Research

Long terminal repeat retrotransposons of *Oryza sativa*

Eugene M McCarthy*, Jingdong Liu†, Gao Lizhi* and John F McDonald*

Addresses: *Department of Genetics, University of Georgia, Athens, GA 30602, USA. †Monsanto, St. Louis, MO 63198, USA.

Correspondence: Eugene M McCarthy. E-mail: gm@uga.edu

Published: 13 September 2002

Received: 28 December 2001

Genome Biology 2002, 3(10):research0053.1–0053.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/10/research/0053

© 2002 McCarthy et al., licensee BioMed Central Ltd

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Long terminal repeat (LTR) retrotransposons constitute a major fraction of the genomes of higher plants. For example, retrotransposons comprise more than 50% of the maize genome and more than 90% of the wheat genome. LTR retrotransposons are believed to have contributed significantly to the evolution of genome structure and function. The genome sequencing of selected experimental and agriculturally important species is providing an unprecedented opportunity to view the patterns of variation existing among the entire complement of retrotransposons in complete genomes.

Results: Using a new data-mining program, LTR_STRUC, (LTR retrotransposon structure program), we have mined the GenBank rice (*Oryza sativa*) database as well as the more extensive (259 Mb) Monsanto rice dataset for LTR retrotransposons. Almost two-thirds (37) of the 59 families identified consist of *copia*-like elements, but *gypsy*-like elements outnumber *copia*-like elements by a ratio of approximately 2:1. At least 17% of the rice genome consists of LTR retrotransposons. In addition to the ubiquitous *gypsy-* and *copia*-like classes of LTR retrotransposons, the rice genome contains at least two novel families of unusually small, non-coding (non-autonomous) LTR retrotransposons.

Conclusions: Each of the major clades of rice LTR retrotransposons is more closely related to elements present in other species than to the other clades of rice elements, suggesting that horizontal transfer may have occurred over the evolutionary history of rice LTR retrotransposons. Like LTR retrotransposons in other species with relatively small genomes, many rice LTR retrotransposons are relatively young, indicating a high rate of turnover.

Background

Retrotransposons are mobile genetic elements that make up a large fraction of most eukaryotic genomes. They are particularly abundant in plants, where they are often a principal component of nuclear DNA. In maize 50-80%, and in wheat fully 90%, of the genome is made up of retrotransposons [1,2]. In animals this percentage is generally lower than in plants but can still be large. For example, more than 40% of the human genome is now known to be composed of retroelements [3,4]. All retrotransposons are distinguished by a life cycle involving an RNA intermediate. The RNA genome of a retroelement is copied into a double-stranded DNA molecule by reverse transcriptase and is subsequently integrated into the host’s genome. Retrotransposons fall into two main categories, those with long terminal repeats (LTRs), such as retroviruses and LTR retrotransposons, and those that lack such repeats, (for example, long interspersed nuclear elements or LINEs).
Our laboratory is in the process of screening the GenBank rice (*Oryza sativa*) database (GBRD) and the Monsanto rice dataset (MRD) for the presence of LTR retrotransposons. We have chosen to scan the rice genome because, as the most important food crop in the world, much of its sequence data is already available. With a haploid content of 430 million base pairs (Mbp), the rice genome is the smallest among cultivated cereals [5,6] and only about three times larger than the smallest known genome among angiosperms, that of *Arabidopsis thaliana* (~130 Mbp). *O. sativa* has one of the smallest genomes among grasses as a whole [6]. Genomes of other cereals are far larger. For example, the maize (*Zea mays*) genome is 2,500 million base pairs (2.5 Gbp) and that of wheat (*Triticum aestivum*), 16 Gbp. The molecular genetic resources for rice are excellent, including detailed physical and genetic maps, large YAC and BAC libraries, an efficient transformation system, and an extensive collection of expressed sequence tags (ESTs).

We have used a new search program, LTR_STRUC (LTR retrotransposon structure program; E.M.M. and J.F.M., unpublished work), as the initial data-mining tool in our survey. Structural features important to the algorithm on which LTR_STRUC is based include two sites critical to replication, the primer-binding site (PBS) and polypurine tract (PPT), as well as the presence of canonical dinucleotides at the ends of each LTR (typically TG and CA). Particularly important are the direct or ‘target-site’ repeats (TSRs). When an LTR retrotransposon inserts itself into host DNA, a short (usually 4-6 bp) segment of host DNA is replicated at the site of insertion. This feature allows LTR_STRUC to make an exact demarcation of the limits of a putative element. Because it searches for retroelements on the basis of their generic structure, LTR_STRUC eliminates much of the bias inherent in BLAST searches based on a known retroelement query. After elements were initially identified using LTR_STRUC, sequence analyses were carried out to identify open reading frames (ORFs) encoding reverse transcriptase (RT) and other retrotransposon proteins. Subsequent RT sequence alignments were carried out, followed by construction of phylogenetic trees.

RTs from elements identified in our survey fall into numerous distinct families, where ‘family’ is defined as a group of elements with RTs having mutual similarity of at least 90% at the amino-acid level [7]. In addition, four types of non-autonomous elements discussed here lack RT sequences (*Osr25, Osr37/Rire4, Osr43, and Osr44*), and were classified as distinct families on the basis of their unique structures (see below).

Currently, there is no consensus with respect to rice retrotransposon nomenclature. In our method of nomenclature, rice LTR retrotransposons are specified by the appellation *Osr* (*Oryza sativa* retrotransposon). Distinct families are indicated by number (for example, *Osr1, Osr2, Osr3, ...*).

There have been four different nomenclatures previously used in reference to rice LTR retrotransposons: *Tos* (transposon *Oryza sativa*) [8], *Rire* (rice retrotransposon) [9] *Rrt* (rice retrotransposon) (S. Wang, submission to EMBL database: *Rt3* (accession number T03666), *Rrt5* (T03669), and *Rrt8* (T03671), and *Osr* (*Oryza sativa* retrotransposon) (N. Jwa, submission to GenBank: *Osr1* (AB046118)). We have chosen to adopt the *Osr* nomenclature in this study because it is consistent with the systematic logic (indicative of genus and species of host organism) used in previous genomic studies of LTR retrotransposons and includes the letter ‘r’ to indicate retrotransposon. However, in every case where we use the *Osr* acronym in this paper to refer to a previously named family, we also include any pre-existing name(s) for the family (for example, *Osr15/Tos12, Osr26/Rire2*).

**Results and discussion**

As is the case for most eukaryotic species analyzed to date, rice LTR retrotransposons fall, for the most part, into two major categories, *gypsy*-like and *copia*-like (two exceptions are discussed below). *Copia*-like elements in the rice genome are usually 5-6 kb in length; however, certain families are composed of longer elements so that the mean length is around 6.2 kb. For example, elements in *Osr7* and *Osr8* are about 9,000 bp in length. Results of our study indicate that the TSSRs of all rice LTR retrotransposons are 5 bp long (Table 1). The dinucleotides terminating the LTRs are similar invariant: across all families, the 5’ nucleotide pair is consistently TG, and the 3’ end, consistently CA (except for a few mutated copies). In the rice genome, normal *gypsy*-like elements (that is, those that lack a deletion or insertion) are typically in the 10 to 13 kb range, but some do bear large insertions or internal deletions. Their mean length of 11.7 kb is larger than that of typical *gypsy*-like elements in other species, which are usually in the range of 7-8 kb [7,10]. The reason for this larger mean length of *O. sativa* LTR retrotransposons is presently unknown. Duplication of retroelement sequences during the process of reverse transcription has been previously observed in mammalian systems [11] and nested insertions of transposons into LTR retrotransposons are not uncommon in plants [12]. However, none of the full-length LTR retrotransposons reported here has a substructure consistent with nested LTR retrotransposon insertions. For example, none of the elements we report in Table 1 encode more than one region of RT homology and none contain nested pairs of putative LTRs. Of course, we cannot eliminate the possibility that the larger size of *O. sativa* *gypsy*-like elements is, at least in part, due to insertions of unrecognized elements or ancient insertions of known elements that can no longer be recognized. Whatever, the reason for the exceptional size of *O. sativa* *gypsy*-like elements, it apparently does not inhibit function, as sequence analysis (see below) indicates that the majority of these elements have transposed in the recent evolutionary past. *Gypsy*-like elements in *O. sativa* also have larger LTRs.
### Table 1

**Summary of rice LTR retrotransposons characterized in this study**

| Family | Pre-existing name(s) | Accession number of exemplar | Location | Chromosome number | LTR length (bp) | Inserted element length | TSR | %LNI (mean for family) | Approximate copy number (haploid genome) |
|--------|----------------------|-------------------------------|----------|-------------------|-----------------|-------------------------|-----|------------------------|------------------------------------------|
| Osr1   | Tos14/Rire15         | AC023240                      | 100410-106807 | 10                | 965             | 6,398                   | AGTCC | 98.1                  | 250                                      |
| Osr2   | AL442110             |                              | 95121-100070 | 4                 | 267             | 4,950                   | ATATT | 98.5                  | <50                                      |
| Osr3   | AF458765             | 51-525                        | ?         | 146               | 5,200           |                         | CATTC | 99.3                  | 50-100                                   |
| Osr4   | AB026295             | 160208-165872                 | 6         | 350               | 5,665           |                         | GTTAC | 98.9                  | <50                                      |
| Osr5   | AC021891             | 56044-62135                   | X         | 1608              | 8,920           |                         | AGTTT | 98.8                  | <50                                      |
| Osr6   | AP001369             | 25869-28634                   | I         | 1220              | 9,216           |                         | TAAAT | 97.2                  | 1100                                     |
| Osr7   | Osr7                 |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr8   | Osr8                 |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr9   | Osr9                 |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr10  | Osr10                |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr11  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr12  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr13  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr14  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr15  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr16  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr17  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr18  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr19  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr20  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr21  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr22  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr23  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr24  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr25  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr26  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr27  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr28  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr29  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr30  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr31  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr32  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr33  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr34  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr35  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr36  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr37  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr38  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr39  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr40  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr41  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr42  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr43  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |
| Osr44  | Rire1               |                              |           | ND                | ND              |                         | ND    | ND                    | ND                                       |

*Location given is for an example RT in the GBRD (no full-length element was identified for this family). †As a full-length element is known in the MRD, the TSR and lengths of the LTR and element (columns 5-7) are taken from an element in the MRD while the location (if given) in columns 2-4 refers to an RT in the GBRD. ‡Percentages based on number of hits using a sample LTR from each family as query to search the MRD. ††N. Jiang and S.R. Wessler (unpublished work) suggest that if pericentric DNA (which is largely heterochromatic) is taken into account, Osr25 elements exist at a higher copy number (~1,000 copies in the entire genome) than our survey, based largely on euchromatic sequences, would suggest. ND, not determined.
Genome Osr43 identified two families of LTR retrotransposons (57 families of gypsy-like elements, a total of 37. In addition to libraries [8].

Our survey has identified numerous LTR retrotransposon families that have not been described previously. These findings show that at least 59 distinct LTR retrotransposon families exist in the rice genome. This result compares with an earlier family estimate of 32 based on screening genomic libraries [8]. Copia-like elements are less numerous than gypsy-like elements in the rice genome, but they still comprise more than half the families, a total of 37. In addition to 57 families of copia- and gypsy-like elements, we have identified two families of LTR retrotransposons (Osr43 and Osr44) that show no significant sequence similarity to any known transposon.

For the purposes of this analysis, a ‘full-length element’ is defined as one that has two complete and recognizable LTRs. Any other LTR retrotransposon sequence is here defined as a ‘fragment’. The results of our survey of the GBRD and MRD suggest that there are in the order of 450 full-length copia-like elements in the entire rice genome. We found full-length copia-like elements both with and without RT domains. We estimate the total copy number (including fragmentary copies) at 3,500, or about 3% of the genome. BLAST searches with representative LTR queries from each of the rice LTR-retrotransposon families against the MDR indicate that gypsy-like elements are twice as common (total copy number ~7000; ~1,400 full-length). Previous estimates of this ratio have been somewhat higher [13]. Owing in part to their large LTRs, gypsy-like elements in rice are twice as long as copia-like elements (11.7 kb versus 6.2 kb) and so make up a proportionately larger fraction of the genome (~14%). That is, a total of about 17% of the genome is composed of LTR retrotransposon sequences. This estimate exceeds those of previous workers [8,13-15]. For example, using a variety of RT probes Wang et al. [14] estimated that around 100 copies of copia-like elements are present in the entire haploid genome. This estimate did not discriminate between full-length and fragmentary copies. From our examination of the searchable portion of the GBRD alone (which represented at the time approximately 10% of the rice genome), we have identified the actual sequences for 46 separate full-length copia-like elements. This implies that the number of full-length copia-like elements in the whole genome should be about ten times higher, that is, around 450 to 500 elements. In an analysis of 340 kb around the Adh1-Adh2 region of the rice genome, Tarchini et al. [16] reported that 14.4% of this region consisted of LTR retrotransposons. This value is in reasonably good agreement with our estimate of about 17%. Mao et al. [15] give a lower figure (9.3%) but we believe our higher figure is more accurate because their study sought homology to known retrotransposon sequences and such homology would be undetectable for the many new families of retrotransposons presented here. Similarly, they give a higher ratio of gypsy- to copia-like elements, but they may not have been aware that gypsy-like elements are significantly larger in rice, which would inflate their estimate of this ratio.

The previous low estimates of copy number given for rice LTR retrotransposons are probably attributable to three factors. First, these earlier studies used an incomplete set of RTs as probes for hybridization (or as queries for BLAST). For example, Osr8, a high copy copia-like family, was not recognized in previous studies. Second, a number of rice LTR retrotransposons lack an RT ORF and would thus go undetected in studies using RT probes. In particular, no member of families Osr25 and Osr37/Rire4 seem to have a RT (yet these two families have a total copy number of around 900 elements). Third, data-mining with LTR_STRUCT (see Materials and methods) allows a higher degree of assurance that the putative RTs detected in the survey actually are RTs because it places putative polyproteins in the context of a canonical retroviral structure. Such is not the immediate result of a simple BLAST with an RT query. Our estimate that LTR retrotransposons make up 17% of the rice genome is conservative, inasmuch as our study was based primarily on euchromatic sequences and did not include elements present within the traditionally retrotransposon-rich heterochromatin [14,17]. Thus, our results bring the rice genome closer to the LTR retrotransposon densities reported for other cereals.

Intra-element percent LTR nucleotide identity

Because of the replication process characteristic of LTR retrotransposons, the LTRs of a given retroelement are sequentially identical at the time the element inserts into the host genome [18]. Thereafter, as an element accumulates mutations, its LTRs become increasing different from each other as substitutions specific for each of the two LTRs increase in number. The level of nucleotide identity seen between LTRs of a particular element, usually referred to as intra-element percent LTR nucleotide identity (%LNI), can be used in determining the relative ages of LTR retrotransposon families [7]. In rice, comparison of the two LTRs of the same element often showed the presence of a 10 to 30 bp regional duplication present in one LTR but not the other. In calculating %LNI, we have considered such duplications as single mutation events.

As the neutral nucleotide substitution rate has yet to be computed for rice, we cannot presently equate %LNI with a divergence time in years. However, the generally low level of sequence divergence between flanking LTRs of rice LTR retrotransposons (1.7%) indicates that most of the euchromatic full-length LTR retrotransposons in rice are relatively young, although significantly older elements were also identified. The seeming preponderance of young full-length LTR retrotransposons in the euchromatin of rice is similar to previous reports on yeast [19,20], Caenorhabditis elegans [7],...
A. thaliana [21] and Drosophila melanogaster [12]. This contrasts with findings in Z. mays [12] and humans [22].

Copia-like families
To date, 23 families of copia-like elements have been reported for rice (S. Wang, submission to EMBL, N. Jwa, submission to GenBank, and [8,9,19,23,24]). Several have been described under more than one name. For example, the amino-acid sequence given for Tos4 in Hirochika et al. [23] is the same as that given for Tosi in GenBank (accession number S22455) so they are really the same. Rire5 described by Kumekawa et al. [25] is the same family as Tos14 previously described by Hirochika et al. [23]. The equivalence between Tos14 and Rire5 became evident when we found the LTR sequence reported by Kumekawa et al. in elements that also contained the RT sequence given by Hirochika for Tos14. In our survey of GenBank and MRDB, we have identified an additional 16 copia-like families that have not been described by previous workers. In addition, exemplars for each of the previously identified families were found (except in the case of certain families that exist at such low copy numbers that no full-length element exists in GenBank or MRDB).

The largest copia-like family
One of the most interesting new finds in our survey was Osr8, one of the oldest families of LTR retrotransposons in the rice genome. On the basis of a survey of the available portion of the GBRD and MRD, we estimate the copy number of Osr8 to be around 1,100 (more than any other copia-like family). Osr8 elements exist far more frequently as fragments (ratio of 1:1) and they display relatively low levels of %LNI in their full-length copies (mean %LNI for the five full-length Osr8 elements present in the GBRD is 97.2%). The RT of Osr8 is 60% similar to an unnamed polyprotein in Z. mays (AAD20397). A closely related family, Osr10 has two full-length copies in the GBRD but scans of the MRD suggest this element, also previously unrecognized, has the third highest copy number (~400) among copia-like elements. Outside rice, the RT of Osr10 shows highest similarity (~65%) to that of the maize retrotransposon Opie-2 (T04112). The broader clade that includes Osr7, Osr8, Osr9, and Osr10 is closely related to Endovir1-1 (AAAG29949) of Arabidopsis (Figure 1, Table 2). These elements are also related (~60% similar) to maize’s PREM-2 as well as to tomato’s ToRTL1. Both Osr7 and Osr9 are present in very low copy number (one full-length and a few fragments in the GBRD).

Osr14/Tosi/Tos4, Osr15/Tos12 and Osr53/Tos18
Although it is present at only a quarter of the copy number of Osr8, the unrelated Osr14/Tosi/Tos4 is also composed primarily of highly fragmented elements. Those that are full length have low %LNI (family mean 97.6%). Thus, Osr14/Tosi/Tos4 and Osr8 seem to be of similar age and to have followed a similar evolutionary pattern, albeit with less intense amplification in the case of Osr14/Tosi/Tos4. Osr14/Tosi/Tos4, Osr15/Tos12, and Osr53/Tos18 form a well defined clade and are more closely related to Ta1-2 (S23315) of Arabidopsis than to any other rice retroelement family outside their clade (Figure 1, Table 2). Osr15/Tos12 and Osr53 are only just sufficiently different to constitute distinct families.

A quartet of closely allied families
Osr1/Tosi14/Rire5, Osr13/Tos5, Osr51/Tos15, and Osr52/Tos16 have been described as distinct families but, inasmuch as their RTs are all 85% similar to each other, these groups are only marginally distinct. Searches of GenBank show that elements in this group are much more closely related to (75-80% at the amino-acid level) to maize retrotransposon Fourf (AAK73108) than to any rice LTR retrotransposon outside their clade. If the elements belonging to this group were considered to be a single family, it would be almost as large (~900 elements) as Osr8. In the GBRD the majority of these elements are fragmentary, but the estimated copy number of full-length elements in the rice genome for this quartet still exceeds 100.

A Hopscotch-like clade of fragmented elements
Osr18, Osr19, Osr20, Osr22, Osr23, Osr24, Osr45/Tos7, and Osr46/Tos8 form a clade of low copy number families composed primarily of fragmentary copies. Our results suggest that each of these families has a copy number in the range of 50-100 elements. Members of this clade are closely related to maize’s Hopscotch element (T04112) (Figure 1, Table 2).

Low copy number copia-like families
Osr2 and Osr12 are low-copy families and are represented in the GBRD by two and three copies respectively, all of which are full length (although one copy of Osr12 contains a large internal deletion), suggesting that these elements may have recently invaded the rice genome. The high level of LTR nucleotide identity (>99%) seen in these elements is consistent with this recent invasion hypothesis. Members of Osr12 and Osr2 are potentially active because they have large, intact polyprotein ORFs, usually in excess of 1,000 amino acids. All three Osr12 elements detected in the GBRD are on chromosome 10. Similarly, both Osr2 elements are inserted within 50 kb of each other on chromosome 4. Nonetheless, these two families are not closely related (their RT sequences are only ~50% similar at the amino-acid level). Osr12 RTs differ from those of all other rice copia-like elements by 50%. And yet RT sequences of elements in Osr12 are 60% similar to certain elements in the maize genome (Zmri (S27768) and mzecopia (M94481.1)).

One full-length, and one fragmented copy of Osr6 are present in the GBRD. Osr5 is slightly more common than Osr6, to which it is most closely related, but it is currently represented in the GBRD by only a single full-length copy and a few fragments. Osr5 is 60% similar to the tobacco retrotransposon Tnt1-94 at the amino-acid level (RT comparison). Osr4 is
Figure 1
RT-based neighbor-joining tree for copia-like retrotransposons. Distances (uncorrected p) appear next to each branch. RT sequences from plant species other than rice are included for comparison.
another low-copy family. It has several fragmented representatives in the GBRD, and is probably somewhat older than Osr12 and Osr2, but it has only three full-length copies in the GBRD, Osr4 elements have an exceptionally large polyprotein ORF (~1,600 amino acids). The RT of Osr4 shows 50% similarity to that of retroelements in the Arabidopsis genome (for example, BAB01972, NP_175303).

Although the RT of Osr3 was detected during our survey, elements in this family are fragments with ill defined LTRs. TBLASTN reveals the RT of Osr3 to be the single representative of its type in the GBRD. Both Osr3 and the equally aberrant Osr21/Tos17 differ from those of other copia-like elements found in our study by about 55%. Osr11/Rire1 is a low-copy family closely related (75% similarity) to a retroelement in the Arabidopsis genome (At-2, T01860). Two other closely related families are Osr16/Tos6 and Osr17, both of which are similar to Sto-4 (T17429) of maize (Figure 1, Table 2). Nine additional low-copy families identified by earlier workers are Osr47/Tos9, Osr48/Tos10, Osr49/Tos11, Osr50/Tos13, Osr54/Tos19, Osr55/Tos20, Osr57/Rtr3, Osr58/Rtr5, and Osr59/Rtr8. Source references for each of these nine families are given in Table 3.

### Table 2

| Name of retrotransposon | Accession number | Host organism |
|-------------------------|------------------|---------------|
| Opie-2                  | T04112           | Z. mays       |
| Hopscotch               | T02087           | Z. mays       |
| Fourf                   | AAK73108         | Z. mays       |
| Sto-4                   | T17429           | Z. mays       |
| Znr-1                  | S27768           | Z. mays       |
| Endowir1-l              | AAG52949         | A. thaliana   |
| Ta1-2                  | S23315           | A. thaliana   |
| Atr-1*                 | NP_175303        | A. thaliana   |
| Atr-2*                 | T01860           | A. thaliana   |
| Atr-3*                 | NP_178752        | A. thaliana   |
| Atr-4*                 | NP_174802.1      | A. thaliana   |
| Atr-5*                 | AAF13073.1       | A. thaliana   |
| Atr-6*                 | NP_179047        | A. thaliana   |
| Retrosor1               | AAD19359         | Sorghum bicolor |
| Retrosor3               | AAD22153         | S. bicolor    |
| Daniela                 | AF326781†        | Triticum aestivum |
| Acr-1*                  | CAAA73042        | Ananas comosus |

*Previously unnamed RT found by BLAST searches of the GBRD, using rice RTs found in our study as queries. Acr, Ananas comosus retrotransposon; Atr, A. thaliana retrotransposon; Znr, Z. mays retrotransposon.

### Table 3

| Family | Pre-existing family name | Accession number (or source) of sequence |
|--------|--------------------------|----------------------------------------|
| Osr45  | Tos7                     | T03709                                 |
| Osr46  | Tos8                     | T03704                                 |
| Osr47  | Tos9                     | T03705                                 |
| Osr48  | Tos10                    | T03706                                 |
| Osr49  | Tos11                    | T03707                                 |
| Osr50  | Tos13                    | Hirochika et al. [23]                  |
| Osr51  | Tos15                    | T03711                                 |
| Osr52  | Tos16                    | T03712                                 |
| Osr53  | Tos18                    | T03716                                 |
| Osr54  | Tos19                    | T03721                                 |
| Osr55  | Tos20                    | T03723                                 |
| Osr56  | Rire3                    | Kumeukawa et al. [25]                  |
| Osr57  | Rtr3                     | T03666                                 |
| Osr58  | Rtr5                     | T03669                                 |
| Osr59  | Rtr8                     | T03671                                 |

### Table 3

| Family | Pre-existing family name | Accession number (or source) of sequence |
|--------|--------------------------|----------------------------------------|
| Osr45  | Tos7                     | T03709                                 |
| Osr46  | Tos8                     | T03704                                 |
| Osr47  | Tos9                     | T03705                                 |
| Osr48  | Tos10                    | T03706                                 |
| Osr49  | Tos11                    | T03707                                 |
| Osr50  | Tos13                    | Hirochika et al. [23]                  |
| Osr51  | Tos15                    | T03711                                 |
| Osr52  | Tos16                    | T03712                                 |
| Osr53  | Tos18                    | T03716                                 |
| Osr54  | Tos19                    | T03721                                 |
| Osr55  | Tos20                    | T03723                                 |
| Osr56  | Rire3                    | Kumeukawa et al. [25]                  |
| Osr57  | Rtr3                     | T03666                                 |
| Osr58  | Rtr5                     | T03669                                 |
| Osr59  | Rtr8                     | T03671                                 |

**Gypsy-like families predominate in O. sativa**

Osr27/Rire9 [26] is the third largest family in the rice genome, with an estimated copy number of 900 elements, mostly full length. Li et al. [26] estimated the copy number of this family at 1,600. The typical Osr27/Rire9 element is quite large (~12.8 kb total length). Having intact polyprotein ORFs and high mean %LNI (99%), these elements probably are, or recently have been, actively transposing. Yet the presence of a few members of this family that are more mutated (short ORFs, low LTR-LTR nucleotide identity) suggests that this may also be an ancient family. Two other families, Osr40 and Osr41, are also members of the same clade as Osr27/Rire9, Osr25 and Osr26/Rire2 (Osr25 and Osr26/Rire2 are discussed below), but both have RTs that are about 30% different from those of Osr26/Rire2 and Osr27/Rire9. Neither Osr40 nor Osr41 has been previously identified, but with approximate copy numbers of 600 and 300, respectively, these are both large families. The RTs of members of this clade show about 60% similarity to that of Retrosor1 (Sorghum bicolor, AAD19359).

With approximately 1,500 elements, Osr30 constitutes 14% of all LTR retrotransposons in the rice genome. Although Osr30 is the largest family of LTR retroelements in the genome, it has not been previously named. These elements are slightly larger (~13.1 kb) than those of Osr27/Rire9. A higher proportion of fragmented copies and lower level of LTR-LTR nucleotide identity suggest that Osr30 is older
than Osr27/Rire9. Osr29, which is closely allied to Osr30, is also a large family with more than 500 member elements. Taken together, the elements of the Osr29 and Osr30 clade are unusual, because they are as closely related to other major rice clades as they are to any elements outside rice. Osr28 is a low-copy family that is most closely related to Osr29 and Osr30 (Figure 2).

Two other large gypsy-like families are Osr33/Rire8 [25] and Osr34. These two families each have copy numbers of approximately 500. Two low-copy families belonging to the same clade are Osr32 and Osr56/Rire3 [27] (Figure 2). Members of these families have large LTRs, typically in the range 3,000-3,500 bp. RTs of families in this clade show high sequence similarity to an LTR retrotransposon in pineapple (~70% to Acr-1; CAA73042) and to one in Sorghum bicolor (~77% to Retrosor3, AAD22153) (Figure 2).

Low-copy gypsy-like elements
Osr31/Rire7 is an aberrant low-copy family that is much more closely related (77% similarity) to an Arabidopsis element, Atr-4 (see Table 2), than to any other LTR retroelement families in the rice genome (Figure 2). In the clade of five low-copy families, composed of Osr35, Osr36, Osr38, Osr39, and Osr42, an RT was found in the GBRD for only two families, Osr35 and Osr36. The other elements were identified in scans of the MRD and their full sequences have since been submitted to GenBank (for accession numbers, see Table 1). This clade is closely related to Arabidopsis element Atr-5 (Figure 2, Table 2).

Families of non-autonomous elements
Members of family Osr25 are all internally deleted and thus non-autonomous (mean length 4.3 kb). Although Osr25 elements have typical LTRs, PBS, and PPT, the inter-LTR region contains only non-coding, repetitive DNA. The LTRs of Osr25 display 65-70% sequence similarity to the autonomous elements of the gypsy-like family Osr26/Rire2. Elements with LTRs having such a high degree of similarity are usually considered members of the same family. Nevertheless, because members of Osr26/Rire2 have the usual coding structure typical of other gypsy-like elements (while Osr25 elements entirely lack typical retroviral genes) and members of these two families fall into two sharply distinct, non-overlapping clades, we report these two types of elements as separate families. Estimates based on scans of the MRD and the GBRD suggest that the rice genome contains about 500 copies each of Osr25 and Osr26/Rire2. Osr25 and Osr26/Rire2 display 98.9 and 97.9% LNI respectively.

Osr37/Rire4 is also aberrant compared to other rice LTR retrotransposon families. The typical element in this family is 4.4 kb long, about the same length as Osr25 elements. Members of Osr37/Rire4 usually carry a large ORF (up to 600 amino acids) just upstream of the 3′ LTR. This ORF shows no significant similarity to any known RT sequence. Up to the present in the GBRD, where these ORFs are generally identified simply as hypothetical proteins, the large ORF of Osr37/Rire4 seems not to have been recognized as a retroviral gene. This ORF may serve an integrase function as BLAST searches show that it has low homology to a putative integrase in A. thaliana (28%; ACX05171). There are about 600 copies of Osr37/Rire4 in the entire rice genome.

In addition to the foregoing copia- and gypsy-like families, our scans identified two families, Osr43 and Osr44, of small elements (overall length < 2,000 bp). With LTRs only 148 bp long and an overall length of 1,207 bp, Osr44 elements are especially small. Members of Osr43 and Osr44 are unique because, although they possess all of the canonical LTR-retrotransposon structural features (LTRs, PBS, PPT, and TSRs), they are internally deleted and either completely lack or encode only very small ORFs with no similarity to any known protein. Both families contain on the order of 100 copies genome-wide.

Conclusions
Rice LTR retrotransposons are a significant component of the rice genome. We estimate that LTR retrotransposons constitute at least 17% of the O. sativa genome. Although this value is lower than the estimated percentage of LTR retrotransposons in the genomes of other cereal plants [2,12], it is more than tenfold greater than the estimated percentage of LTR retrotransposons in A. thaliana, a species with a genome one-third the size of the rice genome [21]. This disproportionate increase in the percentage of LTR retrotransposons as a function of genome size is consistent with the view that genome size variability in plants is often heavily dependent on variation in LTR retrotransposon content [27,28].

We have determined that individual full-length LTR-retrotransposons present in the sequenced euchromatic regions of the rice genome are all relatively young, displaying, on average, greater than 98% sequence identity between their LTRs. Comparative genomic studies of LTR retrotransposons in both plants and animals have revealed that species with smaller genomes [7,10,19-21] do not harbor older families of LTR retrotransposons, as do species with larger genomes [12,22]. It has been hypothesized that the rate of turnover of retroelements may be higher in small genomes as a result of the presence of less effective epigenetic silencing mechanisms [10]. It remains to be determined whether or not this hypothesis is an adequate explanation of the apparent lack of older full-length LTR retrotransposons in the euchromatic portion of the rice genome.

In general, the major clades of rice LTR retrotransposons are more closely related to elements present in other species than to the other clades of rice elements, suggesting that horizontal transfer may have occurred over the evolutionary history.
of rice LTR retrotransposons. Further analysis is required to definitively test the horizontal transfer hypothesis.

The newly developed search algorithm (LTR_STRUC) we have used in this study to initially identify LTR retrotransposons in the rice genome is not dependent upon sequence homology as are standard search methods such as BLAST. As a consequence, we identified several previously unreported families of rice LTR retrotransposons consisting of non-coding and, in some cases, repeating, sequence motifs. LTR retrotransposons of similar structure have recently been identified within the genomes of both monocotyledonous and
Automated characterization of LTR retrotransposons using LTR_STRUC

LTR_STRUC identifies new LTR retrotransposons on the basis of the presence of characteristic retroelement features (E.M.M. and J.F.M., unpublished work). It scans nucleotide sequence data for putative LTR pairs, aligns the putative pairs, and scores them on the basis of the presence/absence of expected motifs such as TSRS, canonical dinucleotides, PBS, PPT, and so on. When a given pair receives a score above a (user-specified) cut-off, an output record is generated that specifies salient information about the putative element, such as the length of the transposon and its LTRs, its position within the contig, an alignment of its LTRs, the nucleotide sequence of the transposon, its LTRs and target-site repeats, as well as a file listing all ORFs. In our study, once putative elements were identified, sequence analysis was carried out on the individual output files to identify those that described actual LTR retrotransposons. Additional elements were identified by BLAST searches using elements located by LTR_STRUC as queries.

Datasets scanned

Initial scans with LTR_STRUC were conducted on a dataset consisting of the 29.8 Mb of O. sativa BAC-derived sequence data available in GenBank at the time of the initial scan (December 2000). This dataset (TDS) was obtained from the TIGR website [30]. Subsequently, LTR_STRUC was used to scan the non-redundant MRD, a product of the Monsanto Rice Genome Sequencing Project. The MRD is based on an initial dataset of 3,391 BACs distributed across the genome of O. sativa cv. Nipponbare - the same cultivar used by the International Rice Genome Sequencing Project. Removal of contaminants and redundancies from this initial dataset produced the MRD (consisting of 52,202 contigs, totaling 259 Mb of the 430-Mb rice genome). More recently, in an effort to determine the relative copy numbers of the various families and identify additional elements not picked up in our initial survey with LTR_STRUC, we have used representative sequences from each retrotransposon family identified in this study as queries to conduct BLAST searches against both the MRD and the GBRD. Thus, the results reported here constitute a reasonably unbiased survey of LTR-retrotransposon diversity in rice. Both the MRD and GBRD are heavily weighted toward euchromatic sequences. The amount of data scanned was significantly less than the total amount of nucleotide sequence contained in the MRD and GBRD. Much of the MRD (~36%) is composed of contigs that are less than 10 kb long and are therefore of limited utility for the LTR_STRUC program, which finds only full-length elements (rice gypsy-like elements are typically longer than 10 kb and are not entirely contained in such short contigs). In the case of the GBRD, the amount of rice nucleotide sequence available for search was less than one-third of the 174 Mb released to the public (because of 15% redundancy, the GBRD sequences amounted to a total of only about 150 Mb, of which only some 50 Mb were actually available for BLAST search because most of these sequences were in the process of being 'finished'). RT sequences were identified according to previously described criteria [31,32].

Multiple sequence alignments and phylogenetic analyses

The RT domains of the Osr elements were aligned with previously reported RT sequences (Table 2). The ClustalW analysis [33] extension to MacVector 7.0 was used to generate two amino-acid alignments, one for gypsy-like, and one for copia-like elements. Draw N-J Tree and Bootstrap N-J commands of ClustalW were then used to generate non-boosted and boosted trees, respectively.

Acknowledgements

This work was supported by grant DBl-0077009 from the National Science Foundation. We thank Rebecca McCarthy for editorial assistance and Eric Ganko for constructive criticism.

References

1. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Benetzen JL: Identification and characterization of two tandem repeat sequences (TrsB and TrsC) and a retrotransposon (Rire1) as genome-general sequences in rice. Genes Genet Syst 1992, 76:373-382.
2. Bowen N, McDonald JF: Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 2000, 11:1527-1540.
3. Hirochika H, Fukuchi A, Kokuchi F: Retrotransposon families in rice. Mol Genet 1992, 233:209-216.
4. Nakajima R, Noma K, Ohtsubo H, Ohtsubo E: Identification and characterization of two tandem repeat sequences (TrsB and TrsC) and a retrotransposon (Rire1) as genome-general sequences in rice. Genes Genet Syst 1996, 71:373-382.
5. Bowen N, McDonald JF: Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res 1999, 9:924-935.
6. Burns DP, Temin, HM: High rates of frameshift mutations within hom-o-homologous runs during a single cycle of retroviral replication. J Virol 1994, 68:4196-4203.
7. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Benetzen JL: The paleontology of intergene retrotransposons of maize. Nat Genet 1998, 20:43-45.
8. Turchotte K, Sririvasan S, Bureau T: Survey of transposable elements from rice genomic sequences. Plant J 2001, 25:169-180.
9. Wang SP, Liu N, Peng KM, Zhang, QF: The distribution and copy number of copia-like retrotransposons in rice (Oryza sativa L.) and their implications in the organization and evolution of the rice genome. Proc Natl Acad Sci USA 1999, 96:6824-6828.
15. Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, et al.: Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 2000, 10:982-990.

16. Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A: The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted co-linearity with maize chromosome 4. Plant Cell 2000, 12:381-392.

17. Haslop-Harrison JS, Brandes A, Takata S, Schmidt T, Vershinin AV, Alkhimova EG, Karum A, Doudrick RL, Scwarzacher T, Katsiotis A, et al.: The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetics 1997, 100:197-204.

18. Boeke JD, Stoye JP: Retrotransposons, endogenous retroviruses and the evolution of retroviruses. In Retroviruses, edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997: 343-435.

19. Jordan IK, McDonald JF: Tempo and mode of evolution in Saccharomyces cerevisiae genome. Genetics 1999, 151:1341-1351.

20. Promislow DE, Jordan, IK, McDonald JF: Genomic demography: A life history analysis of transposable element evolution. Proc R Soc Lond B Biol Sci 1999, 266:1555-1560.

21. Kapanov VV, Jurka J: Molecular paleontology of transposable elements from Arabidopsis thaliana. Genetica 1999, 107:27-37.

22. Tristem M: Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 2000, 74:3713-3730.

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

24. Noma K, Nakajima R, Ohtsubo H, Ohtsubo E: Rire1, a retrotransposon from wild rice Oryza australiensis. Genes Genet Syst 1997, 72:131-40.

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

26. Li ZY, Chen SY, Zheng XW, Zhu LH: Identification and chromosomal localization of a transcriptionally active retrotransposon of Ty2-gypsy type in rice. Genome 2000, 43:404-408.

27. Kumar A, Bennetzen JL: Plant retrotransposons. Ann Rev Genet 1999, 33:497-532.

28. Wendel JF, Wessler SR: Retrotransposon-mediated genome evolution on a local ecological scale. Proc Natl Acad Sci USA 2000, 97:6250-6252.

29. Witte CP, Le QH, Bureau T, Kumar A: Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 2001, 98:13778-83.

30. The Institute for Genomic Research rice gene index [http://www.tigr.org/tdb/ogi/].

31. Xiong Y, Eickbush TH: Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 1988, 5:675-690.

32. Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse-transcriptase sequences. EMBO J 1990, 9:3353-3362.

33. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997, 25:4876-4882.