Formation of Solo-LTRs Through Unequal Homologous Recombination Counterbalances Amplifications of LTR Retrotransposons in Rice *Oryza sativa* L.

C. Vitte and O. Panaud

Laboratoire Ecologie, Systématique et Evolution, Université Paris XI, Orsay Cedex, France

We studied the dynamics of *hopi*, *Retrosat1*, and *RIRE3*, three *gypsy*-like long terminal repeat (LTR) retrotransposons, in *Oryza sativa* L. genome. For each family, we assessed the phenetic relationships of the copies and estimated the date of insertion of the complete copies through the evaluation of their LTR divergence. We show that within each family, distinct phenetic groups have inserted at significantly different times, within the past 5 Myr and that two major amplification events may have occurred during this period. We show that solo-LTR formation through homologous unequal recombination has occurred in rice within the past 5 Myr for the three elements. We thus propose an increase/decrease model for rice genome evolution, in which both amplification and recombination processes drive variations in genome size.

Introduction

The observation that variation in genome size is not correlated with the biological complexity of higher eukaryotes, referred to as the C-value paradox (Thomas 1971), has been explained in the plant kingdom by the finding that nongenic regions, which make a large proportion of complex plant genomes, are the main source of variation in genome size (Bennetzen et al. 1998). In addition, there is now much evidence that the activity of transposable elements (TEs) is at the origin of most of the structural genomic diversity observed in angiosperms (Kidwell and Lisch 1997; Bennetzen 2000). One of the main goals in today's plant evolutionary genomics is therefore to unravel the processes through which TE activity drives structural changes in complex genomes.

The Poaceae family is a good model in which to study such processes. In this family, genomes are conserved in terms of gene content and gene order (Ahn and Tanksley 1993; Barakat, Carels, and Bernardi 1997), whereas they greatly vary in size (from 0.5 pg/2C for *Oropetium thomaeum* to 27.6 pg/2C for *Lygeum spartum*). Such variations cannot be explained solely by differences in terms of ploidy level or large duplications (Bennett 1998). In addition, several microcolinearity analyses of large contiguous genomic sequences have shown that, whereas genes and gene order are well conserved, there is no correspondence between the TEs that make most of the intergenic regions (SanMiguel et al. 1996; Bennetzen et al. 1998; Tikhonov et al. 1999).

Several recent studies have shown that long terminal repeat (LTR) retrotransposons make a large part of Poaceae genomes (for a review, see Feschotte, Jiang, and Wessler 2002). Because of their copy-and-paste transposition mechanism, active retrotransposons can potentially induce large increases in genome size. For example, in the barley genome (*Hordeum vulgare*), the *BARE-1* family represents on average 16.6×10^3 copies, which corresponds to about 3% of the nuclear genome (Vicient et al. 1999). Observations such as these have lead some authors to propose an increase-only model for the evolution of genome size in the Poaceae family (Bennetzen and Kellogg 1997), where genomes undergo large amplification events that cannot be reversed, thus increasing their size. It has also been proposed that genome size could be reduced through recombination mechanisms, that is, the formation of solo-LTRs through unequal recombination in barley (Shirasu et al. 2000), and/or the formation of deletions through illegitimate recombination in *Drosophila* (Petrov, Lozovskaya, and Hartl 1996; Petrov et al. 2000) and *Arabidopsis thaliana* (Devos, Brown, and Bennetzen 2002). These results suggest that such decreasing forces may have to be taken into account in a model of genomic evolution, leading to an increase/decrease model instead of the increase-only model proposed earlier (Bennetzen and Kellogg 1997). However, as large amplification events have been reported (SanMiguel et al. 1998), it is not yet clear whether these mechanisms could be efficient enough to reverse massive genomic increases. In order to build a model for plant genomic evolution, we thus need to determine the relative extent of these two counteracting mechanisms (retroelement amplification and LTR recombination).

The timing of both retroelement amplification and LTR recombination might also be a parameter that should be taken into account. SanMiguel et al. (1998) have shown that the maize genome has undergone successive massive amplifications, each one being relatively limited through time, corresponding to bursts of retrotransposons amplifications. As for the elimination of the numerous copies produced by such rapid and extensive bursts, it is not yet clear whether recombination occurs continuously through time, thus slowly and regularly decreasing large amounts of DNA, or if there is any mechanism that would activate large recombination events following bursts of amplification, as proposed by some authors (Rabinowicz 2000).

Rice is a Poaceae with a small genome (about 450 Mb) that contains several LTR retrotransposons (Hirochika, Fukuchi, and Kikuchi 1992; Hirochika et al. 1996; Noma et al. 1997; Kumekawa et al. 1999; Ohtsubo, Kumekawa, and Ohtsubo 1999; Tarchini et al. 2000; Kumekawa et al. 2001; Panaud et al. 2002). In addition, the availability of the genomic sequence of the Nipponbare cultivar makes

Key words: *Oryza sativa*, genome evolution, LTR retrotransposon, solo-LTR.

E-mail address: clementine.vitte@ese.u-psud.fr.

Mol. Biol. Evol. 20(4):528–540. 2003
DOI: 10.1093/molbev/msg055

the species a good model for the characterization of TEs and provides a good opportunity to test the increase/decrease model.

The use of representational difference analysis (Lisitsyn, Lisitsyn, and Wigler 1993) as a tool to study genomic differentiations allowed us to isolate 11 rice clones corresponding to seven transposable elements, among which six were LTR retrotransposons (Panaud et al. 2002). The results show that these elements might derive from recent amplification events and could explain part of the genomic differentiations between several *Oryza* species. They are thus good candidates for testing the dynamics of retroelement amplification and LTR recombination.

In this paper, we analyze in detail three *gypsy*-like LTR retrotransposons from the six above-mentioned elements. For each element, we use the rice genomic sequence available in the public database to extract complete copies and solo-LTRs. Through the analysis of both clustering and insertion time of the copies, we study the dynamics of the LTR retrotransposons amplification process, the extent of the LTR unequal homologous recombination process, and the relative timing of these two processes.

Materials and Methods

Data Mining

Three *gypsy*-like LTR retroelements, *hopi*, *Retrosat1*, and *RIRE3*, were analyzed using the 30% of *O. sativa* japonica cv. Nipponbare available in the GenBank database until November 2001. These elements were chosen because they have distinct LTR size, respectively about 1200 bp, 400 bp and 3200 bp, a parameter that, we anticipated, may influence the LTR recombination process. For each of these elements, the "reference" copy given by Panaud et al. (2002) was used as query to perform a BlastN search (http://ncbi.nlm.nih.gov/blast) on the rice genomic sequence. These reference copies are AF537364 for *hopi*, AC020666 (nt 90174–nt 78873) for *Retrosat1*, and AC022352 (nt 55264–nt 43177) for *RIRE3*. Using the output of this search, we created for each element a database with all the BAC and PAC clones that contain a region of homology with this copy. In parallel, the sequence of the 5′ LTR of the reference copy was used to perform the same procedure, in order not to bias the sample against solo-LTRs. We then determined whether each paralog corresponded to a complete copy of the retroelement or to a solo-LTR. Every paralog copy was then used to perform a BlastN2 (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) comparison with the reference copy (the one used as a query for the BlastN search) in order to determine the exact LTR boundaries and the whole structure of each paralog copy (insertions, rearrangements), and to confirm the type of copy. Retroelements were characterized on the basis of the presence of at least a part of the *gag*/*pol* region, and solo-LTRs were characterized by the absence of any adjacent internal region of the corresponding retroelement.

In addition, because this first database was constructed using the reference copy of each element as query for the BlastN search, we anticipated that the results might be biased towards the paralogs most closely related to this reference copy and therefore incomplete. We thus used the first sample to build a preliminary Neighbor-Joining dendrogram and ran additional BlastN searches using paralogs that were distantly related to the reference copy as queries. Reiteration of such BlastN searches was done until no new group was coming out. The copies with truncated ends were not included in the analysis.

The flanking regions of the copies were analyzed in order to determine their duplicated target site. When the two flanking sequences were different, we analyzed the copy sequence and flanking sequences further in order to identify eventual conversion or recombination events. This allowed us to detect copies of our database that had undergone conversion and/or recombination (which could lead to a misestimation of the timing results) and to estimate the proportion of solo-LTRs that may have formed through interelement recombination.

LTR Sequences Alignment and Phenetic Analysis

LTRs from the final sample file were aligned using the Clustal_X multiple alignment mode program (Thompson et al. 1997). Both solo-LTRs and LTRs from complete elements were included in the alignment. For the latter category, the two LTRs of each copy were represented. In order to avoid artifactual clustering due to bad alignment, each alignment was corrected by hand using the SEAVIEW software (Galtier, Gouy, and Gautier 1996). Microsatellite regions were removed, as were unstable regions such as CT-rich regions and very divergent regions that could not be properly aligned. We thus eliminated 707 bp over 1441 bp for the *hopi* LTR alignment, 172 bp over 629 bp for the *Retrosat1* LTR alignment, and 286 bp over 3284 bp for the *RIRE3* LTR alignment. In addition, when a transposable element was found inserted within a sequence, the corresponding indel was considered as a simple insertion event and replaced by an "X" in the sequence of the copy within which it was found, together with the duplicated target site. Final LTR alignments were used to construct a Neighbor-Joining dendrogram using the PHYLO_WYN software (Galtier, Gouy, and Gautier 1996), using the "observed divergence" distance and performing 500 bootstrap replicates.

Reverse-Transcriptase, Integrase, and RNaseH Sequences Alignments and Phenetic Analysis

The sequence of the *gag*/*pol* polyprotein gene of each copy was identified on a BlastX2 analysis (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) comparing the nucleotide sequence of the copy to the *gag*/*pol* polyprotein sequence of the *gypsy* retrovirus of *Drosophila* (GenBank accession number AAC82604). Final alignment of *gag*/*pol* nucleic sequences was performed using the Clustal_X multiple mode alignment program (Thompson et al. 1997). The Clustal_X profile alignment mode was then used to align the reverse-transcriptase (RT) nucleic sequence of the reference copy (described in Panaud et al. 2002) with the preceding *gag*/*pol* sequences alignment.

Table 1
Distribution of *hopi*, *Retrosat1*, and *RIRE3* Copies Extracted from the 30% Rice Genomic Sequence Available on the 12 Chromosomes of Rice

	Rice Chromosomes											
Family	1	2	3	4	5	6	7	8	9	10	11	12
hopi	53	1	4	1	3	8	1	—	1	12	—	—
Retrosat1	12	—	3	—	1	—	—	—	1	4	—	—
RIRE3	18	—	1	—	—	1	—	—	—	5	—	—

NOTE.—For each retroelement, the table gives the number of copies found on each of the 12 chromosomes of the rice genome. Dashes correspond to an absence of copy on the chromosome.

The same procedure was performed for Integrase (Int) and RNaseH domains.

For each element, final RT, Int, and RNaseH alignments were independently used to construct Neighbor-Joining dendrograms using the PHYLO_WYN software (Galtier, Gouy, and Gautier 1996), using the observed divergence distance and 500 bootstrap replicates.

Determination of Phenetic Groups Within the Three Families

For each family, we independently analyzed the topology of the four dendrograms (LTRs, RT, Int, and RNaseH). The three dendrograms obtained from the coding-domain sequences showed similar topologies, although minor differences could be observed. Groups were therefore defined by compiling the data of the three dendrograms, although only the RT dendrogram is presented in the results. The topology of the LTR-based dendrogram only differs from the other three by the fact that subgroups can be defined within a given group, resulting in a more complex topology, which may come from the fact that LTR sequences evolve more rapidly than coding sequences.

Dating of Insertion Events

In order to date insertion events of the copies from our database, we analyzed the LTR nucleotide divergence rate of the copies. This method was first used to date the insertion events of LTR retrotransposons in maize (SanMiguel et al. 1998) and subsequently extended to other species (Jordan and McDonald 1998; Promislow, Jordan, and McDonald 1999; Bowen and McDonald 2001; Jiang et al. 2002) and to human endogenous retroviruses (HERVs) (Tristem 2000).

For each complete copy, the two LTRs were aligned using the Clustal_X algorithm (Thompson et al. 1997). The alignments were checked and eventually corrected by hand using the SEAVIEW software (Galtier, Gouy, and Gautier 1996). The nucleotide divergence rate between the two LTRs was determined using the PAUP software (Swofford 1999). Note that indels and microsatellites were not taken into account to estimate these divergence rates. LTR divergence rate were converted into dates using the average substitution rate of the *Adh1* and *Adh2* loci of grasses, which has been estimated at 6.5×10^{-9} substitutions per synonymous site per year (Gaut et al. 1996). In order to estimate the timing of insertion of retroelements that have undergone recombination and became solo-LTRs, we first analyzed the insertion date of complete retroelements, and then used clustering of the solo-LTRs to specific groups in order to estimate their date of insertion.

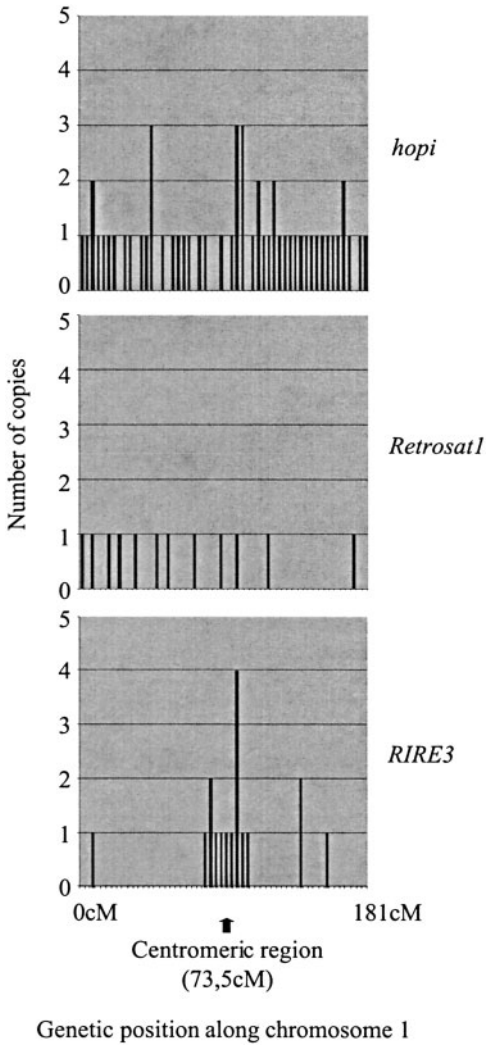


FIG. 1.—Genetic map of copies from *hopi*, *Retrosat1*, and *RIRE3* families along rice chromosome 1. Genetic positions are given in centimorgans (cM) from the top of rice chromosome 1. The arrow shows the position of the centromere at 73.5 cM from the top of the chromosome.

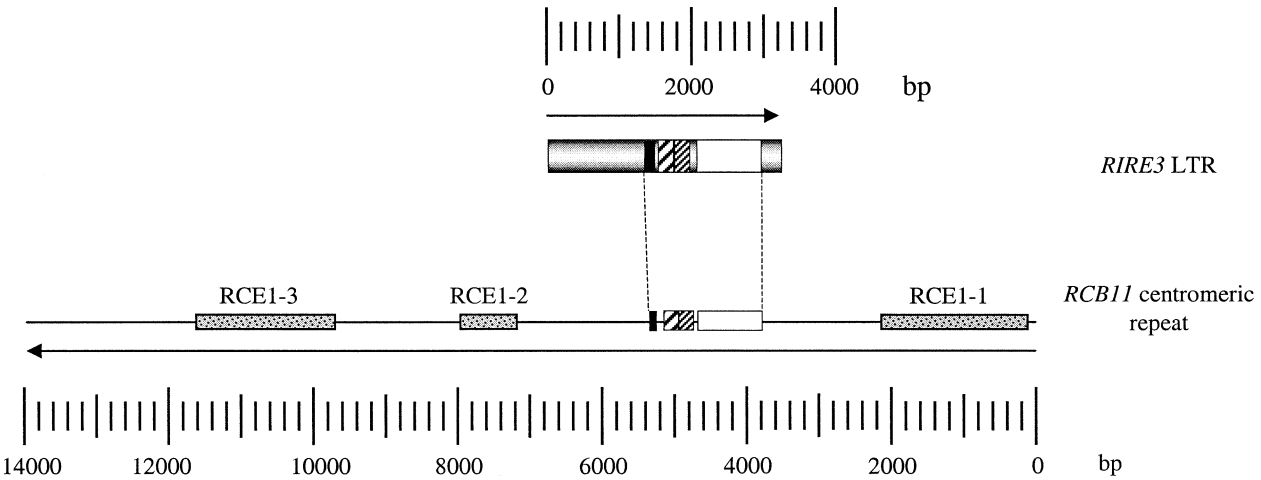


FIG. 2.—Homologous regions of the *RIRE3* LTR and the *RCB11* rice centromeric repeat. The *RIRE3* LTR sequence from the reference copy AC022352 and the *RCB11* centromeric repeat sequence (GenBank accession number AB013613) are presented at the same scale. Note that the *RIRE3* LTR is presented in the 5′ to 3′ orientation, whereas the *RCB11* sequence is presented in reverse, as indicated by the two arrows. Characteristic repeats *RCE1-1*, *RCE1-2*, and *RCE1-3* from the *RCB11* repeat are indicated as gray dotted blocks on the *RCB11* sequence. Homologous regions are represented by black, black and white large hatched, black and white thin hatched, and white blocks on both *RIRE3* LTR and *RCB11* sequences. The black region (87 bp) shows 80% identity (E-value = 0.03), the black and white large hatched region (142 bp) shows 95% identity (E-value = 3×10^{-58}), the black and white thin hatched region (182 bp) shows 85% identity (E-value = 8×10^{-44}), and the white region (814 bp) shows 90% identity (E-value = 0.0).

Age of the Phenetic Groups

For each group previously defined, mean and standard error were calculated for the LTR divergence and subsequently for the date of insertion. As conversion and recombination processes may influence the divergence rate between the two LTRs of a copy, and thus the estimation of the age of the corresponding copy, copies for which signature of conversion or recombination events were detected (*hopi* copies AP001129 and AC079021A and *RIRE3* copy AC022352C [see fig. 3 for details]) were not taken into account for these computations. For *hopi* copy AP003204 and *RIRE3* copies AC080019E and AC080019F, corresponding values were also not included, because such processes were suspected even if no traces could be detected (see *Discussion*). In order to determine whether LTR divergence rate differed significantly within a retroelement family (i.e., comparing groups of the same element), we performed a Mann-Whitney test using Statistica software (Statsoft 1997).

Distribution of the Three Elements on Rice Chromosomes

The genetic position of the BAC and PAC clones used by the rice genome sequencing consortium along the rice (*O. sativa* L. cv. Nipponbare) genetic map is publicly available (http://rgp.dna.affrc.go.jp/publicdata/physicalmap2001/YACall2001.html). This allowed us to retrieve the genetic position of each copy of the three retroelements.

Results

Characterization and Insertion Distribution of Rice *gypsy*-like Retroelements *hopi*, *Retrosat1*, and *RIRE3*

Our database contains 85 copies of the *hopi* family, corresponding to 48 retroelements and 37 solo-LTRs; 22 copies of the *Retrosat1* family, corresponding to 20 retroelements and two solo-LTRs; and 34 copies of the *RIRE3* family, corresponding to nine retroelements and 25 solo-LTRs. These copies seem to be dispersed throughout the genome of *Oryza sativa*, as we found copies of *hopi* on chromosomes 1, 2, 3, 4, 5, 6, 7, 9, and 10; copies of *Retrosat1* on chromosomes 1, 3, 5, 9 and 10; and copies of *RIRE3* on chromosomes 1, 3, 6, and 10 (table 1). The majority of the copies are however found on chromosome 1, mainly because the sequencing of this particular chromosome was far more advanced than the other 11 chromosomes at the time of the study. In addition, the availability of the almost complete sequence of chromosome 1 made possible the analysis of the distribution of the copies along this chromosome for the three families. Results, presented in figure 1, show that copies from the *RIRE3* family are mostly concentrated around the centromeric region, whereas copies from the *hopi* and *Retrosat1* families seem to be distributed evenly along the chromosome.

In addition, the BlastN searches using *RIRE3* copies as query revealed that the *RIRE3* LTR shares homology with rice *RCB11* centromeric sequence (GenBank accession number AB013613), which is composed of *RCE1* repeats. A BlastN2 analysis between *RIRE3* paralogs and clone AB013613 revealed four homologous regions (1.2 kb in total [see fig. 2]). There is however no homology between the *RIRE3* LTR and either of the *RCE1-1*, *1-2*, or *1-3* repeats that are contained in *RCB11*.

For most of the 141 copies, a duplicated target site of 5 bp was detected, but no particular consensus insertion site could be found (data not shown). For seven copies, the duplicated site showed one substitution or one 1-bp deletion and was thus still recognizable. However, for seven other copies, the flanking sequences were different, and no duplicated target site could be established. A

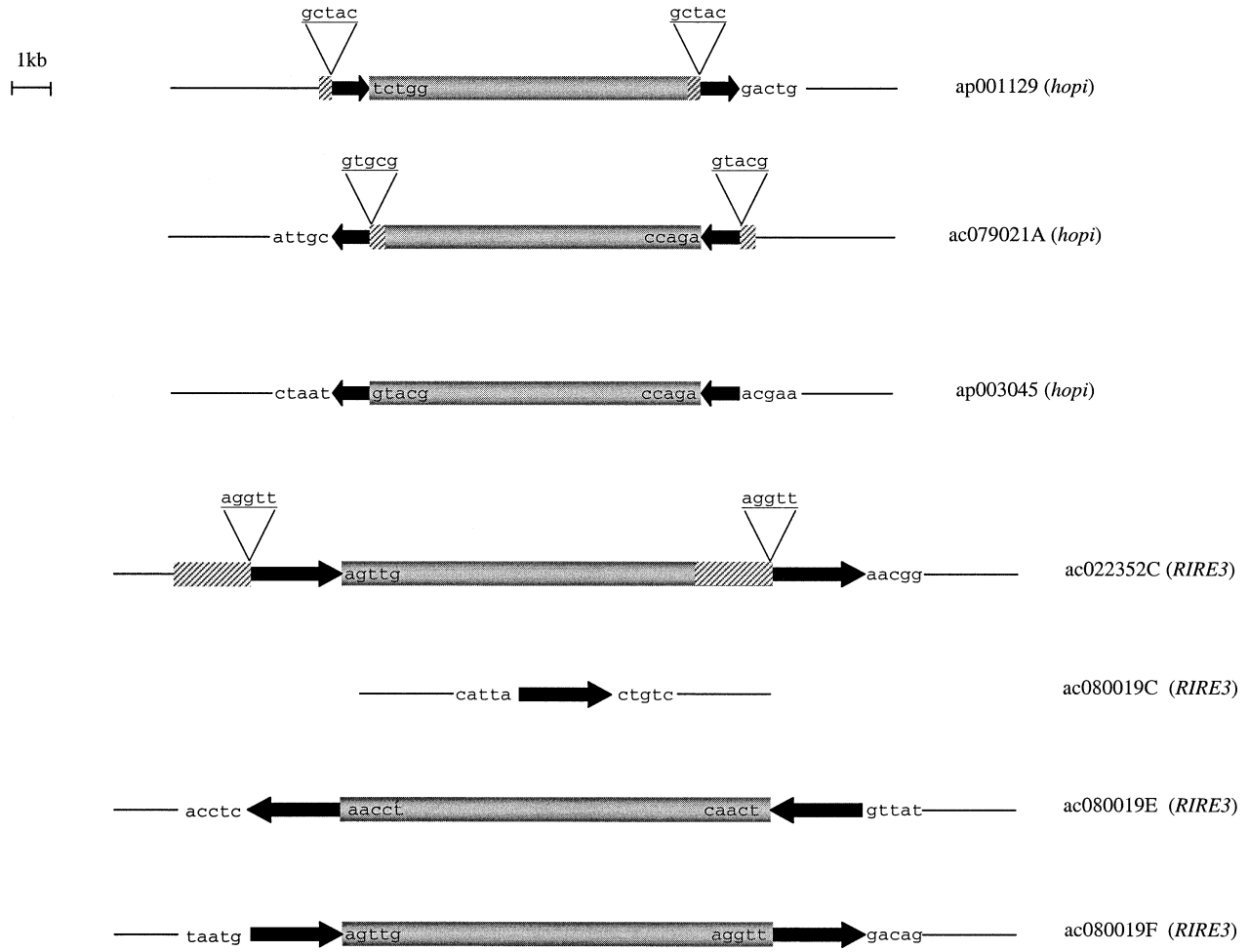


FIG. 3.—Detailed analysis of the copies showing no duplicated target site. All sequences are represented in the same scale. LTRs are represented by black arrows that indicate the orientation of the element comparatively to the BAC clone. The gray boxes correspond to internal regions of the copies. The 5 bp sequences flanking LTRs are represented in order to show the lack of identity sequence between the two fragments that flank the copy and eventually to show the identity of sequence between an internal fragment flanking one LTR and one of the flanking regions of the copy. In such cases, the two identical sequences are underlined. The hatched gray boxes correspond to a region of homology that could be enlarged from the 5 bp sequences. This is the case for copies from BAC clone AP001129, AC079021A, and A022352 but not for copies from BAC clone AP003045 and AC090018 (C, E, and F).

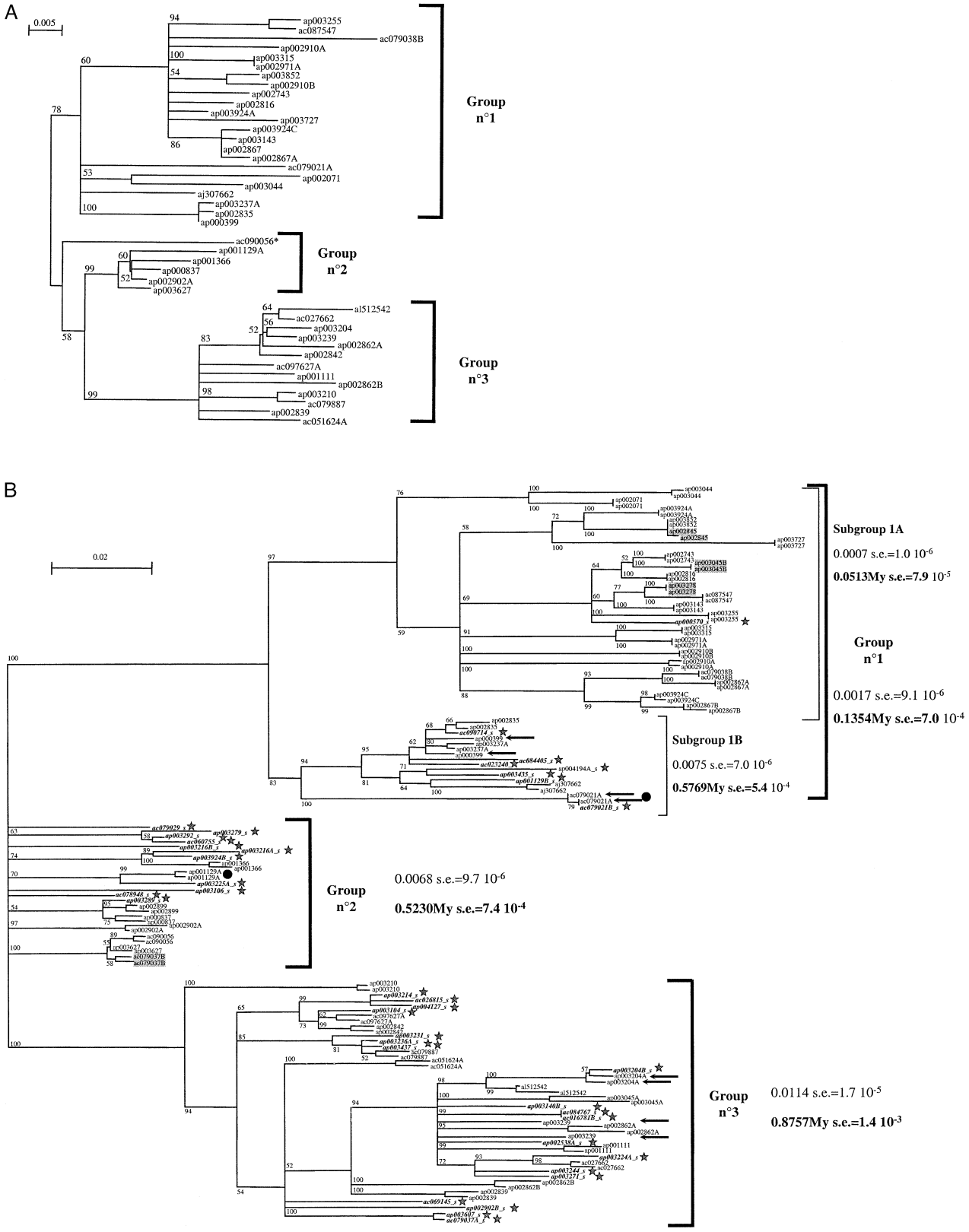
further analysis of these copies revealed traces of intraelement conversion for three of them, whereas for the last four, no particular signs could be revealed (fig. 3). The regions involved in conversion are at least 210 bp, 762 bp, and 2488 bp long, respectively (their exact size cannot be determined because of the sequence identity between the two LTRs of the copy). As conversion could cause a misestimating of the timing data, these copies were removed from the data set for further analysis.

Families of *hopi*, *Retrosat1*, and *RIRE3* Are Structured in Distinct Phenetic Groups That Are Significantly Different in Terms of LTR Nucleotide Divergence

Figures 4*A*, 5*A*, and 6*A* show the three Neighbor-Joining trees that were constructed using RT domain alignments of *hopi, Retrosat1*, and *RIRE3*. Figures 4*B*, 5*B*, and 6*B* show the three Neighbor-Joining trees that were built using LTR alignments of *hopi*, *Retrosat1*, and *RIRE3,*

→

FIG. 4.—Neighbor-Joining trees obtained with the RT sequences data (*A*) and LTR sequences data (*B*). Distance scales of the Neighbor-Joining trees are presented on the left high corner and correspond to a percentage of nucleotide substitution. Bootstrap values superior to 50 are given on the trees. Nodes that showed bootstrap values below 50 were treated as unresolved nodes. Large lines indicate the boundaries of the groups. Accessions with a thick asterisk indicate that the grouping has been determined considering the Int and RNaseH domain sequences trees. Solo-LTRs are written in bold thick characters and represented by stars. LTRs in gray boxes correspond to elements where neither RT, Int, nor RNaseH was completely sequenced and for which the group was inferred from the LTR tree. Arrows indicate couples of LTRs from the same copy that shows clustering discrepancy (i.e., for which the two LTRs do not cluster together), which may be due to intercopy conversion or recombination or to divergence. Copies where an intraelement genic conversion event was found are represented with a black disk. For each group, mean and variance of the LTR divergence parameter are given under the group number; the corresponding estimated insertion time is written in bold.


A
0.005
Group n°1
Group n°2
Group n°3
B
0.02
Subgroup 1A
0.0007 s.e.=1.0 10-6
0.0513My s.e.=7.9 10-5
Group n°1
0.0017 s.e.=9.1 10-6
0.1354My s.e.=7.0 10-4
Subgroup 1B
0.0075 s.e.=7.0 10-6
0.5769My s.e.=5.4 10-4
Group n°2
0.0068 s.e.=9.7 10-6
0.5230My s.e.=7.4 10-4
Group n°3
0.0114 s.e.=1.7 10-5
0.8757My s.e.=1.4 10-3

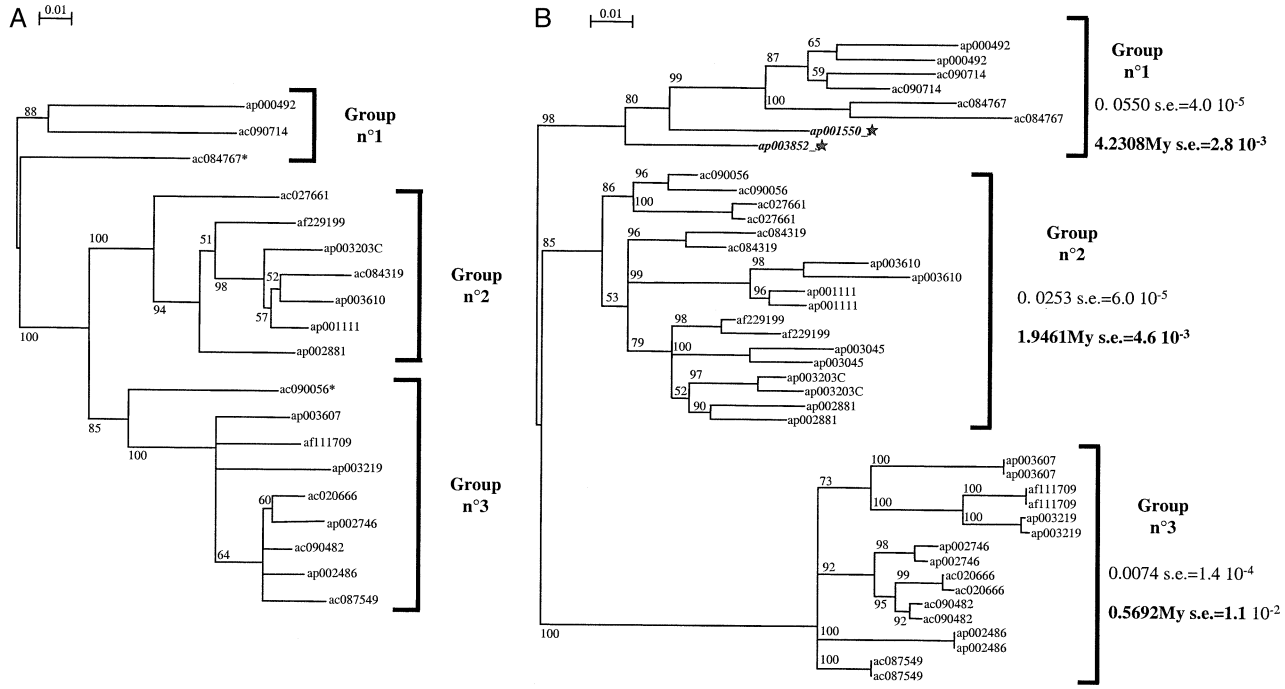


FIG. 5.—Neighbor-Joining trees obtained with the RT sequences data (*A*) and LTR sequences data (*B*). Same legend as figure 4, except that at least one of the three domain sequences has been analyzed for each element, and no sequence showed phylogenetic discrepancies, recombination, or conversion events.

respectively. Mean and variance of LTR divergence rate were calculated for each group and for subgroups 1A and 1B of *hopi* (see figs. 4*B*, 5*B*, and 6*B*). Results of the Mann-Withney test, presented in table 2, show a significant difference between every pair of each group ($P < 0.05$), except for subgroup 1B and group 2 of the *hopi* family and for the two groups of the *RIRE3* family ($P > 0.05$). In this last case, the lack of significance of the test may be due to the small sample size.

Dating of the Insertions of Complete Copies

Out of a total of 71 complete copies, we found 18 copies with identical LTRs and 53 copies with low levels of nucleotide divergence, indicating that *hopi*, *Retrosat1*, and *RIRE3* families have amplified very recently. Figures 4*B*, 5*B*, and 6*B* show the mean insertion time of each of the subgroups.

Discussion

Repartition of the Three Elements in Rice Genome

Copies of *hopi*, *Retrosat1*, and *RIRE3* seem to be present on the 12 chromosomes of rice. The distribution of the three elements along chromosome 1 is variable: *hopi* and *Retrosat1* seem to be dispersed along the chromosome, whereas it appears that most *RIRE3* copies are clustered around the centromere. Even if the clustering of the copies on the genetic map may not be representative of their physical clustering, the difference observed between the three families may be relevant to the predominance of the *RIRE3* element in the centromere region comparatively to the two other elements. In addition, the *RIRE3* LTR is homologous to the *RCB11* centromeric sequence over about 1.2 kb. Both these results suggest that there might be some structural relationships between centromeric regions and the LTR of *RIRE3*. Similarly, the *RIRE7* family shares homology with several centromeric repeats and is located around rice centromeres (Kumekawa et al. 2001), and the *RCS1* repeat family of the rice centromeric region shares homology with *gypsy*-like retroelements from maize (GenBank accession number AF030633) and *Lilium henryi* (GenBank accession number X13886) over 95 bp and 57 bp, respectively (Dong et al. 1998).

Timing of Insertion Events and the History of Rice

In order to convert LTR nucleotide divergence into dates of insertion events, a substitution rate is needed for each retroelement. However, as copies have inserted at different time and different genomic locations, a global rate is difficult to estimate, and such data were not available for these three retrotransposon families. In addition, no synonymous substitution rate is known for any rice sequence. Hence, in order to estimate the insertion time of each copy from our database, we used the average substitution rate of the *Adh1* and *Adh2* loci of grasses, which has been estimated to be 6.5×10^{-9} substitutions per synonymous site per year (Gaut et al. 1996) and had been used previously to date LTR retroelements insertions in maize (SanMiguel et al. 1998). There is a concern that our timing results might be misestimated because we use this rate. LTRs show both very well conserved regions, which might be involved in the replication cycle and thus under selective pressure, and very dynamic regions, which might not be under selective constraint. Hence, retrotransposon LTRs sequences and genic synonymous sites

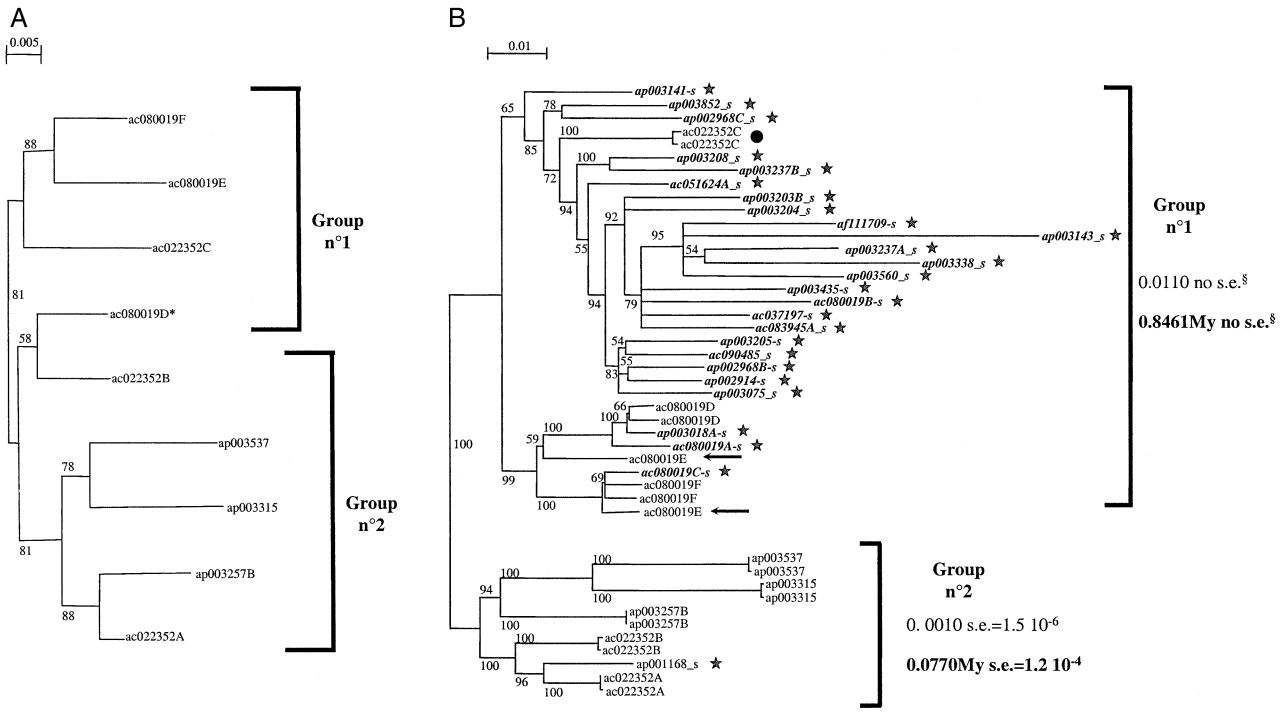


FIG. 6.—Neighbor-Joining trees obtained with the RT sequences data (*A*) and LTR sequences data (*B*). Same legend as figure 4, except that at least one of the three domain sequences has been analyzed for each element. The symbol § indicates that for group n°1, the value given corresponds to the value of the AC080019D copy and does not correspond to a mean, so no standard error could be calculated.

may not evolve identically, and the use of this rate gives a very rough estimate of the insertion time of the copies that has to be reinforced by other data.

The genetic relationships within *Oryza* genus have been well characterized by several authors over the past decades (Wang, Second, and Tanksley 1992; Ge et al. 1999; Bautista et al. 2001). The two cultivated species *O. sativa* (Asia) and *O. glaberrima* (Africa) and their closest wild relatives (*O. rufipogon* and *O. breviligulata*, respectively) have been classified as AA-genome species, based on the chromosomal behavior of their hybrids (i.e., showing a normal pairing of the chromosomes during meiosis; Katayama 1967, 1982). Figure 7 shows their genetic relationships and history: the age of the radiation of the African gene pool from the Asian gene pool is estimated at 2 to 3 Myr (Second 1985). We thus examined the dates of insertion that we found for *hopi*, *Retrosat1*, and *RIRE3* families in the view of both these paleontological data (fig. 7) and the Southern hybridization–based data on the dynamics of these elements in *Oryza* genus that we recently published (Panaud et al. 2002). The three phenetic groups of *hopi* have amplified within the past Myr, that is, after the radiation of the African gene pool (see fig. 7). *Retrosat1* has amplified mainly within the last 2 Myr, although three copies appear to have inserted around 4 Myr, that is, before the radiation of the African species. These results are thus in accordance with the result obtained by Panaud et al. (2002), as no hybridization signal was obtained when *hopi* was probed on *O. glaberrima* genomic DNA and *O. glaberrima* showed a fainter hybridization signal than the one obtained with *O. sativa* when hybridized with the *Retrosat1* probe. All the copies of *RIRE3* that we have analyzed have inserted within the past Myr. This result is incongruent with the results of Panaud et al. (2002) who clearly showed that *RIRE3* present a hybridization signal with *O. glaberrima* genomic DNA, although fainter than in the case of *O. sativa*. However, among the three elements studied here, *RIRE3* presents the highest intraelement recombination rate, particularly for group1 (composed of 25 solo-LTRs and only four complete elements, three of them harboring traces of conversion). Our dating data are therefore based on a much smaller sample than in the case of the other two elements. We thus need to extend this study to more copies of *RIRE3* to reinforce our dating estimation.

Globally, our data seem to be congruent with the Southern hybridization results of Panaud et al. (2002). This suggests that the use of the average synonymous substitution rate of *Adh1* and *Adh2* loci may be appropriate for the timing estimation of insertions of *hopi*, *Retrosat1*, and *RIRE3*. Another source of error in dating insertion events is the possible occurrence of intracopy and intercopy genic conversion and intercopy recombination. We found only few cases were conversion was clear, and we removed the corresponding copies.

Nature of the Amplification Process in Rice

The analysis of the estimated average group insertion times shows that all groups have inserted during the past 5 Myr, but most inserted within the past 1 Myr (figs. 4*B*, 5*B*, and 6*B*). One might however argue that since the method

Table 2
Results of the Mann-Whitney Test

	Hopi				*Retrosat1*				*RIRE3*		
	Subgroup 1A	Subroup 1B	Group 2	Group 3		Group 1	Group 2	Group 3		Group 1	Group 2
Subgroup 1A	—	***	***	***	Group 1	—	**	*	Group 1	—	NS
Subgroup 1B		—	NS	*	Group 2		—	***	Group 2		—
Group 2			—	***	Group 3			—			
Group 3				—							

NOTE.—* corresponds to $P < 0.05$, ** corresponds to $P < 0.01$, and *** corresponds to $P < 0.001$. Dashes correspond to an absence of copy on the chromosome. NS indicates nonsignificant.

used to retrieve retroelement sequences is based on a BlastN search, which depends on threshold parameters, sequences that have undergone extensive rearrangements or that are too divergent from the query sequence could not be retrieved. Hence, if one supposes that old sequences may have undergone more alterations than newer ones, and under a constant evolution rate model, our database may be biased towards recent copies. In addition, we have only considered sequences that could be assigned as retroelements or solo-LTRs, but did not take into account the ones that showed only a partial region of homology with the reference copy and that may correspond to remnant retroelements, which further increases the bias towards recent copies. We therefore consider our analysis as a study of the recent history of the rice genome.

Even in the case of such recent amplification events, if the amplification process has occurred continuously, we should observe for each family a continuous decrease of copy number with time. For the *hopi* family, the global distribution (fig. 8) shows at least two peaks, which suggests that gain of retroelements sequences in the rice genome did not act continuously through time, but rather by distinct amplification events, as it has been shown in maize (SanMiguel et al. 1998), even if they seem to be of less important extent. In the case of rice, it is yet difficult to clearly assess whether these amplification events can be

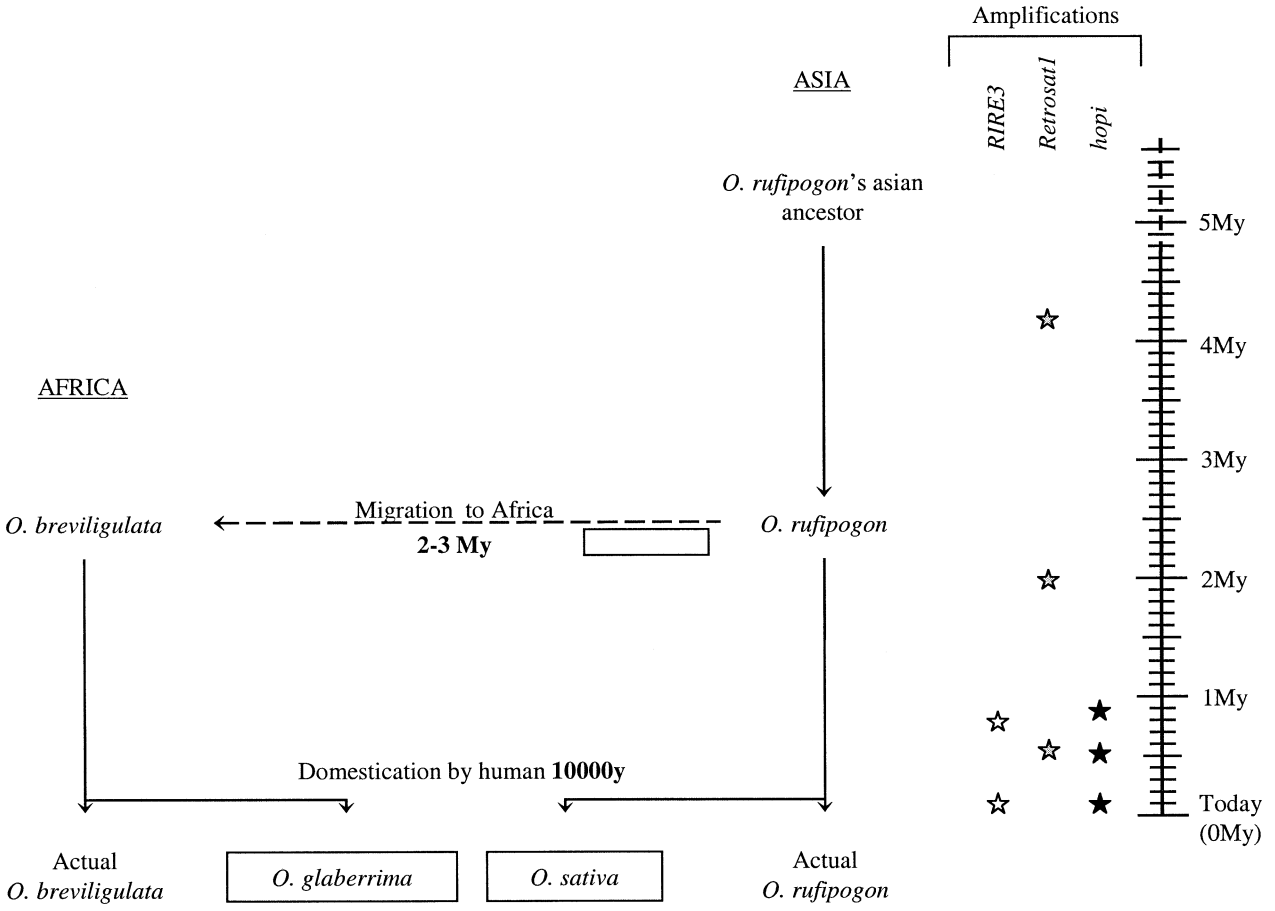


FIG. 7.—Comparison of dates of insertion to the paleontological data of rice. The two cultivated species are shown in boxes, and the locations of the species are written underlined capital letters. Migration from Asia to Africa is represented by a hatched arrow, whereas lineage relationships are shown by continuous arrows. The corresponding time scale is represented on the right in million years (Myr). Amplification events of the three retroelement families are represented by stars.

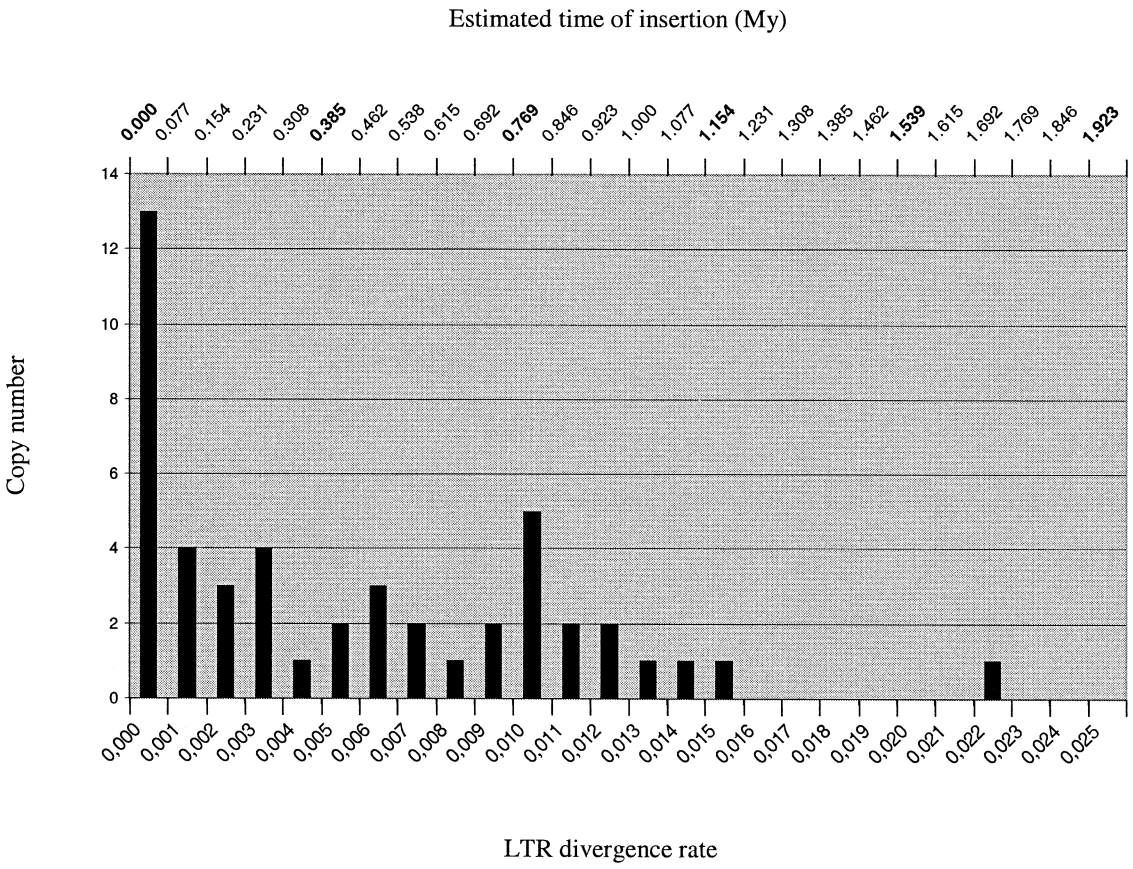


FIG. 8.—Distribution of the insertion events of copies from the *hopi* family. The LTR divergence is given as an estimate of the insertion time of the copies. Corresponding times are shown in million years (Myr).

considered as bursts because our sampling was based on 30% of the total rice genomic sequence.

Insertion Events Are Phenetically Structured

Groups have been defined within each retroelement family on the basis of a consensus topology of the RT, Int, and RNaseH trees. Thus, each group might reflect the transpositional history of one parental copy. For the three families of element analyzed (except for group 1 of *RIRE3*), intragroup variances are very low, suggesting that most of the copies belonging to one group have inserted within the same time period. The slight differences observed between copies of the same group might be due to differences in terms of evolution rate, considering that copies have inserted in distinct genomic environments. But we cannot rule out that these differences effectively correspond to distinct insertion times. In particular, in the *hopi* family, group 1 variance is in majority due to five copies that show the highest divergence rate of the group and which cluster together on the LTRs tree. This suggests that group 1 may have amplified twice, and we thus subdivided this group into two subgroups, 1A and 1B.

Results of the Mann-Whitney test for the *hopi* family suggest that groups 1B and 2 may have amplified concomitantly, whereas they have amplified at independent times compared with subgroup 1A and group 3, which also have amplified independently one from the other.

These results suggest that for a given retroelement family, several master copies may have amplified at different times, each one leading to concomitant insertions of phenetically close copies (corresponding to the groups observed) at distinct periods of time (corresponding to the mean LTR divergence observed for each group), although with rare exceptions (subgroup 1B).

Extent of the LTR Recombination Process

For the three retroelement families studied in this paper, we found solo-LTRs. This shows that unequal homologous recombination between two LTRs does occur in rice genome, as was shown in barley (Shirasu et al. 2000) and *Arabidopsis* (Devos, Brown, and Bennetzen 2002), and thus reinforces the existence of a genome-decreasing force driven by solo-LTRs formation. Whereas solo-LTRs appear to be rare in maize (SanMiguel et al. 1996; Devos, Brown, and Bennetzen 2002), it has been shown that *Arabidopsis* has an approximately 1:1 solo-LTR to intact elements ratio (Devos, Brown, and Bennetzen 2002) and that the *BARE-1* retroelement of the barley genome shows 16-fold more LTRs than internal domains. The excess is mainly due to solo-LTRs (Vicient et al. 1999; Shirasu et al. 2000). In rice, Vicient and Schulman (2002) showed an approximately 1.6:1 ratio of global solo-LTRs to complete copies for *copia*-like elements and an approximately 6.3:1 ratio for the *RIRE1 copia*-like family, and an approximately

0.3:1 ratio for the *RIRE2* family (corresponding to the *Retrosat1* family). Here, considering the three *gypsy*-like retroelement families, we found 114 complete copies and 75 solo-LTRs, leading to an approximately 0.7:1 ratio solo-LTRs to complete copies. Considering each family separately, we found approximately 0.6:1, approximately 0.1:1 (which is close to the 0.3:1 ratio found for *RIRE2* by Vicient and Schulman [2002]) and approximately 2.5:1 ratios, for the *hopi*, *Retrosat1*, and *RIRE3* families, respectively. Thus, the ratio between solo-LTRs and complete elements varies considerably among the *gypsy*-like retroelement families. Considering the results of Vicient and Schulman (2002), this feature also seems to be true for *copia*-like elements, as the ratio for the *RIRE1* family is distinct from the one obtained for the *copia* elements all together. These differences may be due to specific characteristics of the retroelements, such as preferential insertion regions or LTR size. Here, we show that the *RIRE3* family inserts preferentially in the centromeric regions of chromosome 1, whereas the two other families seem to be dispersed on this chromosome. This feature could have an impact on the unequal homologous recombination process and could lead to a different ratio between the *RIRE3* family and the two others.

Occurrence of recombination between LTRs may also be influenced by LTR size. If one considers that recombination occurs randomly along LTRs, then it could be hypothesized that retroelements with large LTRs would tend to recombine more than smaller ones. Here, we compared three elements with different LTR sizes (i.e., about 400 bp for *Retrosat1*, 1200 bp for *hopi*, and 3200 bp for *RIRE3*). Although our sample size is small, our data suggest that, globally, the proportion of solo-LTRs may increase with LTR size. Since we do not know yet how the recombination process takes place through time (see *Relative Timing of Amplification and Reduction*), the evaluation of the impact of LTR size on LTR recombination is nevertheless difficult to directly assess from the ratio of solo-LTRs to complete copies from retroelements that have not inserted at the same time. These results nevertheless clearly show that, in order to estimate the extent of the decreasing process, we have to take into account the specific features of each retroelement family that a genome contains, instead of analyzing globally the occurrence of solo-LTRs. For this reason, we cannot yet make comparisons between the ratios of solo-LTRs to complete copies for the two retroelement *gypsy* and *copia* types within the rice genome or for one retroelement family in different genomes, because we lack data concerning both timing of the copies and the ratios of LTR to complete copies for individual families.

In order to study the occurrence of solo-LTRs originating from the recombination of two different copies, we analyzed the flanking regions of the 64 solo-LTRs. All solo-LTRs but four showed perfect duplicated target sites. For three among these four, the duplicated target site is imperfect (gcgga/gtgga, ggcgt/ggcat, and ccgca/ctgca, respectively) but still recognizable. For the last copy (AC080019C), no duplicated target site could be revealed, as the flanking regions are different (catta/ctgtc). Hence, this copy might result from the recombination between two different copies, whereas the others might be the result of intracopy recombination events. This suggests that solo-LTRs form preferentially by unequal homologous recombination of two LTRs of the same copy. If such copies have not been the target of other(s) element(s), solo-LTR formation might thus reduce genome size, although not sufficiently to reverse the amplification process.

Relative Timing of Amplification and Reduction

In order to analyze the timing of the decreasing process, we examined the clustering of the solo-LTR sequences with the LTRs of the complete elements. Results presented in figures 4*B*, 5*B*, and 6*B* show that all solo-LTRs cluster with the groups of complete copies described. Hence, solo-LTR formation seems to be concomitant with the amplification of active copies.

Considering each family independently, solo-LTRs seem to be more abundant in old groups than in young ones (figs. 4*B*, 5*B*, and 6*B*). For example, in the *hopi* family, it appears in fig. 4*B* that the three oldest groups have more solo-LTRs (ratios of solo-LTR to complete copies are 7:5, 11:9, and 19:14, respectively, for subgroup 1B, group 2, and group 3) than the very recent one (subgroup1A, ratio 1:21). Nevertheless, in absence of a larger sample, we cannot yet determine if these differences are correlated with the age of the groups. It is thus not clear whether the decreasing process is continuous through time or driven by large recombination events as proposed by Rabinowicz (2000).

Finally, the extensive characterization of the retroelements that compose the nongenic compartment of the rice genome will allow to further elaborate our increase/decrease model, by the precise determination of both extent and timing of the amplification and LTR recombination processes.

Conclusion

The analysis of copies from three rice LTR retroelements retrieved from the rice genomic sequence shows that the rice genome has undergone LTR retrotransposon amplification events over the past 5 Myr. During these events, only a few master copies seem to have amplified, leading to the formation of structured groups within each family. Since their insertion, some copies have undergone unequal homologous recombination events that lead to the formation of solo-LTRs. Recombination seem to have occurred preferentially in old groups of copies and could be due to a continuous or to a massive process. We thus propose an increase/decrease model of grass genome evolution, in which both increasing and decreasing mechanisms drive genome size variations. Nevertheless, this evolutionary model has to be completed by the analysis of the extent of both these counteracting mechanisms. This will be possible through the extensive analysis of copies from a large number of retroelements, as soon as the rice genomic sequence is complete.

Acknowledgments

We thank Eric Coissac for his help in bioinformatics and Marie-Angèle Grandbastien and Dominique Higuet for their valuable suggestions and comments on the manuscript and for their friendly support.

Literature Cited

Ahn, S., and S. D. Tanksley. 1993. Comparative linkage maps of the rice and the maize genomes. Proc. Natl. Acad. Sci. USA **92**:7980–7984.

Barakat, A., N. Carels, and G. Bernardi. 1997. The distribution of genes in the genomes of Gramineae. Proc. Natl. Acad. Sci. USA **94**:6857–6861.

Bautista, N., R. Solis, O. Kamijima, and T. Ishii. 2001. RAPD, RFLP and SSLP analyses of phylogenetic relationships between cultivated and wild species of rice. Genes Genet. Syst. **76**:71–79.

Bennett, M. D. 1998. Plant genome values: how much do we know? Proc Natl. Acad. Sci. USA **95**:2011–2116.

Bennetzen, J. L. 2000. Transposable element contribution to plant gene and genome evolution. Plant Mol. Biol. **42**: 251–269.

Bennetzen, J. L., and E. A. Kellogg. 1997. Do plants have a one-way ticket to genomic obesity? Plant Cell **9**:1509–1514.

Bennetzen, J. L., P. SanMiguel, M. Chen, A. Tikhonov, M. Francki, and Z. Avramova. 1998. Grass Genomes. Proc. Natl. Acad. Sci. USA **95**:1975–1978.

Bowen, N. J., and J. F. McDonald. 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. **11**:1527–1540.

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

Dong, F., J. T. Miller, S. A. Jackson, G. L. Wang, P. C. Ronald, and J. Jiang. 1998. Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc. Natl. Acad. Sci. USA **95**: 8135–8140.

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

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

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

Ge, S., T. Sang, B. Lu, and D. Hong. 1999. Phylogeny of rice genomes with emphasis on origin of allotetraploid species. Proc. Natl. Acad. Sci. USA **96**:14400–14405.

Hirochika, H., A. Fukuchi, and F. Kikuchi. 1992. Retrotransposon families in rice. Mol. Gen. Genet. **233**:209–216.

Hirochika, H., K. Sugimoto, Y. Otsuki, H. Tsugawa, and M. Kanda. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. USA **93**: 7783–7788.

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

Jordan, I. K., and J. F. McDonald. 1998. Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae Ty* elements. J. Mol. Evol. **47**:14–20.

Katayama, T. 1967. Cytogenetical studies on the genus *Oryza*. IV. Cytogenetical studies on the first backcross generation of the (A × BC) × A and (A × CD) × A genomes. Jpn. J. Genet. **42**:160–174.

———. 1982. Cytogenetical studies on the genus *Oryza*. XIII. Relationship between the genomes E and D. Jpn. J. Genet. **57**:613–621.

Kidwell, M. G., and D. Lisch. 1997. Transposable elements as sources of variation in animal and plants. Proc. Natl. Acad. Sci. USA **94**:7704–7711.

Kumekawa, N., N. Ohmido, K. Fukui, E. Ohtsubo, and H. Ohtsubo. 2001. A new *gypsy*-type retrotransposon, *RIRE7*: preferential insertion into the tandem repeat sequence *TrsD* in pericentromeric heterochromatin regions of rice chromosomes. Mol. Genet. Genomics **265**:480–488.

Kumekawa, N., H. Ohtsubo, T. Horiuchi, and E. Ohtsubo. 1999. Identification and characterization of novel retrotransposons of the *gypsy* type in rice. Mol. Gen. Genet. **260**: 593–602.

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

Noma, K., R. Nakajima, H. Ohtsubo, and E. Ohtsubo. 1997. *RIRE1*, a retrotransposon from wild rice *Oryza australiensis*. Genes Genet. Syst. **72**:131–140.

Ohtsubo, H., N. Kumekawa, and E. Ohtsubo. 1999. *RIRE2*, a novel *gypsy*-type retrotransposon from rice. Genes Genet. Syst. **74**:83–91.

Panaud, O., C. Vitte, J. Hivert, S. Muszlak, J. Talag, D. Brar, and A. Sarr. 2002. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using representational difference analysis. Mol. Gen. Genomics **268**:113–121.

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

Petrov, D., T. Sangster, J. Johnston, D. Hartl, and K. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. Science **287**:1060–1062.

Promislow, D. E., I. K. Jordan, and J. F. McDonald. 1999. Genomic demography: a life-history analysis of transposable element evolution. Proc. R. Soc. Lond. B. Biol. Sci. **266**:1555–1560.

Rabinowicz, P. D. 2000. Are obese plant genomes on a diet? Genome Res. **10**:893–894.

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

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

Second, G. 1985. Relations évolutives chez le genre *Oryza* et processus de domestication des riz. Paris, Paris Sud, Orsay.

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

Statsoft. 1997. Statistica 5.1 software.

Swofford, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey, and A. Rafalski. 2000. The complete sequence of 340 kb of DNA around the rice *ADH1-ADH2* regions reveals interrupted colinearity with maize chromosome 4. Plant Cell **12**:381–391.

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

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 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., P. J. SanMiguel, Y. Nakajima, N. D. Gorenstein, J. L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous *ADH* regions of maize and sorghum. Proc. Natl. Acad. Sci. USA **96**:7409–7414.

Tristem, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. **74**:3715–3730.

Vicient, C., and A. H. Schulman. 2002. Copia-like retrotransposons in the rice genome: few and assorted. Genome Lett. **1**:35–47

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

Wang, Z., G. Second, and S. Tanksley. 1992. Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. Theor. Appl. Genet. **83**:565–581.

Pierre Capy, Associate Editor

Accepted November 18, 2002