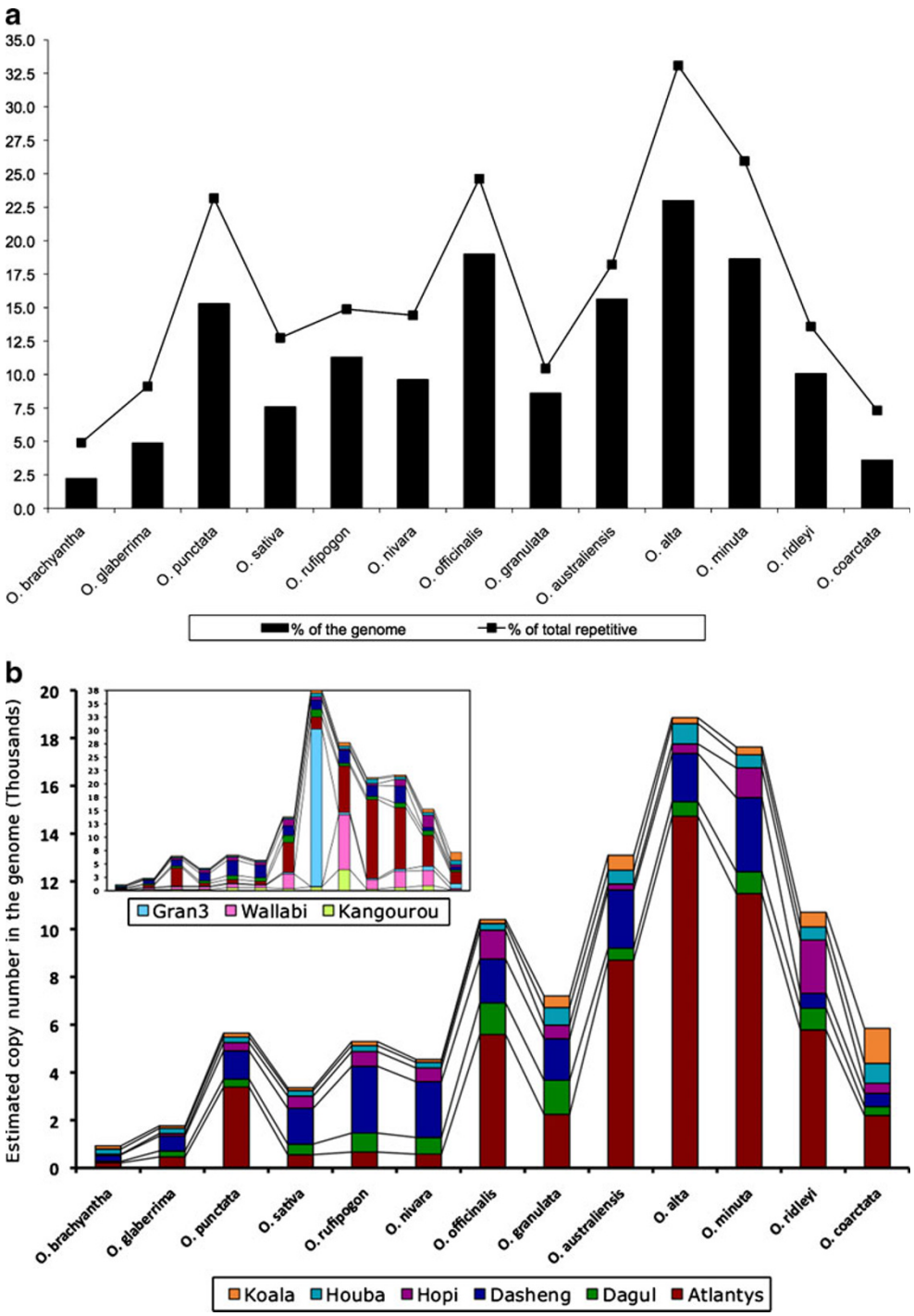


Fig. 5 **a** Six specific LTR retrotransposons (Ty3-gypsy—Atlantys, Dagul, Dasheng, and Hopi; Ty1-copia—Houba and Koala) as percent of the genome and as percent of total repetitive in the genome. **b** Comparison of preferential amplification of the six specific LTR retrotransposons in *Oryza* based on estimated copy numbers in the genome. *Inset* shows all nine LTR retrotransposons (including Kangourou, Wallabi, and Gran3) analyzed for differences in their estimated copy numbers in the genome. Species in both **a** and **b** are arranged in order of their increasing genome sizes.

numbers of the 111 LTR-RTs families was [2670, 110.1± 373.9] and [2270, 136.1±357.3] for Ty3-gypsy and Ty1-copia, respectively, indicating the amplification of specific families within each class. Besides the three Ty1-copia families previously described, we also identified four Ty3-gypsy families (SZ21, SZ240, GypsyA, and an uncharacterized family), forming 64.7% of the total gypsy LTR-RT copies in the *O. brachyantha* genome (Table S3). Elements from these seven families comprise ~60% of the total LTR-RT copies in the *O. brachyantha* genome, whereas elements belonging to remaining 104 families account for ≤40% of the total LTR-RTs. These results indicate that *O. brachyantha* is not exempt to LTR-RT amplification, as might be incorrectly interpreted from its overall low LTR-RT content. Amplification of specific LTR-RT families was seen in the *O. brachyantha* genome.

Rapid burst of MITEs in *O. brachyantha*

O. brachyantha, the smallest diploid genome has an exceptionally high amount of Class II MITE DNA transposons (4% of the genome) as compared to the other oryza genomes (Table 1). Based on the distribution of MITE families (Fig. S2), three families were identified in *O. brachyantha* that were each found to be greater than three percent of the total MITEs in *O. brachyantha*. Of these three families, rapid bursts of two MITE families, "OLO24" and "EXPLORER" was seen in *O. brachyantha* but not in the other species (Fig. 6). Lower numbers of MITEs in all the *Oryza* genomes, except *O. brachyantha* can be attributed to post-speciation sequence divergence of these elements such that they mutated beyond recognition. Their presence in high numbers in *O. brachyantha*, however, can be explained by formation of new MITEs through deletion of their corresponding Class II DNA transposons at a higher rate than occurring in other *Oryza* species.

To determine if MITEs in all the species except *O. brachyantha* are diverged from the *O. sativa* "MITE pool", or if they retain sequence similarity but are still present in low copy numbers, we analyzed the "OLO24" and "EXPLORER" families in all the species and calculated the percentage of total MITEs that were greater than 50% diverged in sequence and also the ones which were greater than or equal to 75% similar to the corresponding *O. sativa* MITEs (Table S4). This was done to determine if the failure to detect MITEs is due to sequence divergence or if they are preferentially amplified in *O. brachyantha*. Despite the observation that ~65% of the *O. brachyantha* "OLO24s" are greater than 50% diverged from *O. sativa* "OLO24," we could still identify them using the *O. sativa* dataset.

Correlaion between autonomous DNA transposons and repetitive content of genome

Autonomous and non-autonomous variants of seven types of non-MITE DNA transposons were identified using similarity searches to *O. sativa* MITEs (Fig. 7a, b). Of the seven types of non-MITE DNA transposons analyzed; CACTA, PILE, POLE and Tc1 had higher amounts of autonomous DNA transposons as compared to the non-autonomous fraction as opposed to hAT, Helitrons and MULEs, whose non-autonomous content was higher with the exception of *O. australiensis*, *O. granulata*, and *O. ridleyi* (Fig. 7b).

In general, throughout the genus, the four A-genome species and *O. brachyantha* had a higher percentage of non-autonomous DNA transposons as compared to autonomous DNA transposons. Interestingly, *O. brachyantha* happens to be the least repetitive and has the smallest genome in *Oryza*. For species with higher repetitive content, the amount of autonomous DNA transposons was higher (Fig. 7a).

Among the tetraploids, the general trend of higher percentage of autonomous DNA transposons is maintained in all the species, although *O. coarctata* is exceptional with the lowest repetitive content and the lowest amount of autonomous (49%) and the highest amount of non-autonomous (44%) DNA transposons among the tetraploids. On the other hand, *O. ridleyi*, the highest repetitive

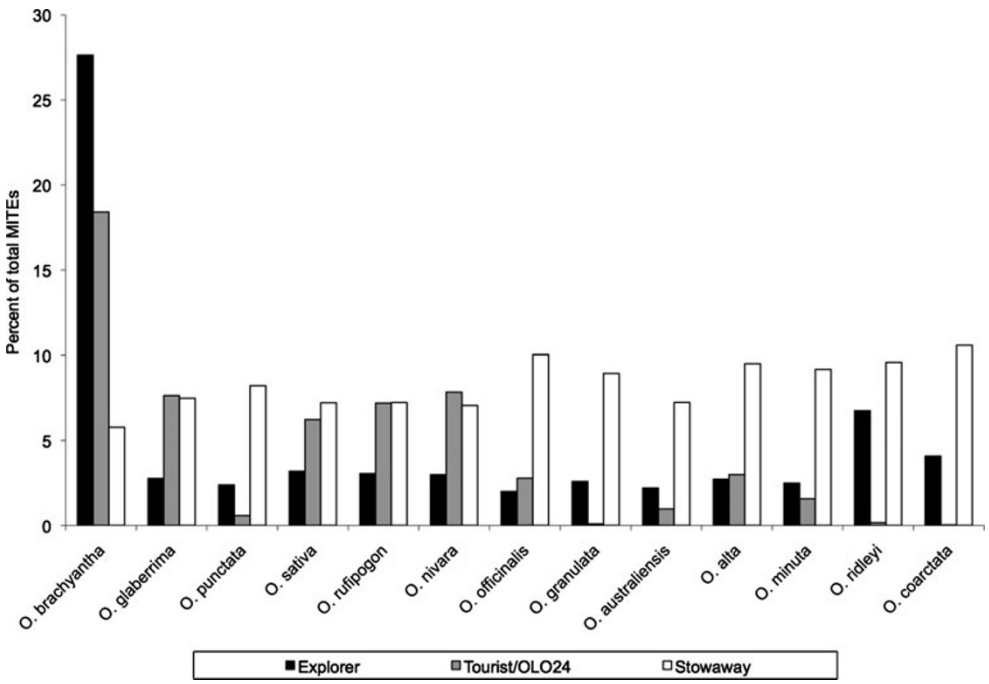
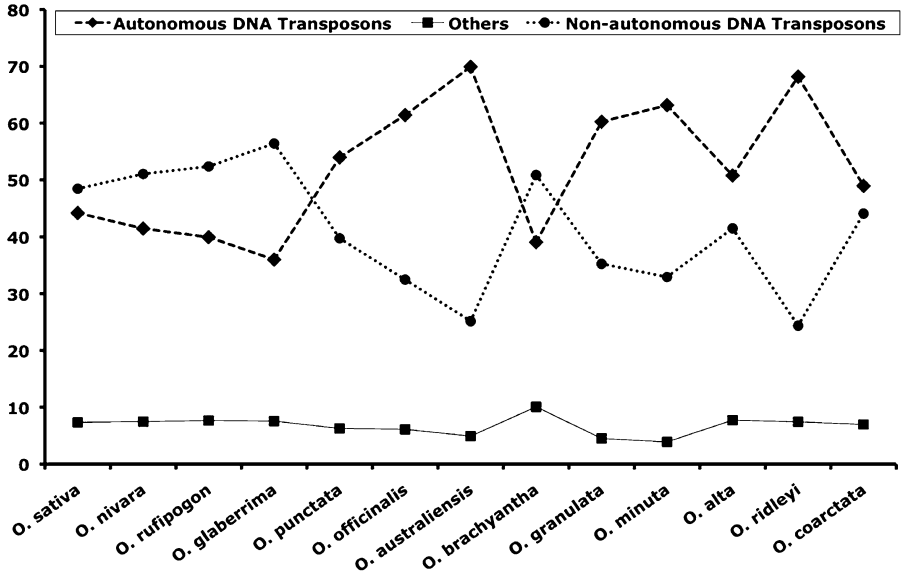


Fig. 6 Rapid burst of "OLO24" and "Explorer" in *O. brachyantha* as compared to other species, expressed as percentage of total MITEs in each species. The species are arranged in order of their increasing genome sizes.

a



b

Species	AUTONOMOUS (NON AUTONOMOUS) DNA Transposons*						
	CACTA	hAT	Helitron	MULE	PILE	POLE	Tc1
O. sativa	23.8% (4.34%)	3.21% (9.91%)	2.09% (14.72%)	10.06% (18.65%)	2.28% (0.53%)	1.78% (0.13%)	0.97% (0.17%)
O. nivara	21.81% (4.36%)	3.01% (9.84%)	2.06% (13.81%)	9.97% (22.17%)	1.68% (0.55%)	1.89% (0.16%)	1.04% (0.17%)
O. rufipogon	20.75% (5.16%)	2.94% (9.48%)	1.67% (14.85%)	10.16% (22.01%)	1.56% (0.63%)	1.84% (0.14%)	1.0% (0.13%)
O. glaberrima	17.62% (4.97%)	2.36% (10.04%)	2.0% (16.11%)	9.72% (24.39%)	1.56% (0.62%)	1.53% (0.15%)	1.2% (0.16%)
O. punctata	31.92% (5.22%)	3.5% (7.87%)	3.71% (8.25%)	11.02% (17.83%)	2.09% (0.48%)	1.23% (0.06%)	0.52% (0.04%)
O. officinalis	38.46% (4.53%)	4.29% (6.44%)	1.87% (6.64%)	12.73% (14.12%)	1.92% (0.59%)	1.6% (0.05%)	0.55% (0.09%)
O. australiensis	38.77% (5.14%)	7.4% (5.93%)	0.87% (3.66%)	18.28% (9.78%)	1.24% (0.34%)	3.16% (0.1%)	0.21% (0.19%)
O. brachyantha	15.25% (3.52%)	5.39% (14.16%)	2.08% (5.97%)	9.63% (25.83%)	2.51% (1.12%)	3.64% (0.25%)	0.55% (0.02%)
O. granulata	10.09% (2.29%)	13.8% (11.84%)	4.35% (6.48%)	21.04% (13.34%)	3.4% (0.54%)	6.72% (0.62%)	0.84% (0.12%)
O. minuta	36.9% (4.14%)	4.71% (6.88%)	2.74% (6.93%)	14.51% (14.3%)	2.01% (0.5%)	1.62% (0.1%)	0.69% (0.08%)
O. alta	23.8% (4.81%)	5.17% (8.11%)	2.91% (8.57%)	13.99% (19.27%)	2.4% (0.56%)	1.57% (0.14%)	0.94% (0.03%)
O. ridleyi	27.9% (4.49%)	4.83% (6.36%)	1.51% (2.71%)	29.76% (10.16%)	1.9% (0.45%)	1.82% (0.06%)	0.48% (0.14%)
O. coarctata	23.72% (2.82%)	4.79% (10.76%)	2.13% (8.34%)	14.24% (20.7%)	1.81% (1.13%)	1.49% (0.2%)	0.76% (0.12%)

*Values are expressed as percentage of the total Non-MITE DNA Transposons

Fig. 7 **a** Autonomous and non-autonomous DNA transposons as percentage of total non-MITE DNA transposons in *Oryza*. (Others include the non-MITE DNA transposons that could not be classified as either of these two classes.) The species are arranged according to the ploidy level. **b** Detailed analysis of autonomous and non-autonomous subtypes of seven non-MITE DNA transposons (CACTA, hAT, Helitron, MULE, PILE, POLE, Tc1) expressed as percentage of total non-MITE DNA transposons in *Oryza*.

tetraploid, had the highest percent autonomous (68%) and the lowest percent non-autonomous (25%).

O. australiensis with the highest repetitive content of all species had 70% and 25% of autonomous and non-autonomouus DNA transposons, respectively which is the highest autonomous and the lowest non-autonomous of all species (Fig. 7a). The percent autonomous non-MITE DNA transposons was found to be positively correlated with the total repetitive content of the genome with a correlation coefficient of 0.81.

Discussion

BESs, approximating 932Mbp, representing about 8–17% of each *Oryza* genome and corresponding to one sequence tag per every 4–8 kb (Kim et al. 2008), were used to analyze repetitive DNA across the genus. We reported the extent and distribution of 111 Class I LTR-retrotranspososn families, 98 subtypes of MITEs, and seven subtypes of Class II non-MITE DNA transposons throughout *Oryza*, and how each of these classes of TEs can be associated with the variation in genome size that *Oryza* has.

Since the BAC libraries are build using partial *Hin*dIII digestion, a bias in the sampling of the genomic sequences is expected. To make sure that this bias is not affecting the total% repetitive content of each genome, we used the Nipponbare whole genome sequence as well as randomly generated 800 bp fragments as controls to determine the total repetitive content and compared it to the Nipponbare BES repeat content (results not shown).

Repeat identification strategies

De novo and similarity-based detection are the two main criteria upon which most repeat identification strategies are based. As similarity-based searches are contingent upon the existence of precompiled repeat databases, they have a limited application for genomes lacking such a repeat anthology. The de novo approach is therefore the

method of choice for undescribed genomes. Most of the currently available de novo methods, such as RECON (Bao and Eddy 2002), REPuter (Kurtz et al. 2001), RepeatFinder (Volfovsky et al. 2001), and PILER (Edgar and Myers 2005), are being based on self-alignment approaches and are effectual only where sequence information is not limited in terms of sequence coverage or contiguity. Mathematically defined repeats thus provide an alternative to traditional similarity-based repeat finding methods that rely on precompiled repeat libraries as well as to most self-alignment based approaches. Even with the paucity of sequences available, k-mer frequencies can capture a rich statistical information on the repeat profiles of many plant genomes (Kurtz et al. 2008).

With the limited and fragmentary sequence information available for the *Oryza* species (Kim et al. 2008), we employed a combination of homology-based and de novo methods for repeat detection and categorization. Mathematically defined repeats calculated on the basis of frequency of overlapping 20-mers (Kurtz et al. 2008) in the BES datasets enabled us to catalog our BAC clones as mid, low, and high repetitive. Not surprisingly, the two most repetitive genomes in *Oryza*—*O. australiensis* and *O. granulata*—have the highest percentage of clones falling the mid and high repetitive categories as compared to other species, irrespective of the ploidy level. Such a classification will be useful for physical mapping and eventual sequencing as high repetitive clones can be avoided.

Size variation and repetitive content

Genome size in *Oryza* varies ~4.4 fold from 360 Mbp in *O. brachyantha* to 1,568 Mbp in *O. coarctata*. Other than ploidy, these differences can be attributed to structural changes (Bennetzen et al. 2005; Vitte and Panaud 2005) and genomic obesity caused by TEs (Kumar and Bennetzen 1999). The repetitive content of a species was found to be highly correlated to its genome size by a correlation coefficient of 0.9. Not surprisingly, our analysis by RepeatMasker and RECON showed the predominance of a particular class of TEs, the LTR-RTs, across the entire genus, congruent with other reports (Kim et al. 2008; Zuccolo et al. 2007). Widely documented as an ubiquitous feature of many complex plant genomes (Flavell et al. 1992; Voytas et al. 1992; Hirochika and Hirochika 1993; Suoniemi et al. 1998), LTR-RTs can occupy significant proportions (Ammiraju et al. 2007; Zuccolo et al. 2007; Kim et al. 2008), sometimes even more than half of the genomes of many species (Piegu et al. 2006; San Miguel et al. 1996; Vicient et al. 1999; Kalendar et al. 2000; Schulman and Kalendar 2005).

We also observed a positive correlation between the amount of autonomous DNA transposons to the repetitive content of the genome. One of the many possibilities for this observation could be that the high repetitive species with high amounts of autonomous DNA elements also have a higher rate of replicative transposition. Our data also indicate a near-perfect correlation of the amount of LTR-RTs to the repetitive content (0.9) and genome size of the species (0.8), indicating that they contribute significantly to genome size variation as well as repetitive content of a species. Analyses of the *O. australiensis* [EE] and *O. granulata* [GG] genomes demonstrated retrotranspositional bursts of Ty3-gypsy type of LTR-RTs in the EE (Piegu et al. 2006) and GG (Ammiraju et al. 2007) genomes subsequent to speciation, which accounts for significant proportions of the genome sizes of these species. The Tallymer and RECON data for *O. officinalis* [CC] and *O. alta* [CCDD] also indicate likely amplification of high copy repetitive sequences in these species, presumably retrotransposons also accounting for genome size variation. Apart from *in-silico* comparisons such as reported here, other experiments can be done to look for changes in localization/distribution, which can be detected by in situ experiments such as fluorescence in situ hybridization (FISH). For instance, during the course of time, tandem repeats can diverge and disperse, and the dispersed repeats can cluster together which can be differentiated by FISH.

O. coarctata, an exceptional case

Despite having the largest genome in *Oryza* [1,568 Mbp], *O. coarctata* has a very low repetitive content (43.7% of the genome) corresponding to only 681.7 Mbp of repetitive bases. If *O. coarctata* is excluded from our dataset, the repetitive content of a species is perfectly correlated, with a correlation coefficient of 1.0, to its genome size. Such an observation was also made previously (Zuccolo et al. 2007) but was attributed mainly to an incorrect genome size estimation of *O. coarctata*. We, however, on the basis of very thorough analyses based on mathematically derived repeats, self-alignment based de novo repeat detection, and homology to known *O. sativa* repeats, present a repeat profile for *O. coarctata*, which explains its low repetitive content as compared to other species. The repeats in *O. coarctata* are quite diverged from *O. sativa*, so much so that *O. coarctata* has a higher amount of unique repetitive sequences specific to its genome. *O. coarctata* also has many families of repetitive sequences present in low copies, which were discarded when the de novo repeats were parsed for copy number of five or greater. Tallymer data, based on the frequency distribution of 20-mers supports the abundance of low copy sequences (92.6% of total) in *O. coarctata*. We also observed a dearth of LTR-RTs in this species, in general, and Ty3-gyspsy types, in particular. The Ty3-gyspsies are the most abundant type of LTR-RTs in

Oryza accounting for the majority of repetitive content in all species except *O. brachyantha* and *O. coarctata*. Therefore, the abundance of low copy sequences may serve as a partial explanation for its low repetitive content, although we cannot exclude the possibility of an incorrect genome size estimation.

Repetitive element landscape in genus *Oryza*

Throughout the genus *Oryza*, Ty3-gypsy elements outnumber Ty1-copia except for *O. brachyantha and O. coarctata*, where we found that Ty1-copia and Ty3-gypsy elements are present in comparable amounts with no preferential Ty3-gypsy amplification. This could either be due to no/low rates of amplification or high rates of removal of such elements by unequal/illegitimate recombination. The latter can be tested by looking for the presence of solo LTRs or other remnants of Ty3–gypsy-like elements.

A very preliminary analysis of the solo-LTRs in all the species was done (data not shown). Due to the limitation imposed by the sequence read length, it was difficult to distinguish between the solo-LTRs and intact element when the solo-LTR was located toward the end of a BES. The number of solo-LTRs that were identified were highest in *O. granulata*, approximately 9 fold higher than *O. australiensis*. The results also revealed the highest ratio of solo:intactLTR in *O. brachyantha* and the lowest in *O. australiensis*. Interestingly, these results coincide with the repetitive content and size of these two species, with high rate of LTR deletions and very little LTR amplification in *O. brachyantha* and the reverse scenario for *O. australiensis*. *O. granulata* seems to be the most dynamic genome in *Oryza* with the rapid retrotransposon burst of Gran3 (Ammiraju et al. 2007) and at least nine retrotransposon (two Ty1-copia and seven Ty3-gypsy) families identified in our analysis coupled with the highest number of solo-LTRs present and still 73.8% of its genome is repetitive. Due to the specific repeat databases used for each species in this analysis, the total repeat content of *O. granulata* is higher than previously reported- 40.5% (Kim et al. 2008), suggesting the presence of high amounts of species-specific repetitive sequences in *O. granulata*.

Based on the Tallymer data, the genomes with the least repeat content have majority of their clones in the 0–40% repetitive range and a very few clones greater than 40% repetitive. *O. brachyantha*, with 91.9% of all its clones being low repetitive and a general depletion of LTR-RTs (Kim et al. 2008; Zuccolo et al. 2007), has three families of Ty1-copia and four families of Ty3-gypsy that have amplified to reach 56% and 64% of the total copia and gypsy copies, respectively. This suggests that the *O. brachyantha* genome is not immune to the amplification of LTR-RTs. Therefore, there must be mechanisms that keep its repetitive content low and genome size under check. The rate of removal of TEs by deletions and/or illegitimate recombination may be higher, or the amplification of LTR-RTs beyond a certain level may be detrimental such that any particular element is removed or becomes stagnant. A high rate of removal was indicated by the high ratio of solo:intact LTRs in the *O. brachyantha* genome. Thus, the MITE outburst, the fewer non-MITE DNA transposons, the LTR-RT bursts of only specific families, and the higher ratio of solo:intact LTRs (data not shown) may result in, and be maintenance of, a smaller genome.

Similar to *O. brachyantha*, *O. glaberrima* also has a small genome size [364 Mbp], low repetitive content (43.4% of the genome), and is deficient in both Ty1-copia and Ty3-gypsy type of LTR-RTs. However, we did not observe a MITE expansion in *O. glaberrima* such as seen in *O. brachyantha*. MITEs, in general, were more abundant in *O. brachyantha* as compared to all other species. Analysis of sequence divergence of MITEs shows that, although >50% diverged *O. brachyantha* MITEs could still be identified using the *O. sativa* dataset, the presence of highly similar sequences to *O. sativa* is not necessarily implicated in their increase in copy number as compared to other species. Thus, sequence divergence does not fully explain the depletion of MITEs in all species, except *O. brachyantha*. Post-speciation changes through mutations and/or deletions can render these elements nearly unidentifiable accounting for the lower number of MITEs in some, but not all, cases. However, it should be noted that the *Oryza* "MITE pool" is conserved, although variants from this pool can give rise to "Novel MITEs."

Besides divergence, alternate mechanisms must exist to explain the rapid burst of MITEs in *O. brachyantha.* Due to the structural characteristics of MITEs being similar to the defective Class II DNA transposons (lacking internal region and transposase) and extensive conservation of sequence and size among members of the same subfamily, it is suggested that MITEs have originated from a limited number of progenitor DNA transposons (Feschotte et al. 2002a, b). Due to the inability to encode its own transposase, transposition of MITEs is catalyzed by the transposase encoded by the transposon from which it is derived (Craig et al. 2002; Feschotte et al. 2002a, b) or even a distantly related self-restrained autonomous DNA transposon (Yang et al. 2009). This deletion mechanism happening at a higher rate in *O. brachyantha* may, however, help to explain both the genome size reduction and rapid burst of MITEs in *O. brachyantha*.

So the question arises: Why is such a mechanism exclusive to *O. brachyantha*? Are specific environmental conditions, edaphic factors, or biotic stress involved? If yes, then such factors can be proposed to play a role in genome size variation due to their effect on the amplification/

deletion of specific transposons, although detailed analyses are needed to arrive at any such conclusions. Despite lacking coding capacity, MITEs can amplify in large numbers by manipulating even the distantly related and self-restrained autonomous DNA transposons (Yang et al. 2009).

Preferential amplification of specific elements

It is evident from our analyses that the rate at which different elements amplify with respect to other elements within a species or with respect to the same element across species varies considerably. In *O. australiensis*, the largest diploid genome, an LTR-RT-driven genome expansion had been reported previously (Piegu et al. 2006). Our analysis shows a rapid apmlification of at least 14 families of LTR-RTs (two Ty1-copia and 12 Ty3-gypsy types), which supports the role of these elements in genome size variation. Based on the clustering analysis of the unique de novo repeats (data not shown) and copy number distribution of Ty1-copia and Ty3-gyspy families, we, however, propose the amplification of more than just three families of retrotransposons in *O. australiensis*. Due to limitation on sequence read length which is too small to span full-length retrotransposons, we were unable to identify these full-length elements. Explosive proliferation of one or more LTR-RT families subsequent to speciation has also been observed in other genera such as *Zea* (San Miguel et al. 1996) and *Gossypium* (Hawkins et al. 2006), where amplification of these families in comparison to others occupy significant portions of each genome. However, rapid amplification is not the only determinant of genome expansion. Different factors, such as rapid genomic DNA loss through unequal/illegitimate recombination and internal deletions, also act as counter forces in determining the relative retrotranspositional rates of different elements/families (Devos et al. 2002; Ma et al. 2004).

Thus, besides polyploidization, LTR-RTs can be the primary candidates for variation in genome sizes, owing to their differential rates of proliferation in different species (Fig. 8). Any particular element may be present in all the species but due to varying rates of retrotransposition may amplify only in one or a few species. This could be due to varying selection pressures on different elements. Sequence divergence between elements belonging to the same family can provide clues as to the timing of transposition. Actively and/or recently transposed elements belonging to the same family should be nearly identical in sequence, whereas inactive transposons may be diverged to such an extent that they will be unidentifiable.

Another possibility is that certain elements are exclusive to particular species, which then raises the question: where did they come from and/or why did they get deleted from all other species? Horizontal transfer, a major source of evolution and speciation in bacteria (Lawrence 2002), can be one of the mechanisms for the origin of TEs due to their mobility and capacity to integrate into the host DNA (Roulin et al. 2008). In plants, there are documented cases of horizontal transfer of both mitochondrial (Mower et al. 2004; Richardson and Palmer 2007) as well as nuclear-encoded genes both between (Diao et al. 2006) and within genera (Roulin et al. 2008). To investigate such an origin of

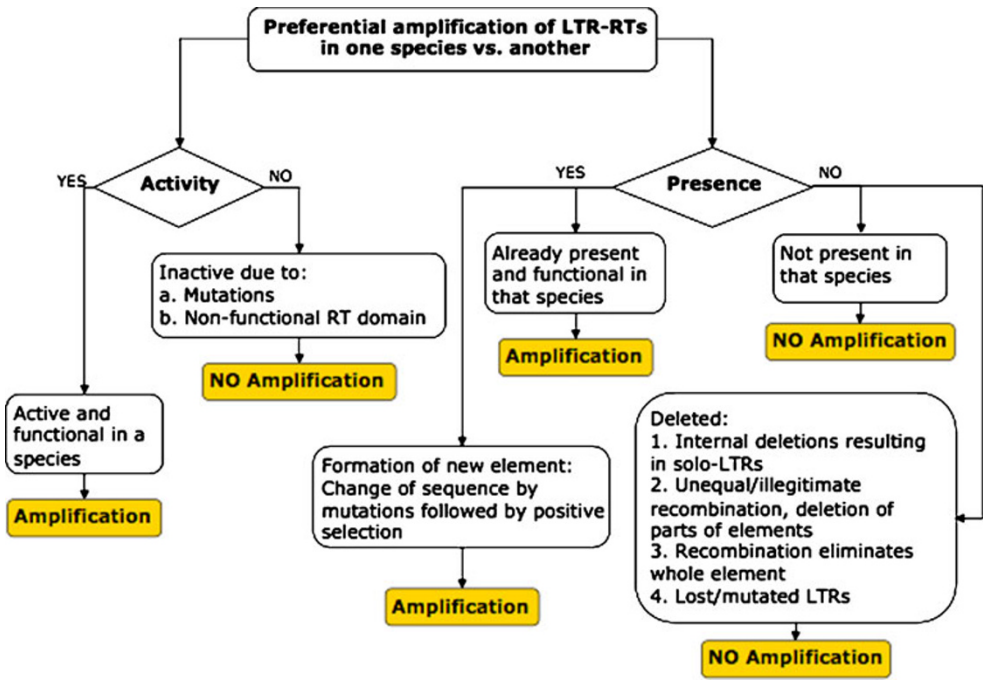


Fig. 8 Preferential amplification of certain LTR retrotransposons as compared to others.

TEs, comparisons can be made using such full-length LTR-RTs and/or look for their remnants in other species. Because of the short reads in our dataset, we were not able to do this, but as genome sequencing progresses, this will be an interesting question to follow.

Based on our analysis, there are three scenarios for amplification of one element as compared to others in the same species or across species. Such a process is either (a) *favored by the genome*—if so, all elements should be high copy in one genome vs. the others, (b) *element-dependent*—if so, a particular element should be high copy in all genomes, or (c) *an interaction between the element and genome*—this seems most plausible in that a particular element in a particular genome environment results in amplification. However, activation/mobilization of TEs as a result of "genomic shock" due to wide hybridization (McClintock 1984; Shan et al. 2005), tissue culture (Jiang et al. 2003; Kikuchi et al. 2003), and γ-ray irradiation (Nakazaki et al. 2003) has also been reported. Therefore, depending on the factors that potentially influence element copy number and/or activity, we can say that the repetitive elements may or may not be predisposed to certain genomes and that the genome × environment interaction may also play a role in regulating their copy number.

Practical applications of the data generated

The data presented here will help further our understanding of genome organization and evolution in *Oryza*. Due to a rapid rate of divergence of repetitive DNA relative to gene sequences (Ma and Bennetzen 2004), they maintain the dynamic nature of the genomes through balancing forces of genome expansion and contraction (Vitte and Panaud 2005; Devos et al. 2002; Ma et al. 2004). Identification and characterization of repetitive sequences therefore will aid the sequence assembly programs and further analysis of genomic data and will simplify gene annotations during future genome sequencing of the wild relatives of *O. sativa*. Characterization of BAC clones into low, mid, and high repetitive will be of constructive use in eliminating the overlapping and redundant high repetitive clones from the BAC-based physical maps of *Oryza* (Kim et al. 2008; Soderlund et al. 2006; SYMAP- http://www.agcol.arizona.edu/software/symap/). The utility of the species-specific repeat databases lies in the fact that association of these repeats with differentially expressed genes in a species will help unravel mechanisms of gene regulation.

Conclusions and future prospects

Analysis of repetitive content of the *Oryza* genomes not only helped us identify and classify repetitive sequences into different classes but also indicated the possibility of how these sequences may be involved in genome size variation. We provide evidence to show that besides the Class I LTR-RTs (Wessler et al. 1995; Piegu et al. 2006; Ammiraju et al. 2007; Zuccolo et al. 2007; Kim et al. 2008), Class II DNA transposons, both MITEs (Wessler et al. 1995; Yang et al. 2009) and non-MITEs, can influence the genome size of a species through their expansion, loss, and movement in the genome. Preferential amplification of specific LTR-RTs in the largest diploid genome and rapid bursts of MITEs in the smallest diploid genome were observed as alternate mechanisms controlling genome size in the genus *Oryza*, apart from polyploidization. Although we identified 38 LTR-RT families that are amplified in 1,500 or more copies throughout *Oryza*, it still remains to be determined if preferential amplification of some of these families is due to the predisposition of its elements to certain lineages or vice versa.

Materials and methods

BAC libraries for wild relatives of *O. sativa*

A set of BAC libraries from 13 species representing the ten genome types of *Oryza* were obtained from Arizona Genomics Institute (AGI) and were used for this analysis. Each library represents a minimum of ten genome equivalents and has an average insert size ranging from 123 to161 kb (Ammiraju et al. 2006). BESs were generated from these libraries, resulting in an average of 731,430 forward, 719,415 reverse, and 690,184 clones with paired reads, with 650 bp as the average read length after trimming (Kim et al. 2008).

Mathematically derived repeats

Tallymer (Kurtz et al. 2008), a program based on enhanced suffix arrays (Abouelhoda et al. 2004), was used to compute the 20-mer occurrence counts and construct a frequency index of each 20-mer for the entire *Oryza* BES dataset. These frequencies were plotted logarithmically on a genomic scale to distinguish regions of high TE content from low copy regions. BAC-end pairs were merged by inserting a gap (stretch of Ns) between the forward and reverse reads, and will be referred to as a BAC clone for the purpose of this analysis. Based on the 20-mer frequency distribution, clones in the BAC libraries of all species were further categorized into low, mid and high repetitive clones.

Compilation of species-specific repeat databases for all *Oryza* species

A comprehensive custom repeat database was compiled, first for *O. sativa* ssp. Nipponbare as the basal dataset. This

was done using *Oryza* repeat database (3,752 sequences) from Dr. Robin Buell's lab at Michigan State University (http://plantrepeats.plantbiology.msu.edu/oryza.html), two TE databases [courtesy of Dr. Tom Bureau from McGill University, Canada (158 sequences) and Dr. Ning Jiang from Michigan State University, USA (1,487 sequences)], CRRs (234 sequences; Nagaki et al. 2005), and LTR-RTs (261 sequences) identified from the whole genome sequence of Nipponbare [International Rice Genome Sequencing Project (IRGSP) pseudomolecule, version 4] using LTR_STRUC (McCarthy and McDonald 2003). Overlapping/redundant elements were removed from these datasets using an 80% similarity index as the cutoff value. Elements greater than or equal to 80% similar were regarded as same elements and were removed. Elements less than 80% similar were identified as being unique and were included in the Nipponbare custom repeat database (5,892 sequences).

RECON (Bao and Eddy 2002) was then used to identify de novo repeats from the *Oryza* BES datasets. To increase the speed and efficiency of the program, the BLAST output was parsed to discard self-hits as well as hits with an e-value greater than 1e-5. The RECON output, which identified repetitive elements and classified them into distinct families, was parsed for sequences greater than 40 bp in length that were found at least five times per family. Overlap between the de novo and the Nipponbare custom library was determined using RepeatMasker. Sequences left unmasked by this process and thus not a part of the custom repeat database were extracted and annotated using BLASTN (Altschul et al. 1997) at an e-value = $1e^{-5}$ against the all-plant repeat database at http://plantrepeats.plantbiology.msu.edu/. For each individual species, these annotated de novo repeats were combined with the Nipponbare repeat library to form a species-specific repeat database. This database was used for homology-dependent repeat search in that particular species using RepeatMasker (Smit et al. at http://repeatmasker.org).

Analysis of repetitive sequences

RepeatMasker (version 3.1.9; WuBlast as the search engine) was used to mask the repetitive sequences for the entire *Oryza* BES dataset, using an exclusive database for each species, as described above. Customized Perl scripts were used to parse the RepeatMasker output and to remove/minimize any overlaps between different repeat co-ordinates. The masked sequences were identified and classified into different types of repeats in each of the species. Low-complexity repetitive regions and SSRs were also identified, and their relative abundance and density (number of SSRs/Mbp of the genome) were determined. The most frequent type of SSR motif within each di-, tri-, and tetranucleotide repeats was further identified for all the species.

Different classes of TEs (Class I retrotransposons and Class II DNA transposons) were analyzed in detail using subsets of the repeat database. A number of 58 subfamilies of Ty3-gypsy and 53 subfamilies of Ty1-copia type were analyzed for preferential amplification in one species vs. all others. Nine specific elements from these families (seven Ty3-gypsy and two Ty1-copia types) were compared across the species to see if they are present/absent or differentially amplified across the species. The autonomous and non-autonomous subtypes of CACTA, hAT, MULE, PILE, POLE, Tc1, and Helitrons belonging to non-MITE DNA transposons were identified by homology-based searches using RepeatMasker. Divergence analysis of MITE subtypes was done using BLASTN at e = $-^{10}$ to examine their preferential amplification within the *O. brachyantha* genome. MITE sequences that are either less than 50% or greater than 75% similar to the corresponding *O. sativa* MITEs were identified to test for sequence divergence of MITEs within *Oryza*.

Acknowledgments This work was funded by the NSF Plant Genome Award # DBI-0321678 to SAJ, RW, and LS. We thank Dr. Ning Jiang from the Michigan State University, USA and Dr. Tom Bureau from McGill University, Canada for generously sharing their TE databases. We also thank the AGI technical staff, especially members of the BAC/EST Resource, BAC Library Construction, Sequencing, Genome Finishing, and Annotation Centers for supporting this project.

References

Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. J Discrete Alg. 2004;2:53–86.

Aggarwal RK, Brar DS, Khush GS. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. Mol Gen Genet. 1997;249:65–73.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 1997;25:3389–402.

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J. 2007;52:342–51.

Ammiraju JSS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, et al. 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. 2006;16:140–7.

Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 1991;9:211–5.

Bao Z, Eddy S. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.

Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol. 2000;42:251–69.

Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 2005;95:127–32.
Charlesworth B. The evolution of sex chromosomes. Science. 1991;251:1030–3.
Charlesworth B, Langley CHA. The population genetics of *Drosophila* transposable elements. Annu Rev Genet. 1989;23:251–87.
Charlesworth B, Langley CH, Stephan W. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics. 1986;112:946–62.
Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371:215–20.
Craig NL, Craigie R, Gellert M, Lambowitz AM. Mobile DNA II. Washington, DC: American Society for Microbiology Press; 2002.
Crow JF, Simmons MJ. The genetics and biology of *Drosophila*. London: Academic; 1983.
Devos KM, Brown JK, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 2002;12:1075–9.
Diao X, Freeling M, Lisch D. Horizontal transfer of a plant transposon. PLoS Biol. 2006;4:e5.
Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21:i152–8.
Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9:397–405.
Feschotte C, Zhang X, Wessler SR. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposon. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington, DC: American Society for Microbiology Press; 2002a. p. 1147–58.
Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002b;3:329–41.
Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A. Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucl Acids Res. 1992;20:3639–44.
Galasso I, Schmidt T, Pignone D, Heslop-Harrison JS. The molecular cytogenetics of *Vigna unguiculata* (L) Walp: the physical organization and characterization of 18s–58s–25s rRNA genes, 5s rRNA genes, telomere-like sequences, and a family of centromeric repetitive DNA sequences. Theor Appl Genet. 1995;91:928–35.
Gao D, Gill N, Kim H-R, Walling JG, Zhang W, Fan C, Yu Y, Ma J, SanMiguel P, Jiang N, Cheng Z, Wing RA, Jiang J, Jackson SA. A lineage-specific centromere retrotransposon in *Oryza brachyantha*. Plant J. 2009;9999.
Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 2006;16:1252–61.
Hirochika H, Hirochika R. Ty1-copia group retrotransposons as ubiquitous components of plant genomes. Jap J Genet. 1993;68:35–46.
Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimeric insertion. Nat Genet. 1994;7:143–8.
IRGSP. International rice genome sequencing project: the map based sequence of the rice genome. Nature. 2005;436:793–800.
Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. Nature. 2003;421:163–7.
Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci USA. 2000;97:6603–7.
Kapitonov VV, Jurka J. Helitrons on a roll: eukaryotic rolling—circle transposons. Trends Genet. 2007;23:521–9.
Khush GS. Origin, dispersion, cultivation and variation of rice. Plant Mol Biol. 1997;35:25–34.
Kikuchi K, Terauchi K, Wada M, Hirano H-Y. The plant MITE mPing is mobilized in anther culture. Nature. 2003;421:167–70.
Kim H, Hurwitz B, Yu Y, Collura K, Gill N, SanMiguel P, et al. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. Genome Biol. 2008;9:R45.
Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet. 1999;33:479–532.
Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl Acids Res. 2001;29:4633–42.
Kurtz S, Narechania A, Stein J, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9:517.
Lawrence JG. Gene transfer in bacteria: speciation without species? Theor Pop Biol. 2002;61:449–60.
Li CB, Zhang DM, Ge S, Lu BR, Hong DY. Differentiation and intergenomic relationships among C, E and D genomes in the *Oryza officinalis* complex (Poaceae) as revealed by multicolor genomic in situ hybridization. Theor Appl Genet. 2001;103:197–203.
Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430:471–6.
Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA. 2004;101:12404–10.
Ma J, Devos KM, Bennetzen JL. Analyses of LTR retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 2004;14:860–9.
Mackay TFC. Genet Res. 1986;48:77–87.
Matyasek R, Gazdova B, Fajkus J, Bezdek M. NTRS, a new family of highly repetitive DNAs specific for the T1 chromosome of tobacco. Chromosoma. 1997;106:369–79.
McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003;19:362–7.
McClintock B. The significance of responses of the genome to challenge. Science. 1984;226:792–801.
Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution: grasses, line up and form a circle. Current Biol. 1995;5:737–9.
Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37:997–1002.
Mower JP, Stefanovic S, Young GJ, Palmer JD. Plant genetics: gene transfer from parasitic to host plants. Nature. 2004;432:165–6.
Nagaki K, Neumann P, Zhang D, Ouyang S, Buell CR, Cheng Z, et al. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. Mol Biol Evol. 2005;22:845–55.
Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, et al. Mobilization of a transposon in the rice genome. Nature. 2003;421:170–2.
Nayar M. Origin and cytogenetics of rice. Adv Genet. 1973;17:153–292.
Okamura K, Nakai K. Retrotransposition as a source of new promoters. Mol Biol Evol. 2008;25:1231–8.
Panaud O, Vitte C, Hivert J, Muzlak S, Talag JD, Brar DS, et al. Genomic differentiation between rice (*Oryza sativa* L.) and foxtail

millet (*Setaria italica* L. Beauv.) revealed through representation difference analysis. Mol Gen Genomics. 2002;268:113–21.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 2006;16:1262–9.

Richardson AO, Palmer JD. Horizontal gene transfer in plants. J Exp Bot. 2007;58:1–9.

Roulin A, Piegu B, Wing RA, Panaud O. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus *Oryza*. Plant J. 2008;53:950–9.

San Miguel P, Bennetzen JL. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot. 1998;82:37–44.

San Miguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996;274:765–8.

Schulman AH, Kalendar R. A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenet Genome Res. 2005;110:598–605.

Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, et al. Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (*Zizania latifolia* Griseb.). Mol Biol Evol. 2005;22:976–90.

Shapiro JA, Sternberg R. Why repetitive DNA is essential to genome function? Biol Rev. 2005;80:227–50.

Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: a system for discovering and viewing syntenic regions of FPC maps. Genome Res. 2006;16:1159–68.

Suoniemi A, Tanskanen J, Schulman AH. Gypsy– like retrotransposons are widespread in the plant kingdom. Plant J. 1998;13:699–705.

Tateoka T. Taxonomic studies of *Oryza* III. Key to the species and their enumeration. Bot Mag Tokyo. 1963;76:165–73.

Tateoka T. Notes of some grasses: XVI. Embryo structure of the genus Oryza in relation to their systematics. Am J Bot. 1964;51:539–43.

Uozu SH, Ikehashi N, Ohmido H, Ohtsubo E, Ohtsubo FK. Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. Plant Mol Biol. 1997;35:791–9.

Vaughan DA, Morishima H, Kadowaki K. Diversity in the *Oryza* genus. Curr Opin Plant Biol. 2003;6:139–46.

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, et al. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell. 1999;11:1769–84.

Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res. 2005;110:91–107.

Volfovsky N, Haas B, Salzberg S. A clustering method for repeat analysis in DNA sequences. Genome Biol. 2001;2:research0027.0021–research0027.0011.

Voytas DF, Cummings MP, Koniczny A, Ausubel FM, Rodermel SR. Copia-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci USA. 1992;89:7124–8.

Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A de novo Alu insertion results in neurofibromatosis type 1. Nature. 1991;353:864–6.

Wang ZX, Kurata N, Saji S, Katayose Y, Minobe Y. A chromosome 5-specific repetitive DNA sequence in rice (*Oryza sativa* L). Theor Appl Genet. 1995;90:907–13.

Wessler SR, Bureau TE, White SE. LTR retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev. 1995;5:814–21.

Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, Kudrna D, et al. The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. Plant Mol Biol. 2005;59:53–62.

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH. Date of the monocot–dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA. 1989;86:6201–5.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008;319:1527–30.

Yang G, Lee YH, Jiang Y, Shi X, Kertbundit S, Hall TC. A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter. Plant Cell. 2005;17:1559–68.

Yang G, Zhang F, Hancock CN, Wessler SR. Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. Proc Natl Acad Sci USA. 2007;104:10962–7.

Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. Tuned for transposition: molecular determinants underlying the hyperactivity of a stowaway MITE. Science. 2009;325:1391–4.

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

Zuccolo A, Ammiraju J, Kim H, Sanyal A, Jackson S, Wing R. Rapid and differential proliferation of the Ty3-gypsy LTR retrotransposon atlantys in the genus *Oryza*. Rice. 2008;1:85–99.