ABSTRACT  Although most of non-long terminal repeat (non-LTR) retrotransposons are incorporated in the host genome almost randomly, some non-LTR retrotransposons are incorporated into specific sequences within a target site. On the basis of structural and phylogenetic features, non-LTR retrotransposons are classified into two large groups, restriction enzyme-like endonuclease (RLE)-encoding elements and apurinic/apyrimidinic endonuclease (APE)-encoding elements. All clades of RLE-encoding non-LTR retrotransposons include site-specific elements. However, only two of more than 20 APE-encoding clades, Tx1 and R1, contain site-specific non-LTR elements. Site-specific non-LTR retrotransposons usually target within multi-copy RNA genes, such as rRNA gene (rDNA) clusters, or repetitive genomic sequences, such as telomeric repeats; this behavior may be a symbiotic strategy to reduce the damage to the host genome. Site- and sequence-specificity are variable even among closely related non-LTR elements and appeared to have changed during evolution. In the APE-encoding elements, the primary determinant of the sequence-specific integration is APE itself, which nicks one strand of the target DNA during the initiation of target primed reverse transcription (TPRT). However, other factors, such as interaction between mRNA and the target DNA, and access to the target region in the nuclei also affect the sequence-specificity. In contrast, in the RLE-encoding elements, DNA-binding motifs appear to affect their sequence-specificity, rather than the RLE domain itself. Highly specific integration properties of these site-specific non-LTR elements make them ideal alternative tools for sequence-specific gene delivery, particularly for therapeutic purposes in human diseases.

OVERVIEW OF NON-LTR RETROTRANSPOSONS

DNA transposons are the mobile elements that move by a “cut and paste” mechanism (1, 2). In contrast, retrotransposons encode reverse transcriptase, and move by a “copy and paste” mechanism. The process of retrotransposon insertion into genomic locations involves an RNA intermediate. Retrotransposons can be classified into long terminal repeat (LTR) and non-LTR retrotransposons. LTR retrotransposons have LTRs at both ends and resemble retroviruses in both structure and integration mechanisms. Non-LTR retrotransposons comprise two subtypes, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Non-LTR retrotransposons are in general 4 to 7 kb long and do not carry LTRs, and their retrotransposition mechanism is different from that of LTR retrotransposons. SINEs are nonautonomous retrotransposons of 100 to 500 bp that do not encode proteins. It has been proposed that the proteins encoded by LINEs are the source of the enzymatic retrotransposition machinery of SINEs (3, 4, 5).

The process of retrotransposition of non-LTR elements is initiated by an encoded endonuclease domain that nicks one strand of DNA at the target site and creates a 3′-hydroxyl end, which is used as a primer for reverse transcription of the retrotransposon mRNA onto the DNA target (6, 7). This unique process, target primed reverse transcription (TPRT), is peculiar to non-LTR retrotransposons (6). These retrotransposons have been classified into two large groups on the basis of...
their structural and phylogenetic features (8, 9). One of the groups encodes a restriction enzyme-like endonuclease (RLE) in the C-terminal region of the single open-reading frame (ORF) (10). The RLE-encoding non-LTR elements are a phylogenetically ancient class and are further categorized into five clades. The other group usually encodes an endonuclease with homology to apurinic/apyrimidinic endonuclease (APE) in one of its two ORFs (11). This group is thought to be younger and is categorized into at least 22 clades (8, 9). One clade (Dualen/RandI), which encodes both APE and RLE, is positioned phylogenetically at the midpoint between the early branched and recently branched groups (9, 12).

Retrotransposons are widespread in metazoan genomes. Recent and large-scale genome projects have revealed that the proportion of the genome that contains retrotransposons often exceeds that containing DNA transposons, particularly in higher vertebrates such as humans (13). Non-LTR retrotransposon L1 integrates essentially throughout the whole genome and occupies more than 16% of the human genome (14). The integration of L1 has been shown to be involved in genetic diseases and cancers and affects genome reconstitution and gene evolution (14, 15). Most retrotransposons integrate into random sites of the host genome, but some have a sequence preference. In particular, a few non-LTR subclades integrate into the genome in a highly sequence-specific manner.

In this chapter, I summarize site-specific non-LTR retrotransposons, focusing on the mechanisms of their integration.

**OVERVIEW OF SITE-SPECIFIC NON-LTR RETROTRANSPOSONS**

Some non-LTR retrotransposons have very restricted integration targets within the genome. Targets of site-specific non-LTR retrotransposons are usually located within multicopy RNA genes, such as clusters of rRNA gene (rDNA) or repetitive genomic sequences (16). Targeted integration of site-specific elements into repetitive sequences is considered to be a symbiotic strategy, which allows spreading through the genome with very little damage to essential genes of the host (Fig. 1). In contrast to this strategy, most transposable elements integrate in a parasitic-like, random manner (Fig. 1). The copy number of target sequences is considered to be a primary factor in restricting various sequence specificity of non-LTR elements (17).

How are site-specific non-LTR elements found in the genome? We could find both the 5’ and 3’ ends of

---

**FIGURE 1** Two types of genomic insertion of non-LTR elements. doi:10.1128/microbiolspec.MDNA3-0001-2014.f1

A. Randomly retrotransposed non-LTR elements

- **Parasitic**
  - L1, L2
  - Chromosome
  - Telomeric repeats
  - rDNA
  - Microsatellite
  - Telomeric repeats

B. Site-specific non-LTR elements

- **Symbiotic**
  - SART, TRAS
  - R1, R2
  - Dong
  - Chromosome
  - Telomeric repeats
  - rDNA
  - Microsatellite
  - Telomeric repeats
genomic copies of non-LTR retrotransposons because they usually end with a poly (A) stretch and have target-site duplication (TSD) at both ends. Thus, it is possible to search for the site-specific non-LTR elements in DNA databases by investigating the boundary sequences flanking multiple genomic copies of the non-LTR elements. If all the junction sequences of the multiple copies are the same, they are accepted as “site-specific non-LTR retrotransposons.” However, the 5′ portion of non-LTR elements is often deleted by incomplete reverse transcription (5′ truncation), which means that the same genomic repetitive sequences flank copies of target-specific non-LTR elements 5′ truncated at different positions. Many new examples of site-specific non-LTR elements have been found from sequenced genomes (16, 17).

There are five clades of RLE-encoding elements based on their RT sequence similarity. Most of three clades (NeSL, R2, and R4) are target specific and two clades (HERO and CRE) have some site-specific elements (Fig. 2 and Table 1). CRE1/CRE2/SLACS/CZAR in trypanosomes (CRE clade) and NeSL-1 in nematodes (NeSL clade) were found in the spliced leader exons (18, 19, 20). R2 (R2 clade) was found at a specific sequence in the 28S rDNA of many invertebrate and vertebrate species (21, 22, 23) (Fig. 3A). R4 in Ascaris and Dong (R4 clade) were found at another site of the 28S rDNA (Fig. 3A) and microsatellite TAA repeats, respectively (24, 25) (Table 1). Apart from these already characterized site-specific elements, some novel site-specific RLE-encoding elements have been found in silico using BLAST searches of databases or degenerate PCR cloning (17). HERO-1_HR (HERO clade) was recently found in a microsatellite, (ATT)n repeats (26) (Table 1). Table 1 lists site-specific non-LTR elements from various clades and their targets.

In contrast to RLE-encoding elements, most of the APE-encoding non-LTR elements do not insert themselves in a sequence-specific manner, but do have weak target-site specificity, e.g., human L1 for TAAA repeats (30, 31, 32). APE is considered to be closely related to class II of the AP endonuclease family of highly conserved multifunctional DNA repair enzymes (33, 34).
Among over 20 clades of APE-encoding non-LTR elements, members of only two clades, Tx1 and R1, are known to be sequence-specific. The Tx1 clade, which has many types of site-specific non-LTR elements, targets rRNA genes (Mutsu), tRNA genes (Dewa), snRNA genes (Keno), telomeric repeats (Tx1-1_ACar) (Fig. 3B), and microsatellites (Kibi and Koshi) (17, 35) (Table 1). Tx1L and Tx2L insert into other transposons, Tx1D and Tx2D, respectively (36). The R1 clade also contains several site-specific elements. Extensive screening of novel R1 clade elements found different targets for sequence-specific integration: different locations of the rRNA genes (e.g., R1) (Fig. 3A), telomeric repeats (e.g., TRAS) (Fig. 3B), and microsatellite (e.g., Waldo) (16). To the best of our knowledge, the sequence-specific R1 clade elements target only

![TABLE 1 Targets of site-specific non-LTR retrotransposons](ASMscience.org/MicrobiolSpectrum)

<table>
<thead>
<tr>
<th>Endonuclease</th>
<th>APE</th>
<th>Clade</th>
<th>RLE</th>
<th>CRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rRNA gene</td>
<td>R1</td>
<td>R2</td>
<td>NeSL</td>
<td>R4</td>
</tr>
<tr>
<td>tRNA gene</td>
<td></td>
<td></td>
<td></td>
<td>R4-2_Sra*</td>
</tr>
<tr>
<td>snRNA gene</td>
<td></td>
<td></td>
<td></td>
<td>CRE</td>
</tr>
<tr>
<td>Spliced leader</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telomeric repeat</td>
<td>TRAS/SART</td>
<td>Tx1-1_Acar</td>
<td>NeSL</td>
<td>MoTeR</td>
</tr>
<tr>
<td>Microsatellite</td>
<td>Waldo</td>
<td>Kibi</td>
<td>HERO-1_HR</td>
<td>Dong</td>
</tr>
<tr>
<td>Transposon</td>
<td></td>
<td></td>
<td>R5-2_SM*</td>
<td>CRE-1_NV*</td>
</tr>
</tbody>
</table>

Only representative elements are shown.
APE: endonuclease with homology to apurinic/apyrimidinic-endonuclease, RLE: restriction enzyme-like endonuclease
* : References are shown in the reference list 27–29. **: Mutsu targets 5S rRNA genes.

![FIGURE 3 rDNA-specific-elements (R-elements) and telomere-specific-elements](ASMscience.org/MicrobiolSpectrum)

A. R-elements (rDNA-specific elements)

B. Telomere-specific elements
three genomic locations: rDNA, telomeric repeats, and ACAY (or AC) repeats (Fig. 3 and Table 1). Phylogenetic trees for R1 clade elements reveal that the target specificity has been independently altered several times during evolution (16) (Fig. 4). We note that some elements (Hida, Noto, HOPE, Kaga, and Hal) of this “site-specific clade” have lost their specificity of integration.

Dr. Eickbush’s chapter describes in detail the retrotransposition mechanisms of R2 and the related site-specific RLE-encoding non-LTR elements. In this chapter, I describe the retrotransposition mechanism of mainly site-specific APE-encoding retrotransposons.

**FIGURE 4** Evolution of target sequences in R1-clade elements. The phylogeny is constructed on the basis of data in K. K. Kojima and H. Fujiwara (16). Nonsequence-specific elements are shown as dotted lines and asterisks. Target and flanking sequences of each site-specific element are shown on the right. In the TPRT model, the bottom strand is first nicked and then the top strand is cleaved. The SART and TRAS elements target the same (TTAGG/CCATT)n, telomeric repeats. However, TRAS first cut the TTAGG bottom strand, whereas SART first cut the CCTAA strand. The bottom-to-top strand cleavage in each target sequence is represented by bent lines. Arrows indicate variation of top-strand cleavage. Broken lines between the top and bottom strands, in TRAS, SART, Waldo, and Mino, indicate unidentified exact cleavage sites because they target the tandem repeats TTAGG and ACAY (or AC). Broken boxes indicate homologous sequences near cleavage sites among RT and R7. [doi:10.1128/microbiolspec.MDNA3-0001-2014.f4](http://doi.org/10.1128/microbiolspec.MDNA3-0001-2014.f4)

**TELOMERE-SPECIFIC NON-LTR ELEMENTS**

Telomeres are the regions at the ends of chromosomes that are essential for the complete replication, meiotic paring, and stability of chromosomes (37). It has been hypothesized that the terminal sequences of chromosome ends shorten with each cell division, and a reduction in telomere length causes cellular senescence associated with the cessation of cell division. The telomeres of most eukaryotic cells are composed of simple repeated sequences called telomeric repeats. One of the strands is G-rich and is synthesized by a specialized enzyme, telomerase (37). In many organisms, the addition of telomeric repeats by
telomerase is essential to compensate for the critical telo-
mere shortening that occurs with aging (37, 38).

**Drosophila** telomere-specific elements

However, alternative telomere maintenance has been
suggested for some insects (39, 40). All dipteran insects
have lost the telomeric repeats; it has been suggested that
they use a telomerase-independent telomere mainte-
nance system (41, 42, 43, 44, 45). In *D. melanogaster*,
telomeres are composed of three specialized non-LTR
retrotransposons, TART, HeT-A, and TAHRE and main-
tained mainly by retrotransposition of these elements,
which have been extensively studied to date (41, 42, 43,
46, 47, 48, 49, 50, 51, 52) (Fig. 3B). Phylogenetically,
the three telomere-specific elements belong to the APE-
encoding Jockey clade, but HeT-A has no RT (pol) do-
main. It is noteworthy that the retrotransposition of
these elements does not depend on a specific DNA at the
target site. Although the mechanism as to how they tar-
get chromosome ends is not understood clearly, a recent
report showed that the ORF1p of HeT-A localizes in the
nucleus and forms spherical structures at telomeres,
which may support targeting to the chromosome ends
(52). In *Drosophila*, repeated but random retrotrans-
position of the three elements with their 5′ ends toward
the chromosomes form long rows of elements.

Distantly related *Drosophila* species, *D. yakuba*, *D. viri-
lis*, and *D. melanogaster*, all have this robust mechanism
of telomere maintenance, suggesting establishment be-
fore the divergence of the *Drosophila* genus (48).

**Telomere-specific non-LTR elements in**
**Bombyx and Tribolium**

In contrast to *D. melanogaster*, the silkworm *B. mori*
retains the telomeric (TTAGG)$_n$ repeats at the ends of
chromosomes (39, 53). Immediately proximal to a 6 to
8-kb stretch of TTAGG repeats, there are more than
1,000 copies of non-LTR retrotransposons, which are
designated the TRAS and SART families of the R1
clade (54, 55, 56, 57) (Fig. 3B and 4). They insert into
both strands of telomeric repeats so these elements are
inserting in different orientations (i.e., CCTAA is the
reverse complement of TTAGG) (Fig. 4). TRAS1, a
predominant element in the TRAS family, is 7.8 kb in
length and is inserted specifically into the site between
the C and T of the CCTAA telomeric strand (58, 59). A
major element in the silkworm SART family, SART1
(SART1Bm), is 6.7 kb in length and is inserted spe-
cifically between the T and A of the TTAGG strand
(60, 61). The TRAS and SART families occupy 3% of
the silkworm genome and more than 300 kb of each
of the chromosome ends (39). These telomere repeat-
specific non-LTR families are also found in other insects
(40, 61, 62) and in arthropods (K. K. Kojima, per-
sonal communication). The TRAS and SART elements
are transcribed abundantly in comparison with other
non-LTR elements (63). The translational regulation of
bicistronic (ORF1 and ORF2) mRNA of SART1 is analo-
gous to translational coupling in prokaryotes and vi-
ruses (64). Although most insects retain the TTAGG
telomeric repeats (except for *Diptera*), some have differ-
etent pentanucleotide telomeric repeats, such as TCAGG
in the flour beetle *T. castaneum* (65) (Fig. 5). Recently,
seven types of non-LTR retrotransposons of the R1 clade
were found in the TCAGG telomeric repeats of *T. cas-
taneum* (61). Based on their amino acid sequences, these
elements constitute a monophyletic group. They are phy-
genetically closer to SART1Bm than to TRAS1Bm, and
named SARTTc1 to SARTTc7 (SARTTc family). All of
these TCAGG-repeat-specific non-LTR elements are in-
serted between C and A of TCAGG telomeric repeats
and the orientation of the insertion is the same as that
observed for the SART1Bm family (61).

To investigate the functional roles of these telomere-
specific non-LTR elements, we measured the telomerase
activity in various insects using a modified TRAP (te-
lomeric repeat amplification protocol) method (66).
Insects retaining standard telomeric repeats undisturbed
by retrotransposons, such as cockroaches or crickets,
showed significant telomerase activities in several tis-
ues. However, the telomerase activity could not be de-
tected in two cultured cell types of *D. melanogaster*,
or in three cell lines and tissues from *B. mori*. Recently,
we also found by TRAP analysis that *Tribolium* shows
only weak telomerase activity in the elongation step
(67; M. Osanai-Futahashi and H. Fujiwara, unpublished
data). These results suggest that telomere-specific non-
LTR elements rescue the telomere crisis in *D. melano-
gaster*, and provide a backup for the telomere function
in *B. mori* and *T. castaneum*, which means that *B. mori*
does not need as much telomerase activity.

We analyzed the gene structure of *B. mori* (Lepidop-
tera) and *T. castaneum* (Coleoptera) TERT (telomerase
reverse transcriptase) and found some unusual shared
characteristics compared to the TERTs of other insects
without TRAS and SART elements: (i) an intronless gene
(Fig. 5), (ii) 3′-poly A tails in the genomic copy (Fig. 6B),
(iii) upstream ATG codons in the 5′ UTR (five ATGs for
*BmoTERT* and four ATGs for *TcTERT*) (Fig. 6A),
(iv) loss of N-terminal GQ motifs (Fig. 5 and 6), and
(v) repressed transcription (40, 65). It is reported that
multiple upstream ATGs (or upper AUG) reduce the translation of some genes (68). In addition, N-terminal GQ motifs are shown to be involved in the telomeric repeat processivity in yeast and human TERT (69, 70). Thus, the above features may cause inefficient elongation of their telomeric repeats. Hymenoptera (such as the honey bee Apis mellifera), an ancestral group among higher insects (71, 72), have an unusual TERT structure with many introns and long telomeric repeats undisturbed by retrotransposons (Fig. 5) (73). This suggests that TERT mRNA has been reverse-transcribed in a common ancestor of Coleoptera and Lepidoptera after branching from Hymenoptera, and integrated into the genome as a processed gene. Owing to 5′ truncation in the process of reverse transcription, the N-terminal including the GQ motif was deleted, and this unusual gene replaced the normal TERT gene (Fig. 6C). The original TERT genes with introns could not be found in the silkworm and Tribolium genome databases, and we do not yet know how these genes were lost. The telomeric repeat-specific non-LTR retrotransposons may have evolved to backup the weak activity of telomerase in Lepidoptera and Coleoptera, which helped them to survive weak telomerase activity during insect evolution (40).

### Telomere-associated non-LTR elements in other organisms

Telomere-associated non-LTR retrotransposons have also been found in organisms other than insects (47) (Fig. 3B and Table 1). Zepp is an L1-clade non-LTR element found in the telomeres of all 16 chromosomes in Chlorella vulgaris; this element exhibits telomere region specificity but not sequence specificity (74). In the genome of Giardia lamblia, two RLE-encoding non-LTR elements capped by TAGGG repeats, GilM and GilT, are located at the ends of the chromosome (75). Recently, two related RLE-encoding non-LTR retrotransposons, MoTeR elements (clade CRE), were found in the telomeric repeat (TTAGGG)ₙ of the fungus Magnaporthe oryzae (76). In the genome of the green anole lizard, an APE-encoding element Tx1-1_ACar (clade Tx1) inserted in the telomeric repeats (TTAGGG)ₙ has been found recently (Table 1) (35).
As mentioned above, many sequence-specific non-LTR elements accumulate in and target the ribosomal RNA (28S, 18S, and 5.8S rRNAs) gene array (rDNA). These elements are called R-elements. Thus far, nine different R elements inserted into various sites of rDNA have been reported (Fig. 3A and Table 1). R7 in R1 clade (16) and R8 in R2 clade (77) insert into 18S rDNA, while the remaining seven R elements, R1 (78, 79), R2 (7, 78), R4 (24), R5 (80), R6 (16), R9 (81), and RT (82), are inserted into the 28S rDNA. It is of interest that all R-element target sites within 28S and 18S rDNA are highly conserved among organisms.

The APE-encoding elements R1, R6, R7, and RT, which are structurally related and classified in the R1 clade, are believed to share a common sequence-specific ancestor (Fig. 4) (16). In phylogenetic trees based on amino acid sequences of the RT and endonuclease domains, these elements and telomeric repeat-specific TRAS and SART are closely related (Fig. 4) (16, 61). The tree also shows that R1, TRAS, SART, and Waldo (ACAY-specific element) are ancient retrotransposons that branched at a very early stage. In fact, these elements exist in a relatively wide variety of insect and arthropod species. The 28S rDNA-specific RT and 18S rDNA-specific R7 are closely related and their target sequences are similar (Fig. 4).

R2, R8, and R9 are the R2 clade elements containing the RLE domain. R2 constitutes a large group that includes many retrotransposons with the same sequence specificity for 28S rDNA and is distributed over at least
six animal phyla: Arthropoda, Nematoda, Chordata, Echinodermata, Platyhelminthes, and Cnidaria (17, 21, 22, 77, 83). R2 elements have been found recently in many fishes and in some birds and reptiles, but not in mammal genomes (22, 23, 84). The R2 clade can be further classified into four subclades, A, B, C, and D. The R2-A, -B, -C, and -D subclades have three, two, two, and one zinc-finger motif(s), respectively, in the N-terminal region (Fig. 2) (22, 23). R8 identified in Hydra magnipapillata in the Cnidaria phylum is categorized as a member of the R2-A subclade and is likely to have been derived from R2, changing its target specificity from 28S rDNA to 18S rDNA (Fig. 3A) (77). In contrast to the target changes between RT and R7 described above, we found no similarities between the sequences around the target sites of R2 and R8. The target recognition of longer sequence may explain a lack of target similarity between R2 and R8 (77). Recent reports by Christensen and co-workers suggest that the members of R2-A and R2-D subclades use different targeting mechanisms (85, 86).

R9 has been recently identified from bdelloid rotifers (81). R9 is a member of R2-A subclade and inserts within the newly characterized 28S rDNA target (Fig. 3A and Table 1). The long TSD (126 bp) found at the insertion site of R9 is unique; a short (up to 20 bp) TSD or a small deletion (R2Bm) are usually observed at the target sites. R4 is closely related to some elements that insert in TAA repeats and to some elements with no evident specificity (clade R4). Another rDNA-specific non-LTR-element, R5, is a member of NeSL clade.

MOLECULAR MECHANISM OF SEQUENCE-SPECIFIC RETROTRANSPOSITION

APE-encoding non-LTR retrotransposons

APE domain

The process of TPRT, which is unique to non-LTR retrotransposons, is initiated by the targeting endonuclease that cleaves one (bottom) strand of the target site and creates a free 3′-hydroxyl end, which is used as a primer for reverse transcription. Thus, target-site selection of sequence-specific non-LTR retrotransposons is primarily or mainly determined by theendonuclease domain itself (87).

L1Tc from Trypanosoma cruzi encodes a domain that has a true AP endonuclease activity; it has been shown to cleave apurinic/apyrimidinic (AP) sites, indicating a possible role of L1Tc in DNA repair (88). However, other APE-encoding non-LTR elements appear to have lost this intrinsic enzymatic activity. Detailed analysis of the APE domain of the human nonsequence-specific retrotransposon L1 has shown that it preferentially cleaves the T–A junction similar to in vivo target sequences (5′-TTTT/AA-3′) (11, 89). APEs purified from R1Bm and Tx1L APEs have been also shown to cleave the 28srDNA and Tx1D, respectively, target sequences in vitro (90, 91). Purified APE of TRAS1 could generate nicks in a highly specific manner on both strands of the telomeric repeats, between T and A on the (TTAGG)n bottom strand and between C and T on the (CCTAA)n top strand (Fig. 4) (58). These sites are consistent with insertion sites expected from the genomic structure of boundary regions of TRAS1. Time-course studies of the nicking activity of TRAS1 APE showed that the cleavages of the bottom strand preceded those on the top strand (58). TRAS1 APE cleaves not only the bottom strand but also the top strand in a sequence-specific manner in vitro, while the mechanism of the top strand cleavage in vivo is not clarified. The cleavage of TRAS1 APE is not affected by the flanking sequence, suggesting that the target-site specificity of TRAS1 is mainly determined by its endonuclease domain. In R1Bm, Tx1L, and TRAS1, the bottom-strand cleavage is faster than the top-strand cleavage. This behavior is consistent with the TPRT model of Luan et al. (6), in which the bottom-strand cleavage defines the initial target site for reverse transcription.

Crystal structures of APEs from some non-LTR elements have revealed folding patterns similar to those of AP endonucleases (34, 59, 92). However, the structures have an extra beta-hairpin at the DNA-binding surface, which is a common feature among APEs of non-LTR elements (34, 59, 92). The crystal structure of APE of human L1 suggests that the prominent betaB6–betaB5 hairpin loop may insert into the DNA minor groove at the T–A junction, which is important for recognizing the DNA target (34). Further mutational approaches have revealed that variations in the loop sequence result in altered DNA-nicking profiles, including novel sites (34). Similarly, the extra beta hairpin (beta10–beta11) of APE of TRAS1 fits into the minor groove at the telomeric repeats, and Asp-130 in the hairpin is suggested to be involved in specific recognition of telomeric repeats (59). Structural analysis of APE of R1Bm has shown that mutations on the DNA binding surface decrease the cleavage activity but do not affect the sequence recognition in most residues. However, amino acids changes at Tyr-98 and Asn-180 had altered cleavage patterns of targets, suggesting important roles of these residues for the sequence recognition (92).
A novel retrotransposition assay system using baculovirus has demonstrated that APE-encoding elements, TRAS1 (60), SART1 (60), R1Bm (93), and R7 (94), are capable of in vivo retrotransposition into their specific targets. Using this system, replacing the APE domain of SART1 with the APE domain from TRAS1 changed the SART1 specificity to that of TRAS1, suggesting that the primary determinant of in vivo target selection is the APE domain (60). More recently, similar swapping experiments have shown that the APE domains of SARTTc1 and SARTBm1 are involved in recognition of the target site for (TCAGG)$_n$ and (TTAGG)$_n$ telomeric repeats, respectively (61). It has also been shown that protein-engineered APE of TRAS1 combined with the telomere-binding proteins can cleave the human telomeric repeat (TTAGGG)$_n$ strands in a sequence-specific manner (95).

**mRNA/target DNA interaction**

It is noteworthy that sequence-specific non-LTR elements usually target the same sequence in different repetitive units in the genome, in contrast to randomly integrated non-LTR elements. This means that a read-through mRNA product of sequence-specific elements can easily bind the same DNA sequence of a different repetitive target (Fig. 7A). Importantly, it has been

**FIGURE 7** Schematic model for interaction between mRNA of a site-specific non-LTR element and target DNA at the 3′ junction. (A) Interaction between the read-through transcript of a non-LTR element (from copy1) and the target site DNA at the 3′ junction of copy 3. (B) Annealing of R2Ol mRNA with the target DNA (28S rDNA) at the 3′ junction. In vitro-transcribed R2Ol with 4 bp of the 28S target sequence at its 3′ end showed the accurate and efficient retrotransposition in the zebrafish embryo. However, the R2Ol mRNA with 3 bp of the 28S target sequence showed insufficient retrotransposition (39%) (H. Mizuno and H. Fujiwara, unpublished data). These observations indicate that annealing of the read-through product of R2Ol to the target DNA at the 3′ junction seems important for its efficient and accurate retrotransposition. doi:10.1128/microbiolspec.MDNA3-0001-2014.77
suggested that some rDNA-specific non-LTR elements are transcribed by RNA polymerase I, not by RNA polymerase II (96). R2 of \textit{D. simulans} can be co-transcribed with 28S rRNA and processed at or near the 5′ end of the retrotransposon unit by a self-cleavage mechanism (97, 98) (see Fig. 7B, which shows similar steps in R2Ol). The self-cleavage site in most R2 elements from different species was shown not at the 5′ end of the element but at 28S sequences up to 36 nucleotides upstream of the 5′ junction (98). It has been postulated that the co-transcribed 28S sequences of the first cDNA strand at the 5′ junction may anneal to the target site and prime the synthesis of the second DNA strand (98).

In R1Bm, the read-through transcript of 28S rRNA target sequence at the 3′ junction is suggested to be base-paired with the DNA target nicked by endonuclease of R1Bm and exposed as a “primer” of reverse transcription (93). During TPRT, this annealing of mRNA to the exposed end of nicked DNA target places the RNA template at the accurate position for site-specific retrotransposition. Reverse transcription of R1Bm starts from the position next to the read-through UGU sequence, the 5′ end sequence of TSD (28S sequence adjacent to the 3′ end of R1Bm) (93). It has been also shown that R2Ol is co-transcribed with 28S rRNA, and its transcription appears to continue past the 3′ end of the retrotransposon unit in several tissues of the fish (H. Mizuno and H. Fujiwara, unpublished data) (Fig. 7B). A similar interaction between mRNA and target DNA and an accurate reverse transcription were observed at R2Ol, and at least a 4-bp interaction at the 3′ junction is necessary for its efficient retrotransposition (Fig. 7B).

It has been also suggested that base-pairing between the 3′-end sequence of the RNA and the DNA target supports the formation of the primer-template complex (90). The 3′-end sequence of CR1 RNA could be involved in target-site selection by hybridization to homologous sequences at nicked chromosomal sites (99, 100). In telomeric repeat-specific SART1Bm, short telomeric repeat-like GGUU sequences in the 3′ UTR of mRNA may anneal the bottom strand of (TTAGG)$_n$ repeats, which helps the element to access the target DNA (101).

Access to the target site

Most of ORF1p appears indispensable for the activity of APE-type elements (60, 102). ORF1 proteins of the mammalian L1 and the silkworm SART1 have been shown to form a ribonucleoprotein (RNP) complex with their mRNA (103, 104, 105), and contain nucleic acid chaperone activity (106, 107). In some site-specific non-LTR elements, evidence has been obtained that suggests ORF1p could be involved in gaining access to the target genomic site. ORF1 of telomere-specific non-LTR retrotransposons of \textit{Drosophila}, He-T-A, TART, and TAHRE encodes a Gag-like protein and transports it back into the nucleus (108). The Gag-like protein from the He-T-A element localizes to chromosome ends in interphase nuclei (109). It has been suggested that the He-T-A ORF1p specifically localizes to telomeric ends in interphase nuclei and that the TART ORF1p moves to chromosome ends if assisted by the He-T-A ORF1p (109). ORF1p of TAHRE also requires the help of He-T-A ORF1p to localize to chromosome ends (110).

A recent report has shown that ORF1p of He-T-A is expressed in the early S phase and forms spherical structures at telomeres undergoing replication (52). In the telomeric repeat-specific SART1 of the silkworm, ORF1p includes telomeric repeat interaction domains in vitro and may be found in the telomere dot structures in the nucleus (111).

Figure 8 summarizes each process of the retrotransposition of APE-encoding elements, especially focusing on three factors involved in the sequence specific integration.

Other factors

Zinc knuckle-like motifs in ORF1

ORF1 proteins with one to three cysteine-rich motifs (CCHC motifs) encoded at carboxy terminal are found in many clades of APE-encoding non-LTR elements (Fig. 2 and Fig. 8). In retroviruses, the similar CCHC motifs in Gag proteins, called zinc knuckle motifs, participate in interaction between retroviral RNA and Gag proteins (112, 113), and thus the CCHC motifs in ORF1p of APE-encoding elements are also called zinc knuckle motifs (zinc knuckle-like motifs in this manuscript). It is intriguing that the zinc knuckle-like motif is conserved in both Tx1 and R1 clades, which include sequence-specific non-LTR elements (2). Detailed mutation analyses have shown that three CCHC motifs in ORF1p from SART1 are involved in the interaction with its mRNA in a sequence-specific manner; this suggests that the motifs may have played important roles in the evolution of closely related site-specific elements (105). Matsumoto et al. have also shown that the SART1 ORF1p includes domains for the ORF1p–ORF1p and ORF1p–ORF2p interactions (105). These data suggest that the formation of SART1 RNP complex composed of ORF1p, ORF2p, and mRNA is mediated by the ORF1p.
Zinc finger-like motifs in ORF2
The ORF2 proteins of most APE-type retrotransposons encode at least one CCHC motif at their C-terminal regions, but the structure is different from that of the CCHC motifs in ORF1 (Fig. 2 and Fig. 8) (105, 114). This motif in ORF2 is also conserved in both Tx1 and R1 clades. Point mutations in this motif in human L1 and Bombyx TRAS1 cause the loss of the retrotransposition activity (60, 102). However, the elements from the CR1 group, which has no CCHC in ORF2, retain the retrotransposition activity (114). It is generally assumed that the zinc finger (ZF)-like motif is involved in the interaction with the RNA template, DNA target, or proteins, but the function of CCHC in ORF2 has not been elucidated yet.

Myb-like domain
A Myb-like domain is found between endonuclease and RT domains in ORF2 of telomere-specific TRAS elements (Fig. 2) (56). Many telomere-binding proteins, like RAP1, TAZ1, TRF1, and TRF2 share the Myb-like three-helix motif (115). Myb-like motifs in TRAS1 may interact with the telomere structure, although further study is needed to confirm this assumption.

RLE-encoding non-LTR retrotransposons
Many of the RLE-encoding non-LTR retrotransposons are site-specific. All clades of RLE-encoding non-LTR elements (R2, R4, NeSL, CRE, and HERO) encode some DNA-binding motifs, one to three ZFs at the N-terminal end and one CCHC (cysteine–histidine) motif C-terminal to the RT domain within a multifunctional single ORF. R2-clade elements also encode a Myb-like domain (Fig. 2). In contrast to APE-retrotransposons, the functional roles of RLE-type endonuclease itself in sequence specificity is not known clearly to date (86, 116, 117), and the target-site recognition is achieved primarily through the DNA-binding motifs mentioned above.

R2 elements, particularly R2Dm and R2Bm, have been used as model systems to investigate the mechanism of target-site selection of RLE retrotransposons. R2Bm (classified as subclade R2-D) encodes a single ZF motif, while some other R2 elements, such as R2Lp, R2Ol, R8, and R9 (classified as subclade R2-A), encode three ZF motifs. A study of R2Bm has shown that two subunits of the ORF protein are bound to the R2Bm mRNA (86, 118). One protein subunit at the 3′ UTR of the RNA is bound upstream of the insertion site of the target DNA through the function of an undetermined domain. It has...
been reported that the RLE of the upstream protein cleaves the bottom strand of the 28S target DNA and the RT domain of this protein catalyzes first-strand cDNA synthesis using the nicked DNA as a primer (118, 119). The other protein subunit bound to the 5′ UTR of the RNA is bound downstream from the insertion site through the N-terminal ZF and Myb motifs and may be involved in the second DNA-strand synthesis (86, 120). Myb is a major contributor to the specificity and ZF provides some DNA contacts in R2, while other clades of RLE elements have no Myb motif (83, 86, 120). Recent studies suggest that R2-A clade elements, such as R2Lp and R9, may use Myb and ZF (mainly the third ZF from the N-terminal) motifs to bind to upstream sequences of the insertion site (86). The different binding modes of R2-A and R2-D elements suggest a plasticity of integration mechanisms of R2-clade elements. The authors of the study also extend the targeting mechanism of R2 elements to the members of another RLE-clade NeSL element (86).

APPLICATION OF SEQUENCE SPECIFIC RETROTRANSPOSONS AS A GENE DELIVERY TOOL

Precise transgene integration to a specific target site is necessary to avoid unpredictable side effects, particularly in therapeutic applications for human cells. Recently developed novel methods such as TALEN (transcription activator-like effector nuclease) or CRISPR (clustered regularly interspaced short palindromic repeats/Cas9 have been used to cleave and edit the genome regions in a sequence-specific manner, but the transgene integration achieved using these methods is still ongoing (121, 122). Also, transposons combined with ZF protein or adeno-associated virus proteins have been used as target-specific gene delivery tools (123).

Site-specific non-LTR retrotransposons may be good candidates for the sequence-specific gene delivery (94). The site-specific non-LTR elements TRAS, SART, R1, R2, R7, and R8, when recombined in the AcNPV vector, could integrate into their respective specific targets very effectively. Many non-LTR elements recognize the 3′ UTR of their mRNA during the initial step of the TPRT process. A sequence such as an EGFP reporter with an added 3′ UTR sequence can be retrotransposated in a target-specific manner with a helper non-LTR element by trans-complementation, a useful feature for a gene-delivery tool (61, 101). Baculovirus AcNPV has a wide host range and can infect many species, including human cells (124). It has been shown that R1 and SARTI recombined in AcNPV can integrate at specific target sites in the genome of various tissues when injected into the larva of B. mori (125); this gene-transfer system is also useful in the honey bee (126).

In vitro-transcribed R2Ol retrotransposons into the 28S rDNA target sequence of the zebrafish embryo with a high frequency and accuracy and is transmitted to the next generation (A. Kuroki and H. Fujiwara, unpublished data). It has been shown that fluorescent signals are seen in the G2 generation of transgenic fish whose 28S gene has been integrated with R2Ol carrying the reporter EGFP gene in its 3′ UTR. Moreover, the adenovirus-mediated and AcNPV-mediated R2Ol retrotransposition results in an accurate integration into the 28S target in several types of cultured human cells. These results demonstrate that site-specific non-LTR retrotransposons, such as R2Ol, could become a novel type of sequence-specific gene-delivery vectors in gene-therapy application.

ACKNOWLEDGEMENTS

I gratefully appreciate Dr. K. K. Kojima for a critical reading of the manuscript and for helpful comments. My research was supported by grants from the Ministry of Education, Culture, Sports and Technology of Japan.

REFERENCES


Site-specific non-LTR retrotransposons


