Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice

Benoit Piegu,[1] Romain Guyot,[1] Nathalie Picault,[1] Anne Roulin,[1] Abhijit Saniyal,[3] Hyeran Kim,[4] Kristi Collura,[4] Darshan S. Brar,[2] Scott Jackson,[3] Rod A. Wing,[4] and Olivier Panaud[1,5]

[1]*Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS-IRD, Université de Perpignan, Perpignan 66860, France;* [2]*Plant Breeding Genetics and Biochemistry Division, International Rice Research Institute, Manila 1099, Philippines, USA;* [3]*Agricultural Genomics, Purdue University, West Lafayette, Indiana 47907, USA;* [4]*Arizona Genomics Institute, Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA*

Retrotransposons are the main components of eukaryotic genomes, representing up to 80% of some large plant genomes. These mobile elements transpose via a "copy and paste" mechanism, thus increasing their copy number while active. Their accumulation is now accepted as the main factor of genome size increase in higher eukaryotes, besides polyploidy. However, the dynamics of this process are poorly understood. In this study, we show that *Oryza australiensis*, a wild relative of the Asian cultivated rice *O. sativa*, has undergone recent bursts of three LTR-retrotransposon families. This genome has accumulated more than 90,000 retrotransposon copies during the last three million years, leading to a rapid twofold increase of its size. In addition, phenetic analyses of these retrotransposons clearly confirm that the genomic bursts occurred posterior to the radiation of the species. This provides direct evidence of retrotransposon-mediated variation of genome size within a plant genus.

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

The considerable diversity of genome sizes in higher eukaryotes, in particular the lack of correlation between genome size and biological complexity, first formulated as the *c-value paradox* (Thomas 1971), was one of the most disturbing discoveries in the early days of structural genomics. It has since been established for many organisms that Transposable Elements (TEs) are the main components of complex genomes and that transposition can be regarded as the predominant force driving their structural changes, besides polyploidy (Bennetzen et al. 2005; Vitte and Panaud 2005). In this regard, a particular class of TEs, the retrotransposons, is considered an important factor of genomic inflation in both plants and animals because of their propensity to increase their copy number during transposition (Kumar and Bennetzen 1999). This is well documented in grasses, where Long Terminal Repeat (LTR)-retrotransposons can compose more than half of the genome of some species (SanMiguel et al. 1996; Vicient et al. 1999; Kalendar et al. 2000; Schulman and Kalendar 2005). Much less is known, however, about the dynamics of this process, i.e., the timing of the genomic expansions caused by the activity of LTR-retrotransposons. Do genomes expand gradually through slow accumulation of retrotransposons, or are the expansions a saltatory process caused by large, sudden bursts of retrotransposition? Many studies have shown that in large plant genomes, LTR-retrotransposon families often contain thousands (or tens of thousands) of copies with high sequence identity, which suggests that they originate from a recent massive retrotransposition event (SanMiguel et al. 1998; Vicient et al. 1999). However, none of these studies have so far provided direct evidence at a genome scale that the activity of LTR-retrotransposons could increase genome size to such a large extent over a short evolutionary time.

The genus *Oryza*, to which the Asian-cultivated rice species *O. sativa* belongs, contains 24 species (17 diploids and seven tetraploids) distributed throughout the world (Ge et al. 1999). This genus has been extensively studied because of the economical importance of Asian rice and because the wild relatives of the cultivated species constitute a useful reservoir of genetic diversity which is exploited for rice breeding (Brar and Khush 1997). The taxonomic status and phylogenetic relationships among the 24 species of the genus have been established using phenotypic, cytogenetic, and molecular data (Ge et al. 1999). In addition, their estimated genome size ranges from 357 Mbp for *O. glaberrima* (diploid, genome type AA) to 1283 Mbp for *O. coarctata* (tetraploid, genomes HHKK) (Ammiraju et al. 2006). It is 390 Mbp for the model species *O. sativa* (diploid, genome AA), whose genome has been sequenced (International Rice Genome Sequencing Project 2005). The four largest genome sizes are found in the tetraploid species for which such data is available, which illustrates well that polyploidy in plants is an important factor of genome size variation. However, there is also a significant variation of genome size within the diploid *Oryza* species, with a 2.7-fold variation between the smallest (i.e., *O. glaberrima* 357 Mbp) and the largest genome (i.e., 965 Mbp for *O. australiensis*, genome EE) (Ammiraju et al. 2006). Moreover, as shown in Figure 1, this large difference in size between the genome of *O. australiensis* compared with that of the most closely related diploid species (i.e., *O. sativa*, *O. glaberrima*, *O. officinalis*, and *O. punctata*) suggests that dramatic and recent structural genomic changes have occurred specifically in the lineage of *O. australiensis*.

[5]Corresponding author.
E-mail panaud@univ-perp.fr; fax 33-04-468664899.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5290206.

16:1262–1269 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

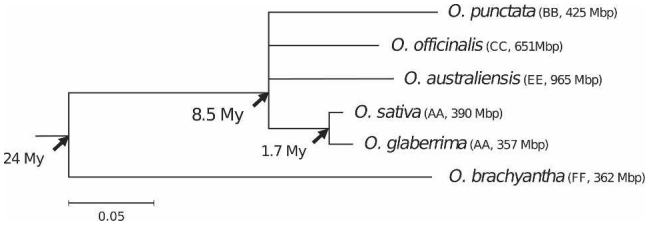

Figure 1. Phylogenetic tree of six diploid *Oryza* species, established with the *ADH2* gene. Only nodes with a bootstrap value >70% are shown. The tentative dates of the radiation event are given for each node. The computational methods are given in the Methods section.

By combining a molecular cloning approach and in silico analyses, we have characterized three LTR-retrotransposon families, whose transcriptional activity has generated over 500 Mbp of genomic sequence for the species. Moreover, detailed phenetic studies of these three families in the genus *Oryza* allowed us to characterize the dynamics of this process by providing evidence for several distinct and successive retrotranspositional bursts which have occurred within the last three million years, after the speciation of *O. australiensis*.

Results and Discussion

Three LTR-retrotransposon families compose 60% of the *O. australiensis* genome

In order to investigate the cause of the genomic expansion in *O. australiensis*, we first applied Representational Difference Analysis (RDA). RDA (Lisitsyn et al. 1993; Panaud et al. 2002) is a PCR-based cloning procedure that allows isolation of sequences which are specific to one genome (the tester) compared with another (the blocker). It is based on subtractive hybridization of genomic fractions (the representation) of both the tester and the blocker. These representations are obtained after digestion of total genomic DNA, followed by ligation with adapters and PCR amplification of the ligated products using a primer homologous to the adapter. Prior to the subtraction, a new set of adapters is ligated to the representation of the tester DNA only, thus allowing the amplification of specific sequences. Using *O. australiensis* genomic DNA as tester and *O. sativa* as blocker we obtained a library primarily composed of a single 359-bp *O. australiensis*-specific fragment (data not shown). This suggested that, at least in the representation we obtained, the difference in composition between these two genomes could be explained by the presence of a sequence which is highly repeated in the genome of *O. australiensis* but which is absent (or present at a much lower copy number) from the *O. sativa* genome. Sequencing of this RDA sequence revealed that it is part of the LTR of *RIRE1*, a previously characterized *TY1/Copia* type LTR-retrotransposon (Uozu et al. 1997). In order to estimate the copy number of *RIRE1*, dot-blot assays were performed using either LTR or internal region probes (Supplemental data #1). We found that there are 30,000 ± 3000 complete *RIRE1* copies and 10,000 ± 1000 apparent single LTRs, considered as recombinational variants called solo-LTRs (Shirasu et al. 2000), which makes this element one of the most highly repeated within a plant genome. *RIRE1* therefore appears to contribute a total of about 265 Mbp, i.e., 27% of the genome of *O. australiensis* (Table 1). This observation led us to use another strategy to identify other repeated sequences that could have contributed to the genomic expansion of the species in addition to *RIRE1*.

The *Oryza* Map Alignment Project (OMAP, http://www.omap.org) has generated a large amount of genomic resources for 12 *Oryza* species, including 137,000 BAC end sequences (BES) of *O. australiensis* (Ammiraju et al. 2006). In silico analyses of these *O. australiensis* BES allowed identification of 25 highly redundant sequences which were not homologous to *RIRE1*. These were successively extended and merged using the BES data, allowing reconstruction of the putative sequence of three *TY3/Gypsy* type LTR-retrotransposons, named *Kangourou*, *Wallabi*, and *Dingo* (Fig. 2; Supplemental data #2). Three BAC clones from the *O. australiensis* genomic libraries were then fully sequenced to validate the structure of these three new elements. One clone harbors the *HD1* locus (and was chosen because it also harbors at least one copy of *RIRE1*). The other two clones were chosen based on the homology of one of their BES with either *Wallabi* or *Kangourou*. Comparative sequence analysis with the *O. sativa* genome suggests that none of the three clones are located in a pericentromeric region. Overall, 350 kb (350,792 bp) of genomic sequence were generated and analyzed (Fig. 3). Seven complete copies and five solo-LTRs of *RIRE1*, *Kangourou*, and *Wallabi* were identified from these genomic clones, thus validating the sequences inferred from the in silico approach. Moreover, *Kangourou* and *Wallabi* were searched for homology with already known rice retrotransposons. We found that *Kangourou* exhibits a low but significant sequence identity with the *Retrosat1/RIRE2* retrotransposon family (65% overall identity in the internal region with 75% in the GAG-POL region). This is an indication that these two families are homologous. However, given the low sequence identity between *Kangourou* and *Retrosat1*, we kept a distinct name for each.

Based on dot-blot assays, *Kangourou* and *Wallabi* were estimated to contribute a total of 90 ± 9 Mbp and 250 ± 25 Mbp, i.e., 9% and 26% of the genome of *O. australiensis*, respectively (Table 1). The copy number of *Dingo* was estimated to be between 3700 and 4300 and could thus clearly be considered as highly

Table 1. Description of the three retrotransposons, *RIRE1*, *Kangourou*, and *Wallabi*, in the genome of *O. australiensis*

		Size in bp	Number of copies	Size in the genome	Total
RIRE1	Full element	8300	30,000 ± 3000	250 ± 25 Mbp	265 ± 26.5 Mbp
	Apparent single LTR	1500	10,000 ± 1000	15 ± 1.5 Mbp	
Kangourou	Full element	9200	9500 ± 1000	87 ± 9 Mbp	90 ± 9 Mbp
	Apparent single LTR	3500	1000 ± 100	3.5 ± 0.5 Mbp	
Wallabi	Full element	9000	27,000 ± 3000	240 ± 24 Mbp	250 ± 25 Mbp
	Apparent single LTR	500	12,000 ± 1000	6 ± 0.5 Mbp	
					605 ± 60 Mbp

The number of copies is estimated based on dot-blot hybridizations. Mean and standard deviation (based on eight repetitions, see Supplemental data #1) are given for each element (either for the full element or for the apparent single LTR).

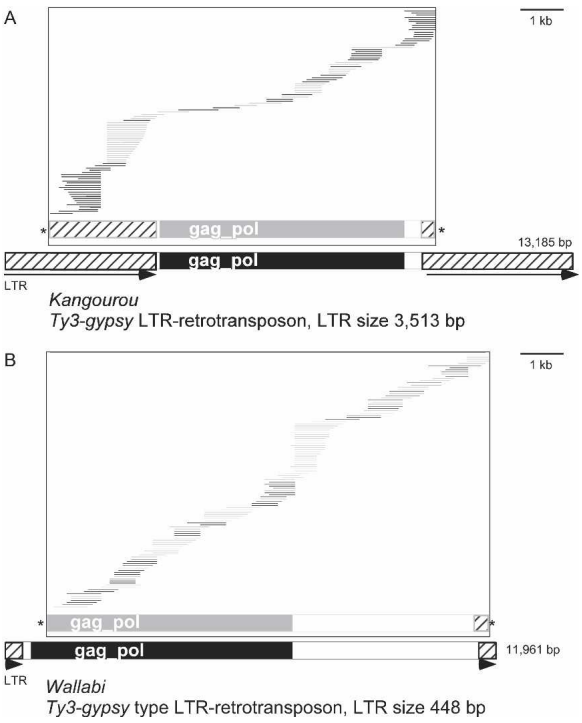

Figure 2. In silico reconstruction of *Kangourou* (*A*) and *Wallabi* (*B*) retrotransposons. The copies of the BES contigs are shown as horizontal lines (alternate black and gray according to their final position on the element). The schematic representation of the element assembled from *O. australiensis* BES is represented in gray. The schematic representation of the elements, as it is found in the *O. australiensis* sequenced BAC clones, is given in black at the *bottom* of the figures. The size scale in bp is given at the *bottom*.

repeated, but its contribution to the genomic expansion of *O. australiensis* (i.e., <5% of the present size) was considered negligible compared with the other three elements. *Dingo* was therefore not included in further analyses. Altogether, *RIRE1*, *Kangourou*, and *Wallabi* contribute ~60% (605 ± 40 Mbp) of the *O. australiensis* genome (Table 1). In contrast, BLAST searches of the genomic sequence of *O. sativa* (*cv*. Nipponbare) revealed that it contains only two, 10, and 16 complete copies of *RIRE1*, *Kangourou*, and *Wallabi* elements, respectively. The retrotransposition bursts of these three elements alone could thus account for the increase in the genome size of *O. australiensis*, compared with that of *O. sativa*. Our data therefore show that, at least in the case of the genus *Oryza*, retrotransposition has contributed to genome size variation to an extent which is comparable with that of polyploidization.

We first anticipated that BES may not be a random representation of the genome of *O. australiensis*, mainly because the BAC library was constructed using the HindIII restriction enzyme, but our results show a posteriori that this approach was nevertheless successful for retrieving the most highly repeated elements from the genome regardless of the representational bias that may have been caused by the restriction enzyme. In fact, it is clear from Figure 2 that some regions of the retrotransposon are overrepresented in the BES database, probably because most of the paralogs harbor a HindIII site at this location. Contrastingly, it is expected that BES corresponding to the regions of the element where the majority of the paralogs do not harbor a cleavage site should be far less frequent. However, in our case, we were able to retrieve at least one BES in all the regions of the three elements, probably because of the high copy number of each family, thus allowing their in silico reconstruction.

Deep annotation of the sequences of the three BAC clones revealed the presence of many TEs which are distinct from *RIRE1*, *Kangourou*, and *Wallabi* (Fig. 3). In all, 61 putative transposable elements were identified, accounting for 195,277 bp of sequence and representing 55.6% of the complete BAC sequences. Among these, a majority belong to the Class I retrotransposon group (accounting for 65% of all the TE and 42.9% of the BAC sequences). The 39 LTR-retrotransposons identified in these sequences belong to 11 distinct families. The LTR-retrotransposons, analyzed and annotated in the three sequenced BAC clones, were found mainly clustered within intergenic regions (Fig. 3). These elements are frequently disrupted by successive insertions of other LTR-retrotransposons, leading to the formation of nested structures as often observed in larger genomes of other cereal species (SanMiguel et al. 1996; Wicker et al. 2001, 2003). Distal parts of BACs OA_59I14 and OA_AB10104J14 and the central part of BAC OA_ABa0008H03 showed a high density of clustered and nested LTR-retrotransposons, with respectively five, four, and eight elements clustered within a distance of 41, 30, and 50 kb. Altogether, *Kangourou*, *Wallabi*, and *RIRE1* account for ~29% of all the BAC sequences which is much less than their overall genome contribution (60%). However, FISH experiments previously revealed that *RIRE1* elements are more abundant in pericentromeric than in distal regions of *O. australiensis* chromosome arms (Uozu et al. 1997), a characteristic shared with other LTR-retrotransposon families in the *O. sativa* genome (Jiang et al. 2002; Vitte and Panaud 2003). Consequently, the distribution of the elements in the three BAC sequences may not reflect their actual distribution in the genome.

The genomic amplification of *O. australiensis* occurred after its speciation

In order to trace the origin of the three elements *RIRE1*, *Kangourou*, and *Wallabi*, we surveyed their presence in nine different genome types of the genus *Oryza* by Southern hybridization using probes corresponding to either the LTR or the internal region (Fig. 4). The results clearly show that all three elements are present in at least one other wild *Oryza* species, indicating an ancient origin in the genus. Moreover, the strong hybridization signals obtained for some species distantly related to *O. australiensis* (e.g., in *O. granulata* [GG] for *Wallabi*) suggest that independent transposition bursts of *RIRE1* and *Wallabi* have occurred in distinct genome types of the genus, although to a lesser extent than in the genome of *O. australiensis*. In order to tentatively characterize and date the transposition bursts of the three elements in the *O. australiensis* genome, we conducted phenetic analyses of the three elements based on the OMAP BES data of the 12 *Oryza* species (Fig. 5; Supplemental data #3): For all three elements, the paralogs found in *O. australiensis* form a cluster which is distinct from those found in other *Oryza* species (i.e., supported by a bootstrap value which is >70%), suggesting that the retrotransposition bursts occurred concomitantly or after the speciation of *O. australiensis*.

In order to test this hypothesis, the dates of retrotransposition of *RIRE1*, *Kangourou*, and *Wallabi* were estimated in the *O. australiensis* genome (Fig. 6). We applied an approach of genomic paleontology, which consists of translating the nucleotide divergence observed between the paralogs mined out from the BES into a radiation date. This approach relies on the estimation of

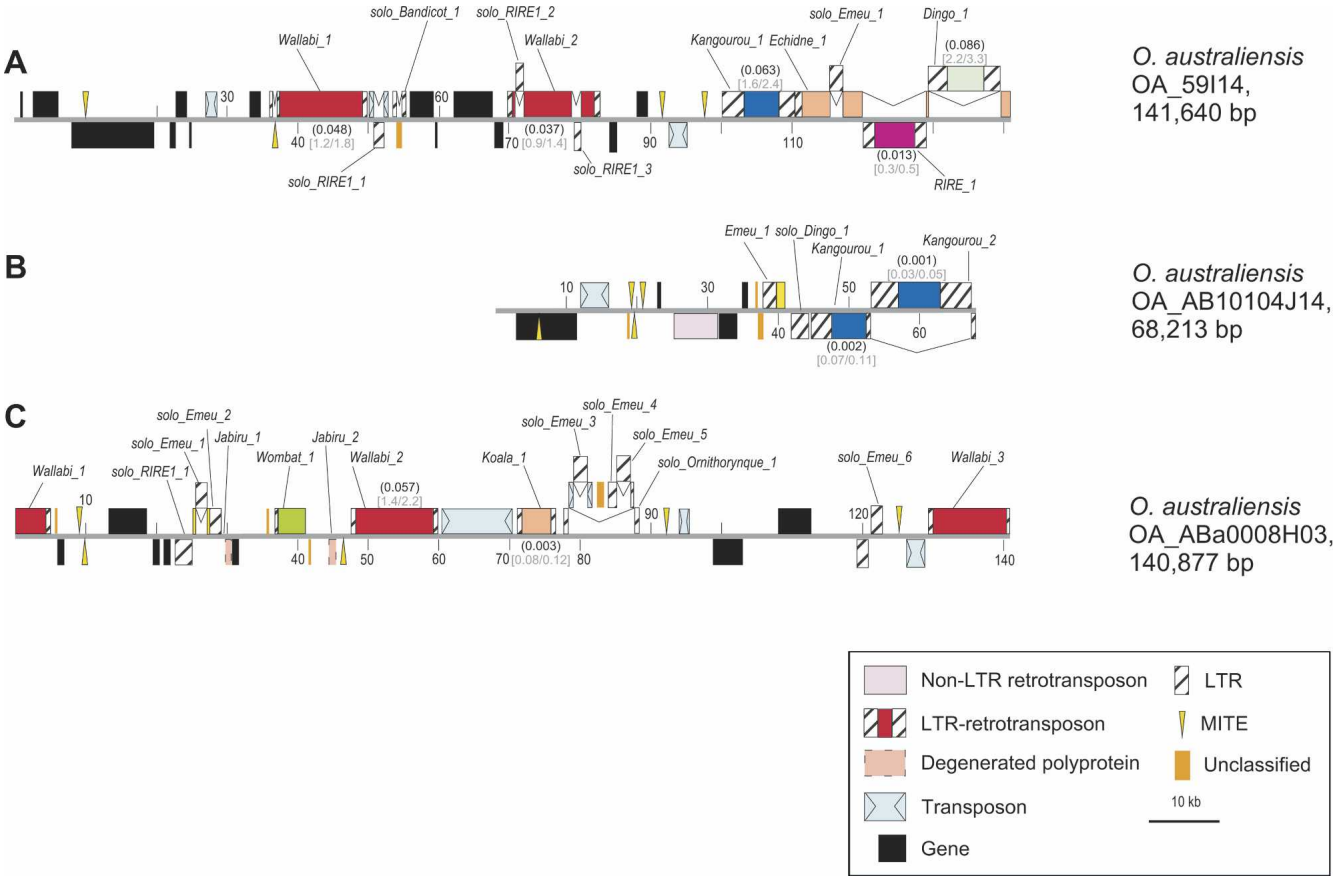

Figure 3. Physical map of three sequenced *O. australiensis* BAC clones. Black boxes represent predicted coding regions. Colored boxes represent different types of TEs as indicated on the figure. Numbers in parentheses indicate the estimated date of LTR-retrotransposon insertions (in million years) using the two molecular clocks MC1 and MC2 (see text). *A,B,C* correspond to the sequence of the BAC clones OA_59114, OA_AB10104J14, and OA_ABa0008H03, respectively.

the rate of the molecular clock (MC) of retrotransposon sequences once they are inserted in the genome. The first examples of such studies in plants were conducted using an MC of 6.5×10^{-9} synonymous substitutions/site/year (SanMiguel et al. 1998), an estimation based on the MC of the *ADH2* gene in the *Poaceae* family (Gaut et al. 1996). Several subsequent studies have led to a re-estimation of the MC of retrotransposons in rice, i.e., 2×10^{-8} subst/site/year (Vitte et al. 2004), referred to as MC1, and 1.3×10^{-8} subst/site/year (Ma et al. 2004), referred to as MC2. The data provided in the present paper are given using both these new MC. In any case, the translated dates can only be considered as rough estimates and only large differences should be retained as putatively significant. The figure clearly shows that the transpositional activity of the three elements has not been continuous during the last 3 to 4 Myr: A peak of activity (defined here as a burst) is indeed observed at ~0.5–0.75, 1.2–1.8, and 2–3 Mya for *RIRE1*, *Wallabi*, and *Kangourou*, respectively. Moreover, the size of the peaks shown in Figure 6 is proportional to the

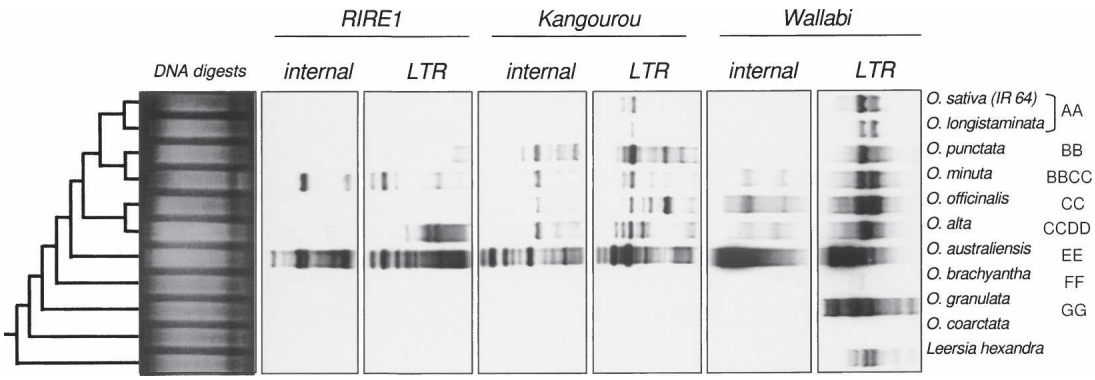

Figure 4. Southern hybridization of the three retrotransposons, *RIRE1*, *Kangourou*, and *Wallabi* on total genomic DNA of *Oryza* species digested with RsaI. The phylogenetic tree given on the figure is extrapolated from Ge et al. (1999). The direction of migration is from *left* to *right*.

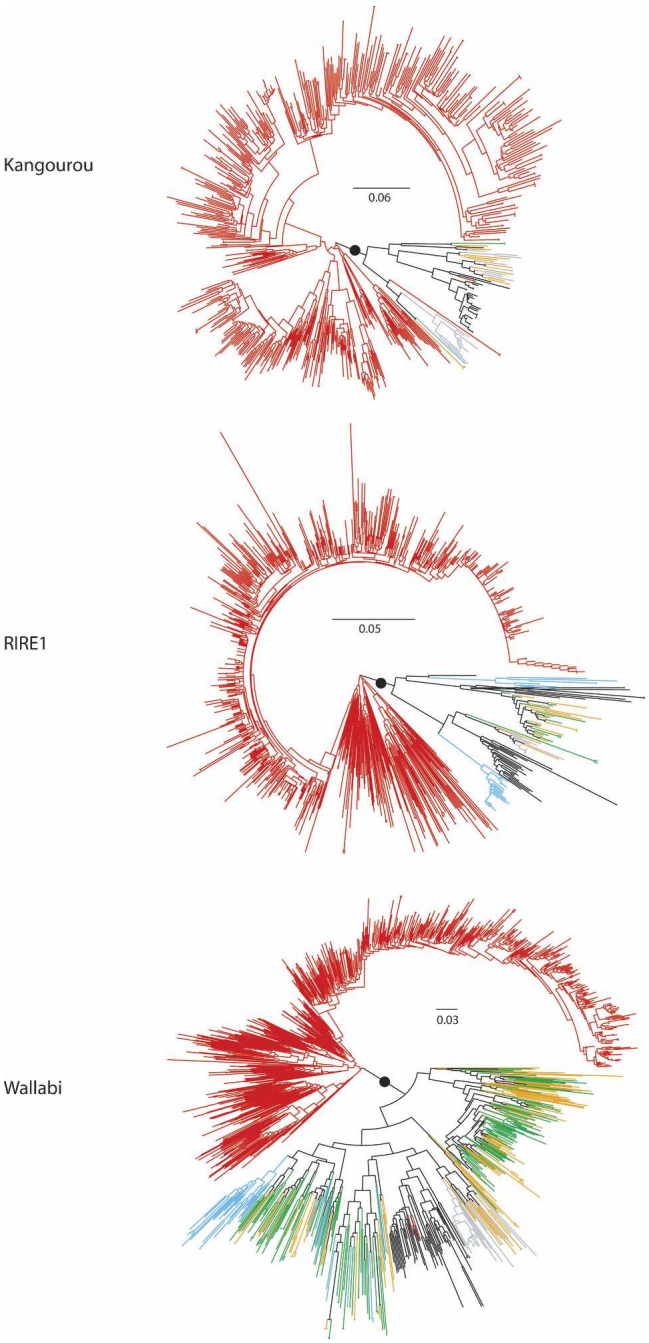

Figure 5. Phenetic relationships of *RIRE1*, *Kangourou*, and *Wallabi* in the genus *Oryza*: The neighbor-joining tree was constructed based on the alignments given in Supplemental data #3. For each tree, the dot shows the branch separating the *O. australiensis* sequences from the others. The number given near the dot corresponds to the bootstrap value. Color coding: black for the A-genome species; gray for *O. punctata*; orange for *O. minuta*; green for *O. officinalis*; blue for *O. alta*, and red for *O. australiensis*. The numbers of aligned sequences used to build the tree were as follows: for *RIRE1*: 752 *O. australiensis* sequences and 113 other *Oryza* sequences; for *Kangourou*: 570 *O. australiensis* sequences and 67 others; for *Wallabi*: 757 *O. australiensis* sequences and 422 others.

number of complete copies of the corresponding elements in the *O. australiensis* genome. This representation clearly shows that the largest bursts are the most recent and, therefore, that most of the genomic expansions that led to the doubling of the genome size of *O. australiensis* are of recent origin (i.e., within the last 3 or 4 Myr). This is further supported by dating of the insertions of *RIRE1*, *Kangourou*, and *Wallabi* elements found in the three BAC sequences (Fig. 3). In the case of full-length LTR-retrotransposons, the date of insertion can be estimated based on the divergence between their two LTRs (SanMiguel et al. 1998; Vitte et al. 2004). The estimated insertion time of these elements ranges from 1.6 to 2.4 Mya for *Kangourou_1* (OA_59I14) to 0.03–0.05 Mya for *Kangourou_2* (OA_AB10104J14). Because the date of the radiation of *O. australiensis* species is estimated at 8.5 Myr (Fig. 1), we conclude from all these lines of evidence that the genomic expansion is posterior to and not concomitant with the speciation. These results also suggest that the strong hybridization signals observed in some other *Oryza* species in the Southern hybridization experiments (e.g., for *Wallabi* in *O. granulata*, Fig. 4) reflect distinct bursts of the corresponding elements in these lineages. In this regard, the large size of the *O. granulata* genome (i.e., 880 Mbp) (Ammiraju et al. 2006), compared with that of *O. sativa*, may be partly accounted for by the retrotranspositional activity of *Wallabi*, although more detailed analyses are needed to quantify precisely this contribution.

The phenetic analyses also provide interesting insights into the dynamics of the genomic expansions observed in *O. australiensis*. The peaks shown in Figure 6 are not overlapping, thus suggesting that the maximum transpositional activity of the three elements did not occur concomitantly. The cause of these successive waves of retrotransposition could be regulatory, i.e., the result of an activation (triggered by external stimuli, such as biotic or abiotic stress) and/or by the repression of silencing of

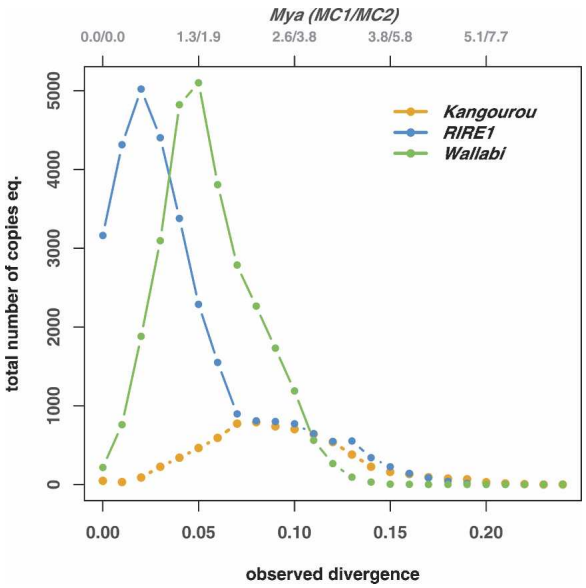

Figure 6. Timing of the bursts of the three retrotransposons, *RIRE1*, *Kangourou*, and *Wallabi*: For each element, the curves represent the distribution of the observed divergence between each paralog (given at the *bottom x*-axis). *Top x*-axis represents the date of divergence in Mya translated from the observed divergence, using the two molecular clocks MC1 and MC2 (see Methods section). The groups of paralogs used to compute the pairwise distances are defined within the phenetic subgroups shown in the phenogrammes given in Supplemental data #4. The *y*-axis represents the total number of copy equivalent, i.e., (the frequency at which the divergence time occurred) × (the number of paralogs in the genome of *O. australiensis*, based on the dot-blot experiments, Table 1).

the corresponding elements. Alternatively, these distinct bursts may be explained by the presence of active elements in the genome only during a short period (corresponding to the peaks). These active elements may have arisen from ectopic recombinations between two defective copies, a mechanism known for retroviruses (Bartosch et al. 2004). The presence of *RIRE1*, *Kangourou*, and *Wallabi* in the genome of many other *Oryza* species (Fig. 4) suggests, however, that functional copies of these three elements have probably been present in the genus since its origin. In order to test whether active copies may still be present in the genome of *O. australiensis*, we determined for each paralog the shortest distance found among all the pairwise distances computed with all the other copies (at the nucleotide level). The distributions of these shortest distances are given in Figure 7. Interestingly, for *RIRE1*, *Kangourou*, and *Wallabi*, several pairs of very closely related paralogs can be found (the first bar of the histogram), suggesting recent transposition (<200,000 yr ago) of active elements. Consequently, transpositional bursts observed in this species may have their origin in a regulatory process, rather than a structural mechanism, although this remains a hypothesis that should be further tested. As a first step one should conduct expression studies of the three elements in *O. australiensis* in order to assess whether its genome still harbors transcriptionally active copies.

Several reports on both plants and animals have shown that transposable elements are efficiently eliminated from eukaryotic genomes, either by recombination or deletion (Petrov et al. 1996; Shirasu et al. 2000; Ma et al. 2004; Chantret et al. 2005). In particular, LTR-retrotransposons tend to be partially eliminated through ectopic recombinations between their two LTRs, leading to the formation of solo-LTRs (Shirasu et al. 2000). This was taken into account in our estimation of the total contribution of the three elements to the genome size increase of *O. australiensis* by using both LTR and internal region probes in the dot-blot assay

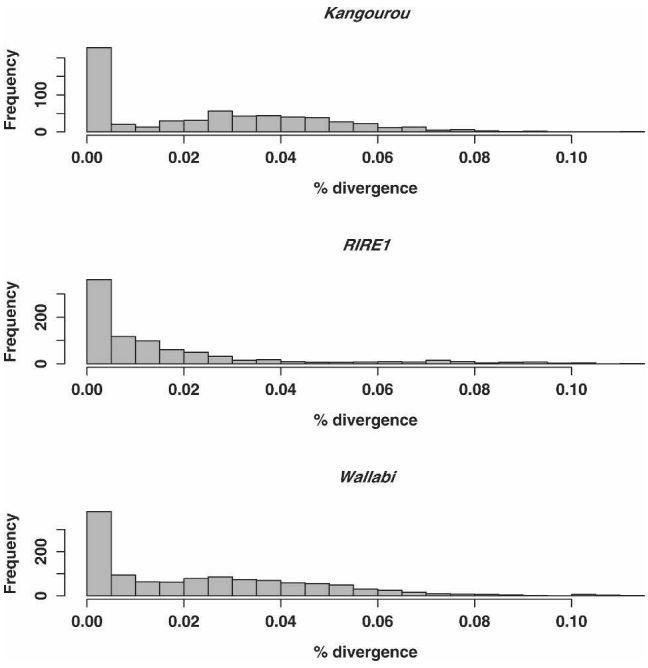

Figure 7. Histograms of the most recent among all observed divergence computed for each paralog (compared with all others) of the three retrotransposons *RIRE1*, *Kangourou*, and *Wallabi*.

(Table 1; Supplemental data #1). Interestingly, the apparent single LTRs represent a significant percentage of the total number of copies that we estimated (i.e., 25%, 9.5%, and 30% for *RIRE1*, *Kangourou*, and *Wallabi*, respectively), regardless of the age of the bursts (that of *RIRE1* being the most recent). This corroborates earlier reports suggesting that the process of partial removal of LTR-retrotransposons through ectopic recombinations leading to solo-LTRs is concomitant with (or occurs shortly after) retrotransposition (Vitte and Panaud 2003). Illegitimate recombination mechanisms targeting LTR-retrotransposons have been identified as inducing considerable loss of DNA and contributing to genome size reduction in *Arabidopsis* and rice (Petrov 2002; Ma et al. 2004). Analysis by sequence alignment of *Wallabi*, *Kangourou*, and *RIRE1* LTR-retrotransposon families from the three fully sequenced BAC clones (Fig. 3) revealed the presence of limited small deletions, corresponding to 2.4%, 6.7%, and 2.6% of the total length of the *Wallabi*, *Kangourou*, and *RIRE1* elements, respectively (data not shown). Altogether, these results indicate that processes leading to the elimination of retrotransposon sequences such as unequal and illegitimate recombinations occurred in the genome of *O. australiensis* similarly to *O. sativa*. Both the extent and timing of these DNA losses need to be investigated in order to clarify their overall contribution to the *O. australiensis* genome size variation following the bursts of the three retrotransposon families. Nevertheless, we anticipate that the overall DNA loss of the three LTR-retrotransposons through small illegitimate recombinations is negligible and did not lead to overestimation of their overall contribution in the genome of the species, based on our dot-blot assay.

The current model of eukaryotic genome evolution in relation to the activity of TEs posits that genome size should result from two balanced forces: increase, induced by retrotransposition, and decrease, caused by recombinations and deletions (Petrov 2002; Vitte and Panaud 2005). The evolutionary dynamics of *RIRE1*, *Kangourou*, and *Wallabi* in the genus *Oryza* provides an opportunity to test this hypothesis. Our Southern hybridization and phenetic data suggest that the three elements were present in the genome of the ancestor of the genus. This is further supported by the presence of homologs of these elements in BES of nearly all the *Oryza* species of the OMAP project. We also show that they have undergone independent amplification in distinct lineages in the genus, leading to one case of genomic obesity (i.e., in *O. australiensis*). In other lineages, their strong regulation and elimination may have led to a decrease in genome size (e.g., in *O. glaberrima*), but this still remains purely speculative. In this regard one should point out that, if all the copies of the three retrotransposons families *RIRE1*, *Kangourou*, and *Wallabi* were to be removed from the genome of *O. australiensis*, about 360 Mbp of genomic DNA would remain, i.e., a size comparable with that of the smallest diploid genomes of the *Oryza* genus. Further examination of complete sequences of high copy number LTR-retrotransposons in *O. australiensis* will provide better insights into the dynamics of the elimination process. The model also predicts that the successive events of TE insertions followed by their elimination should cause a fast turnover of intergenic regions, leading to their rapid divergence among distinct evolutionary lineages. This has been confirmed in several comparative studies between the genomes of maize, sorghum, and rice (Tikhonov et al. 1999; Ma et al. 2005). Comparative genomics studies within the *Oryza* genus, and in particular between closely-related species, would allow us to test this hypothesis and give insight into the dynamics of the process.

Maize, wheat, and barley are large genomes that contain 50%–80% of LTR-retrotransposons. It is now commonly accepted that retrotranspositions have played a crucial role in genomic expansion and architecture and could also have an impact on the transcriptional regulation of genes of these major crop species (Kashkush et al. 2003). However, the relatively older burst of amplification of these elements and the limited genomic sequence information from these large genomes make it difficult to reconstruct the history and study the impact of retrotransposon amplification at the whole genome level. The present study provides the first direct evidence that active LTR-retrotransposons can contribute to large variations in genome size over short periods of time, i.e., at the species level. As in the case of maize, wheat, and barley, LTR-retrotransposons are the main component of the *O. australiensis* genome. Cytologically, the mitotic chromosomes of *O. australiensis* show a twofold size increase compared with the *Oryza sativa* species (Uozu et al. 1997), indicating their dramatic impact on chromosome morphology and suggesting major events of genome reshaping. The availability of comprehensive genomic resources for many species in the genus *Oryza* makes possible physical comparison between closely related genomes contrasting for their size and will allow studies on the evolutionary history of the LTR-retrotransposons at the genus scale. This provides a unique and promising opportunity to unlock our knowledge on the causes of retrotransposition bursts as well as their impact on plant genome architecture and gene expression.

Methods

Research of highly repeated sequences

The 137,000 *O. australiensis* BES were used as query for a BLASTN search (Altschul et al. 1990) against themselves (all-by-all search). Only BES with 200 or more "complete" matches with at least 95% identity were kept. A "complete" match is a match with subject or query completely aligned. A total of 1132 BES were obtained and then used as query for a BLASTN search against known LTR-retrotransposon sequences. Six hundred and forty-nine BES matched perfectly with *RIRE1* (>95% identity). The remaining 483 BES not matching with *RIRE1* were assembled using the Sequencher software (Gene Codes Corporation) with a minimum match of 95%. A total of 25 assembled sequences (seeds) were obtained following this procedure. All these steps were automated with Perl scripts (available on request).

In silico reconstruction of the retrotransposons

The 25 highly redundant seed contigs were submitted to a BLAST search against the 137,000 BES of *O. australiensis*. Seeds and conserved overlapping BES (*E*-value < e-100) were assembled using the Sequencher software with a minimum match of 90%. This rather low threshold was especially necessary for the reconstruction of *Wallabi*, given that most of the paralogs found in the BES are more ancient than those of both *RIRE1* and *Kangourou*. The contigs identified de novo were extended by reiterative comparisons and assemblies with overlapping BES. At each step of the contig extension, the accuracy of the assembly was checked by comparative analysis with the corresponding *O. sativa* homologous LTR-retrotransposons.

Southern hybridizations

Blots were prepared using 2µg of total genomic DNA digested with RsaI transferred onto Hybond-N+ membrane. A nonradioactive procedure for probe labeling and detection signal was used (Panaud et al. 1993). Stringency washes were performed at 65°C in 0.5 × SSC. The probes were obtained by cloning the amplification products of PCR reactions on total genomic DNA of *O. australiensis* using primer pairs corresponding to the selected regions of the three elements. The accessions used for all the species were *cv.IR 64* for *O. sativa*; acc. 110,404 for *O. longistaminata*; acc. 105,690 for *O. punctata*; acc. 101,089 for *O. minuta*; acc. 101,116 for *O. officinalis*; acc. 105,143 for *O. alta*; acc. 100,882 for *O. australiensis*; acc. 101,232 for *O. brachyantha*; acc. 102,118 for *O. granulata*, and acc. 104,502 for *O. coarctata*. No accession number is available for *Leersia hexandra*. The DNA was provided by the International Rice Research Institute, Manila, Philippines.

BAC sequence analysis

BAC sequences were analyzed using BLASTN algorithms (Altschul et al. 1990) against public and local nucleotide databases. Detailed analysis was performed with the EMBOSS package (Rice et al. 2000) and by dot-plot (using DOTTER software; Sonnhammer and Durbin 1995). Putative genes were determined by a combination of coding region prediction software available through the RiceGAAS Web site (http://ricegaas.dna.affrc.go.jp) and similarity searches. Final annotation was performed with the Artemis tool (Rutherford et al. 2000). Putative TEs were both identified and annotated by similarity searches against local databases of plant TEs and by investigating structural properties of the elements. Dating of intact LTR-retrotransposon insertions was conducted according to SanMiguel et al. (1998) with the EMBOSS package using two distinct molecular clocks referred to as MC1 and MC2 (respectively 2×10^{-8} subst/site/year from Vitte et al. 2004 and 1.3×10^{-8} subst/site/year from Ma et al. 2004). Putative TEs were classified according to their mobility mechanisms.

Phenetic analyses

For each of the three elements, *RIRE1*, *Kangourou*, and *Wallabi*, the total sequence was split into subsequences of either 400 bp (for *Wallabi*) or 200 bp (for *RIRE1* and *Kangourou*). Each of these subsequences was used as query for a BLASTN search against all the OMAP BES released in GenBank. Only hits showing homology over at least 90% of the total length of the query were kept. For each element, only one subregion, for which the highest number of *Oryza* species were represented in the data set, was chosen for further analyses. All the subsequences corresponding to the match were aligned using ClustalX (Thompson et al. 1997) and the alignments were modified by hand using SEAVIEW software (Galtier et al. 1996). Final alignments were used to construct a Neighbor-Joining dendrogram using the ClustalX software, using the observed divergence distance and performing 1000 bootstraps. A circular classification tree was drawn using the Treedyn package (http://www.treedyn.org/). The age of the retrotransposition bursts was estimated at the peaks of the distribution of the pairwise nucleotide distances obtained within phenetic groups for each element (Fig. 5; Supplemental data #4). The observed divergence was translated into an insertion date following the method first described by SanMiguel et al. (1998), but using the two molecular clocks MC1 and MC2 (see above).

Phylogenetic analyses of *Oryza* species

The sequence of *ADH2* gene was retrieved from GenBank for six diploid *Oryza* species. We used the GenBank accession numbers AF148623 for *O. australiensis*, AF148632 for *O. brachyantha*, AF148606 for *O. glaberrima*, AF148613 for *O. officinalis*, AF148611 for *O. punctata*, and AF148602 for *O. sativa* (Ge et al. 1999). For each accession, the coding sequence (CDS) was extracted. These

six CDS were then aligned using ClustalX, and a pairwise distance matrix was computed using the Nei and Gojobori method (Nei and Gojobori 1986). Only synonymous substitutions were thus taken into account. A neighbor-joining tree was built using 500 bootstrap replicates. We estimated the date of the nodes using a rate of 6.5×10^{-9} synonymous substitutions/site/year (Gaut et al. 1996). The Mega3 software was used for this work (Kumar et al. 2004).

Acknowledgments

BP, RG, NP, AR, and OP were supported by CNRS funding. HK, RAW, AS, HK, KC, and SJ were supported by NSF grants IOB 0208329 and DBI 0321678. We thank R. Cooke, T. Wicker and V. Colot for their useful comments on the manuscript. The GenBank accession numbers of the LTR-retrotransposons described in this article are DQ365821 for *Kangourou*, DQ365824 for *Wallabi*, and DQ365822 for *Dingo*.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Ammiraju, J.S., Luo, M., Goicoechea, J.L., Wang, W., Kudrna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N.B., Blackmon, B., et al. 2006. The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Res.* **16:** 140–147.

Bartosch, B., Stefanidis, D., Myers, R., Weiss, R., Patience, C., and Takeuchi, Y. 2004. Evidence and consequence of porcine endogenous retrovirus recombination. *J. Virol.* **78:** 13880–13890.

Bennetzen, J.L., Ma, J., and Devos, K.M. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond.)* **95:** 127–132.

Brar, D.S. and Khush, G.S. 1997. Alien introgression in rice. *Plant Mol. Biol.* **35:** 35–47.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., et al. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* **17:** 1033–1045.

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

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. 1996. Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gen *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci.* **93:** 10274–10279.

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.* **96:** 14400–14405.

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

Jiang, N., Bao, Z., Temnykh, S., Cheng, Z., Jiang, J., Wing, R.A., McCouch, S.R., and Wessler, S.R. 2002. Dasheng: A recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics* **161:** 1293–1305.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A.H. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci.* **97:** 6603–6607.

Kashkush, K., Feldman, M., and Levy, A. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33:** 102–106.

Kumar, A. and Bennetzen, J.L. 1999. Plant retrotransposons. *Annu. Rev. Genet.* **33:** 479–532.

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.

Lisitsyn, N., Lisitsyn, N., and Wigler, M. 1993. Cloning the differences between two complex genomes. *Science* **259:** 946–951.

Ma, J., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14:** 860–869.

Ma, J., SanMiguel, P., Lai, J., Messing, J., and Bennetzen, J.L. 2005. DNA rearrangement in orthologous orp regions of the maize, rice and sorghum genomes. *Genetics* **170:** 1209–1220.

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.

Panaud, O., Magpantay, G., and McCouch, S.R. 1993. A protocol for non-radioactive DNA labelling and detection in the RFLP analysis of rice and tomato using single-copy probes. *Plant Mol. Biol. Rep.* **11:** 54–59.

Panaud, O., Vitte, C., Hivert, J., Muzlak, S., Talag, J., Brar, D.S., and Sarr, A. 2002. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Mol. Genet. Genomics* **268:** 113–121.

Petrov, D.A. 2002. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61:** 531–544.

Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276–277.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. 2000. Artemis: Sequence visualization and annotation. *Bioinformatics* **16:** 944–945.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.

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

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10:** 908–915.

Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** 1–10.

Thomas, C.A. 1971. The genetic organization of chromosomes. *Annu. Rev. Genet.* **5:** 237–256.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.D., Bennetzen, J.L., and Avramova, Z. 1999. Colinearity and its exceptions in orthologous ADH regions of maize and sorghum. *Proc. Natl. Acad. Sci.* **96:** 7409–7414.

Uozu, S., Ikehashi, H., Ohmido, N., Ohtsubo, H., Ohtsubo, E., and Fukui, K. 1997. Repetitive sequences: Cause for variation in genome size and chromosome morphology in the genus *Oryza*. *Plant Mol. Biol.* **35:** 791–799.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., and Schulman, A.H. 1999. Retrotransposon BARE-1 and its role in genome evolution in the genus hordeum. *Plant Cell* **11:** 1769–1784.

Vitte, C. and Panaud, O. 2003. Formation of Solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* **20:** 528–540.

———. 2005. LTR retrotransposons and plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **110:** 91–107.

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

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* **26:** 307–316.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J., and Keller, B. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell* **15:** 1186–1197.

Received May 9, 2006; accepted in revised form August 2, 2006.