RIRE2, a novel gypsy-type retrotransposon from rice

Hisako Ohtsubo*, Norikazu Kumekawa, and Eiichi Ohtsubo
Institute of Molecular and Cellular Biosciences, the University of Tokyo,
Bunkyo-ku, Tokyo 113-0032, Japan

(Received 9 July 1999, accepted 20 August 1999)

The 441-bp DNA segment in a PCR-amplified fragment from Oryza sativa cv. IR36 was found to have a sequence with features characteristic of LTRs of retroelements, which was named RIRE2 (Rice retroelement #2) and further analyzed. Cloning and sequencing analyses of the DNA segments connected to LTR-like sequence showed that RIRE2 has a long internal region almost 10 kb long that is flanked by LTR-like sequences. This internal region carries a primer binding site (PBS) and polypurine tract (PPT) which are necessary for cDNA synthesis of retroelements. The PBS sequence is complementary to the 3' end region of tRNAArg. The internal region has an rt gene homologous to that of gypsy-type retrotransposons, evidence that RIRE2 is indeed a retrotransposon related to gypsy from Drosophila. RIRE2 has an extra sequence more than 4 kb long in the region downstream of gag-pol. Phylogenetic analysis of the putative amino-acid sequences of the rt gene as well as the int gene showed that RIRE2 is related to a group of gypsy-type retrotransposons of a large size that include Grande1-4 of teosinte, Tat4-1 and Athila1-1 of Arabidopsis thaliana, and Cyclops-2 of pea, but distantly related to any other group of gypsy-type retrotransposons, including RIRE3 and RIRE8 of rice. RIRE2 and Grande1-4 had the highest homology in the gag-pol region, but the nucleotide sequences of the LTR regions differed. Both elements had significant homology in the middle area of the extra regions downstream of gag-pol, in which they had an open reading frame encoding a protein with no known function on the opposite strand from that coding for gag-pol.

INTRODUCTION

Retroelements are mobile genetic entities that are ubiquitous components of eukaryote genomes. Among them are retrotransposons which are distinguished from other retroelements such as LINEs by their having long terminal repeats (LTR) like those observed in proviruses, the integrated forms of retroviruses. After transcription, they are reverse transcribed into cDNA by the reverse transcriptase-RNase H encoded by the elements then integrated into new sites in the chromosomes by integrase which also is encoded by the elements. Retrotransposons have been found in animals, plants, and fungi (Xiong and Eickbush, 1990; Flavel et al., 1994; Sandmeyer and Menees, 1996).

LTR-retrotransposons, like retroviruses, have the gag gene encoding a structural core protein and the pol region consisting of four genes (pro, rt, rh and int) that respectively encode the four catalytic enzymes, protease, reverse transcriptase, ribonuclease H and integrase. Most of the LTR-retrotransposons characterized have been classified into two groups based on their similarity to the Drosophila retrotransposons, copia or gypsy. These groups can be discriminated phylogenetically by comparison of the amino acid sequences of the four catalytic proteins and by the order of the genes in pol. Copia group elements have int in the region preceding rt and rh, whereas gypsy group elements have int at the distal region of pol (Xiong and Eickbush, 1990).

Retroviruses have an additional gene, env, that encodes proteins consisting of viral virions such as transmembrane and surface proteins (Varmus and Brown, 1989). Some retrotransposons related to gypsy (tom, 17.6 and 297 of Drosophila, Tas of nematode, and TED of vaculovirus) encode proteins that loosely resemble mammalian retroviral envelope (Env) proteins (Kim et al., 1994; Tanda et al., 1994; Song et al., 1994; Ozers and Friesen, 1996; Sandmeyer and Menees, 1996). In plants, several gypsy-type retrotransposons; Tat4-1 and Athila1-1 of A. thaliana (Wright and Voytas, 1998; Pelissier et al., 1995), Cyclops-2 of pea (Chavanne et al., 1998), and even a copia-type retrotransposon SIRE1 of soy bean (Laten et al., 1998), have recently been reported to have an orf(s) coding for a protein(s) that shares some structural similarities to Env.
proteins.

During the course of our work, a PCR-amplified fragment obtained from the total DNA of *Oryza sativa* L. cv. IR36 was found to be associated with tandem repeats of a sequence, called TrsA (Tandem repeat sequence A), located in the subtelomeric regions of the chromosomes of *O. sativa* (Ohtsubo and Ohtsubo, 1994; Zhao et al., 1989). A portion of this sequence (441-bp) appears to have structural features of the LTR of a retrotransposon, here called RIRE2 (Rice retroelement #2). We show that RIRE2 is a very large retrotransposon, about 11 kb long, which is flanked by a pair of LTR-like sequences. The internal region carries a primer binding site (PBS) and polypurine tract (PPT) sequences necessary for cDNA synthesis. The PBS sequence is complementary to tRNAArg and the internal region has homology to the gag-pol regions of gypsy-type retrotransposons. RIRE2 has a unique region, more than 4 kb long, downstream of the gag-pol region. Phylogenetic analysis based on the amino-acid sequence alignment of rt regions, shows that RIRE2 differs from the gypsy-type retrotransposons RIRE3 and RIRE8 previously identified in rice, but is related to the gypsy-type retrotransposons of large size from other plants.

**MATERIALS AND METHODS**

**Plant DNA.** Total DNA was isolated from the leaves of two-week-old seedlings of *Oryza sativa* L. cv. Nipponbare, *Oryza sativa* L. cv. IR36, *Oryza australiensis* W1538, *Oryza brachyantha* W1401, and *Oryza barthii* W1581 by the method described elsewhere (Ohtsubo et al., 1991).

**Southern and slot-blot hybridization.** Southern hybridization was done as described (Kumekawa et al., 1999). Slot-blot hybridization to estimate copy numbers was done as described (Motohashi et al., 1997). Copy numbers were estimated using the LTR DNA (positions 23 – 441 by coordinates given to RIRE2) carried by plasmid pBIN103 as the control. The probe used was amplified by PCR (polymerase chain reaction) with relevant primers using pBIN103 DNA as the template and labeled with \[\alpha\]32P]dCTP (Amersham, 185 TBq/nmol) using a Megaprime DNA Labeling System (Amersham, RPN1604). The free dCTP was removed by the spun-column of Sephadex G50 equilibrated in TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) containing 100 mM NaCl.

**PCR and cloning of PCR-amplified fragments.** The PCR was done with the relevant primers shown in Table 1. Either rTaq DNA polymerase or LA-Taq DNA polymerase (Takara) was used for the amplification reaction. Reaction mixtures of 100 µl contained 250 ng rice DNA, each dNTP at 20 nmol, 100 pmol each primers, 2.5 units of DNA polymerase in the buffer provided by the supplier of the enzyme. Thirty cycles of amplification were done under the following conditions; denaturation for 1 min at 94°C, annealing for 1 min at 50°C or 55°C (depending on the melting temperature of the primers), and DNA synthesis for 2 min at 72°C. A cloning kit for each vector was used to ligate PCR products to vector pCRII (Invitrogen) or pGEM-T Easy (Promega). Ligation products were transformed into *Escherichia coli* XLI-Blue MRF’ cells (Stratagene). Ampicillin-resistant transformants were selected on L-plates containing 100 µg/ml of ampicillin.

**Table 1. Primers used**

<table>
<thead>
<tr>
<th>Primers*</th>
<th>Sequence (5’-3’)</th>
<th>Positionb</th>
<th>Coordinatec</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>TACCCAGTCCAAATACCATCG</td>
<td>LTR</td>
<td>54–34, 10917–10897</td>
</tr>
<tr>
<td>P2</td>
<td>TTCCCACTTGATAGCCACC</td>
<td>LTR</td>
<td>389–408, 11252–11271</td>
</tr>
<tr>
<td>P3</td>
<td>AGGATCCATCGCAGGAG</td>
<td>gag</td>
<td>1558–1575</td>
</tr>
<tr>
<td>P4</td>
<td>GGAGGTGAAGTGAAGCAG</td>
<td>gag</td>
<td>1611–1594</td>
</tr>
<tr>
<td>P5</td>
<td>GAAGTCACCGATTTCGAG</td>
<td>pro</td>
<td>3071–3088</td>
</tr>
<tr>
<td>P6</td>
<td>CGCAAACGTCGAGGAGACG</td>
<td>pro</td>
<td>2992–3011</td>
</tr>
<tr>
<td>P7</td>
<td>AAGTCCCAAGCTTAGCTGAG</td>
<td>rt</td>
<td>4799–4817</td>
</tr>
<tr>
<td>P8</td>
<td>GGGCTCTCCAGGCTGCCCTG</td>
<td>rh</td>
<td>5381–5362</td>
</tr>
<tr>
<td>P9</td>
<td>CAACCTCTGGAGACATTGAGA</td>
<td>int</td>
<td>6562–6581</td>
</tr>
<tr>
<td>P10</td>
<td>GAATGAACAAATGCCAGGACGCT</td>
<td>orf’</td>
<td>6916–6895</td>
</tr>
<tr>
<td>P11</td>
<td>GATGCTAAGCAGAAGGACC</td>
<td>orf’</td>
<td>8950–8931</td>
</tr>
<tr>
<td>P12</td>
<td>ATACCTGCTGCTCTGCC</td>
<td>orf’</td>
<td>9315–9298</td>
</tr>
</tbody>
</table>

* Primers used for PCR are listed, but those for nucleotide sequencing are not. The approximate positions of these primers are shown in Fig. 3.

b LTR, long terminal repeat; pro, the protease gene; gag, the gag gene; rt, the reverse transcriptase gene; rh, the RNase H gene, int, the integrase gene; orf’, the region downstream of pol of RIRE2.

c Coordinates to the sequence of RIRE2, starting with the T residue as the first base at the 5’-end of the 5’-LTR.
A novel gypsy-type retrotransposon RIRE2 from rice

DNA sequencing. DNA sequencing was done by dideoxynucleotide chain termination (Messing, 1983; Sanger et al., 1977) using dye-labeled primers (–21M13 forward, M13 reverse) and an ABI PRISM Dye Primer Cycle Sequencing Ready Reaction kit with AmpliTaq DNA polymerase FS (Perkin Elmer) or a Dye Deoxy Terminator Cycle Sequencing kit with AmpliTaq DNA polymerase FS (Perkin Elmer) and an oligodeoxynucleotide primer hybridizing to a segment of RIRE2. Plasmid DNA was heat denatured and used as the template for the sequencing reaction. The consensus sequence was deduced from the nucleotide sequences of at least three independent clones.

Computer-aided sequence analysis. Sequence analysis was done with the programs, Harr plot 1.2.2 and Genetyx-Mac 7.3 system (Software Development), the BLAST and FASTA search tools (Pearson and Lipman, 1988; Karlin and Altschul, 1990; 1993), and MP search (Smith and Waterman, 1981). Multiple sequences were aligned by using the program CLUSTAL W ver. 1.7, and a dendrogram of the sequences was constructed with the software PHYLIP ver. 3.572 (Thompson et al., 1994; Felsenstein, 1993).

Accession numbers. All new DNA sequence data have been deposited in the DDBJ (DNA Data Bank of Japan) under the following accession numbers, AB030283 (RIRE2) and AB032021 (RIRE2-A26).

RESULTS AND DISCUSSION

RIRE2, a retrotransposon with a large internal region. The 441-bp sequence, associated with the tandem repeat sequence TrsA, is shown in Fig. 1. This sequence starts at TG and ends with CA (Fig. 1), indicative of the long terminal repeat (LTR) sequences of retroviruses and retrotransposons. The LTR sequences of retroviruses are made up of three regions, U3, R and U5 (Temin, 1981). The U3 region has a TATA-box, the promoter for transcription, and an enhancer with repeats of a 10- to 20-bp sequence. The R region has a transcription initiation site, a capping signal sequence to protect the mRNA transcribed, and a poly A addition signal. The U5 region usually carries a stem-loop structure in the end region, which is thought to be a termination site for mRNA synthesis. The 441-bp sequence appears to have these three regions (see Fig. 1).

Southern hybridization was done to determine whether the 441-bp sequence is repeated in genomes of O. sativa cv. Nipponbare and O. sativa cv. IR36. Total DNAs from both strains produced many restriction fragments that hybridized with the probe for the 441-bp sequence (Fig. 2), indicative that the sequence is present in many copies in the O. sativa genomes. Dot blot hybridization also was done to determine how many copies of the 441-bp sequence are present in rice genomes. O. sativa cv. IR36 had this sequence in about 10,000 copies per haploid genome, O. australiensis in 2,600 copies per haploid genome, and O.

<table>
<thead>
<tr>
<th>U5</th>
<th>U3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAATCTTTTGTCTCACAACCAACATACACATTCCAGCTAGCCACCCCTTTTATAAT</td>
<td>ACCTAATGTAGATAGATTACACCGCCGATGGATTTTTTATTGCCTTGACCG</td>
</tr>
<tr>
<td>TGCCGAAATCTAGTTTCGACA</td>
<td>TAGGGGATATAGATATACCTTTCTGTACCTCGGGTACCTTACAG</td>
</tr>
</tbody>
</table>

Fig. 1. The nucleotide sequence of LTR of RIRE2 that starts at TG and ends with CA. Putative U3 and U5 regions are boxed. The R region starts at G and ends with CA between the U3 and U5 regions. Possible promoter elements in the U3 region and a possible poly A additional signal in the R region are shown by boldface letters. The U5 region carries the sequence TTGA, necessary for transcription termination, and a stem-loop structure of 20 bp close to the end region. P1 and P2, shown by thick arrows, were the primers used for PCR (see Table 1).
Fig. 2. Southern blot-hybridization analysis of rice DNA. Total DNA prepared from *O. sativa* cv. Nipponbare (lanes 1–3) or *O. sativa* cv. IR36 (lanes 4–6) was digested with restriction endonucleases, *Hind*III (lanes 1 and 4), *Eco*RI (lanes 2 and 5), or *Bam*HI (lanes 3 and 6), and electrophoresed in a 0.7% agarose gel.

Fig. 3. (A) Schematic representation of the structure of *RIRE2*. Open boxes show the LTRs and hatched boxes the five orfs encoded by each of the two DNA strands. (B) Clones carrying various portions of *RIRE2*. Thick lines indicate the portions of *RIRE2* carried. Dotted lines indicate deleted portions in *RIRE2*. Clones A26 and A55 were obtained by genomic PCR with primers P1 and P2 (large arrowheads). Other clones were obtained by PCR using primers P3–P12 (small arrowheads) (see Table 1).
nial region of RIRE2 (Fig. 3B). Several other clones obtained by PCR and relevant primers which hybridize to the nucleotide sequence of A55 had various portions of the sequence in clone A55 (see Fig. 3B, J101, F1'-4, J4, F8'-1, etc.). The consensus sequence of the internal region of RIRE2 deduced from the nucleotide sequences of all the clones was 10304 bp long.

**Open reading frames in RIRE2.** The RIRE2 sequence had two long open reading frames (orf1 and orf2) in the gag-pol region with the genes gag, pro, rt, rh and int in this order (see Fig. 3). The gene order is the same as that of the gypsy-type retrotransposons and retroviruses, showing that RIRE2 is indeed a gypsy-type retrotransposon.

The sequence of RIRE2 analyzed above is a consensus sequence derived from fragments of the cloned. It should be noted that the nucleotide sequence of each cloned fragment differed from that of the consensus sequences, by about 2 to 3 bp in every 100 bp, at a higher frequency than that for base substitutions (0.16%) induced by PCR. This means that these fragments did not derive from the one RIRE2 copy, rather they originated from divergent copies of RIRE2 present in the rice genome. The identified mutations altered codons, but 98% resulted in synonymous codons. The consensus sequence obtained therefore may be close to that of the active element, as discussed previously in the identification and characterization of the RIRE3 and RIRE8 elements (Kumekawa et al., 1999).

In the consensus sequence, however, one TAG termination codon existed between orf1 and orf2 at the position 3440–3442 by the coordinates to 3A. The TAG codon was present in all the clones carrying the region between orf1 and orf2. The termination codon TAG is present in the pol region of several retroviruses, but the Gag-Pol polyprotein may be synthesized upon suppression by suppressor tRNA (Yoshinaka et al., 1985a; 1985b). RIRE2 therefore may use such a suppression mechanism for the synthesis of its Gag-Pol polyprotein. In the retroviruses MLV, RSV, Mo-MuLV, SNV and BaEV, the TAG codon is present at the junction site of the gag and pol regions (Varmus and Brown, 1989; Levin et al., 1993). RIRE2 however had the TAG codon in the middle of the region coding for the protease gene. In the virus systems, a pseudoknot structure is present in the region downstream of TAG, but in RIRE2, there was no such structure.

RIRE2 had an extra region more than 4 kb long in the region downstream of gag-pol (see Fig. 3A). This 4-kb region encoded no orfs on the strand coding for orf1 and orf2, but did encode three orfs (orf3’, orf4’ and orf5’) on the opposite strand (see Fig. 3A). The significance of these three frames is discussed later.

**Phylogenetic relationship of RIRE2 and other gypsy-type retrotransposons.** A phylogenetic tree was constructed based on the alignment of the amino-acid sequences encoded by the rt regions of various retrotransposons and retroviruses (Fig. 4). RIRE2 clustered with a group of large elements, mostly those of more than
10 kb, including Grande1-4 of teosinte, Tat4-1 and Athila1-1 of A. thaliana and Cyclops-2 of pea (see Plant I group in Fig. 4; see also Fig. 5), which respectively have homology (64.9, 51.1, 35.1 and 33.3% in the amino-acid identity) with the rt region of RIRE2 (positions 3634–4435 by coordinates given to RIRE2). This group is distinct from the other group of plant retrotransposons that includes the previously identified rice retrotransposons RIRE3 and RIRE8 (see Plant II group in Fig. 4), which respectively have homology (33.1 and 31.2%) with RIRE2, and from those of fungi, insects and retroviruses (Fig. 4), which have homology (mostly less than 30.0%) with RIRE2. RIRE2 was most closely related to Grande1-4, and next to Tat4-1, but less closely related to Athila1-1, Cyclops-2 and a retrotransposon-like sequence of broad bean (here called BBRE1) in the same group (see Fig. 4). The phylogenetic tree based on amino-acid sequences of the zinc finger motif regions of the integrases was similar to that based on the amino-acid sequences of the reverse transcriptase (data not shown). The rice gypsy-type retrotransposons were resolved into two lineages as were those from A. thaliana (Wright and Voytas, 1998).

Dot matrix analysis results of the nucleotide sequences between RIRE2 and each of the other four plant retrotransposons (Grande1-4, Tat4-1, Athila1-1 and RIRE3) whose entire sequences with LTRs have been determined, are shown in Fig. 6. Clearly, RIRE2 has significant homology with Grande1-4 and Tat4-1, especially in the pol region, whereas it has poor homology with Athila1-1 (Fig. 6) and Cyclops-2 (data not shown). RIRE2 was homologous to RIRE3 only for rt and int in the pol region (Fig. 6d). Interestingly, RIRE2 and Grande1-4 were homologous in the middle area of the region downstream of pol, whereas RIRE2 and Tat4-1 were not (Fig. 6).

**Characteristic features of RIRE2-related retrotransposons.** As shown above, RIRE2, Grande1-4 and Tat4-1 are very closely related and are distinct from Athila1-1 and Cyclops-2 (see Fig. 4). The former elements have fairly short LTRs (< 600 bp), whereas the latter ones have long LTRs (> 1 kb) (see Fig. 5). The rice elements of RIRE3 and RIRE8, which belong to Plant II group (Fig. 4), have respectively very long LTRs of 2316 and 2948 bp, the largest LTRs so far identified (Fig. 5; Kumekawa et al., 1999).
A novel gypsy-type retrotransposon RIRE2 from rice

LTRs of RIRE2, Grande1-4 and Tat4-1, which have homology in the pol region, are very poorly conserved in the nucleotide sequence (see Fig. 6). These elements however have the PBS sequence that is homologous to tRNAArg (Fig. 5). The PBS sequences of Athila1-1 and Cyclops-2 are homologous to tRNAArg (Fig. 5), evidence that the two groups of elements also can be differentiated by their PBS sequences. The PBS sequence of Plant II group members, including RIRE3 and RIRE8, is homologous to tRNAAsp (Fig. 5), showing that the elements in this group can be differentiated from those of larger size.

**Additional sequences in the region downstream of gag-pol in RIRE2.** RIRE2 and Grande1-4, which had extensive homology in the gag-pol region, also had homology in the middle area of the region downstream of gag-pol (Fig. 6a). RIRE2 had an orf (orf3'; see Fig. 1A) on the opposite strand to that coding for Gag-Pol. Grande1-4 had an orf (called orf24) on the opposite strand to that coding for Gag-Pol (Martinez-Izquierdo et al., 1997), whose nucleotide sequence showed significant homology to that of orf3' in RIRE2. These orfs also had significant homology (36.7% in the identity) to each other at the amino-acid sequence level (Fig. 7). This suggests that they encode products with some function unknown at present. If so, there may be an additional promoter(s) for the transcription of orf3' which differs from the promoter present in LTR. Note that Grande1-4 has two regions of tandem repeat sequences, tandem A and B, and that integration of the region flanked by these tandem repeats from the teosinte genome into the original Grande1 element is assumed to have lengthened the internal region (Martinez-Izquierdo et al. 1997). RIRE2, however, has no such repeat sequences in its internal region, indicative that tandem repeats are not involved in lengthening RIRE2.

Grande1-4 is supposed to have an orf, called orf23, that encodes a product with the nuclear transport signal, in the region downstream of pol on the opposite strand to that coding for Gag-Pol (Martinez-Izquierdo et al., personal communication). The orfs (orf3', orf4' and orf5') in RIRE2, however, appear not to encode any products with this signal. Cyclops-2 is a large retrotransposon (12.3 kb in length) with an orf (orf423) encoding a 423 amino-acid polypeptide, but it has no sequence homology with the Env proteins nor those from known orfs in the databases (Pelissier et al., 1995). The orfs in RIRE2 appear not to encode any products with homology to this orf product. Also, none of the transmembrane motifs found in the products from orfs of Athila1-1 and Tat4-1 (Rost et al., 1995; Wright and Voytas, 1998) were detected in products from orf3', orf4', and orf5' in RIRE2.

This work was supported by a Grant-in-Aid for Scientific Research from Ministry of Education, Science, Sports and Cul-
ture of Japan, and by a grant for the Development of Methods of Rice Genome Analysis and Use of Accurate Genome Map of Rice from the Ministry of Agriculture, Forestry and Fisheries of Japan.

REFERENCES


A novel gypsy-type retrotransposon RIRE2 from rice


