Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 7847–7852, July 1999
Biochemistry

Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements

JIN YANG*, HARMIT S. MALIK, AND THOMAS H. EICKBUSH†

Department of Biology, University of Rochester, Rochester, NY 14627-0211

Communicated by Maxine F. Singer, Carnegie Institution of Washington, Washington, DC, May 12, 1999 (received for review March 25, 1999)

ABSTRACT The non-long terminal repeat (LTR) retrotransposon, R2, encodes a sequence-specific endonuclease responsible for its insertion at a unique site in the 28S rRNA genes of arthropods. Although most non-LTR retrotransposons encode an apurinic-like endonuclease upstream of a common reverse transcriptase domain, R2 and many other site-specific non-LTR elements do not (CRE1 and 2, SLACS, CZAR, Dong, R4). Sequence comparison of these site-specific elements has revealed that the region downstream of their reverse transcriptase domain is conserved and shares sequence features with various prokaryotic restriction endonucleases. In particular, these non-LTR elements have a Lys/Arg-Pro-Asp-X$_{12–14aa}$–Asp/Glu motif known to lie near the scissile phosphodiester bonds in the protein–DNA complexes of restriction enzymes. Site-directed mutagenesis of the R2 protein was used to provide evidence that this motif is also part of the active site of the endonuclease encoded by this element. Mutations of this motif eliminate both DNA-cleavage activities of the R2 protein: first-strand cleavage in which the exposed 3′ end is used to prime reverse transcription of the RNA template and second-strand cleavage, which occurs after reverse transcription. The general organization of the R2 protein appears similar to the type IIS restriction enzyme, *Fok*I, in which specific DNA binding is controlled by a separate domain located amino terminal to the cleavage domain. Previous phylogenetic analysis of their reverse transcriptase domains has indicated that the non-LTR elements identified here as containing restriction-like endonucleases are the oldest lineages of non-LTR elements, suggesting a scenario for the evolution of non-LTR elements.

Eukaryotic retrotransposable elements can be divided into two lineages that utilize completely different mechanisms of integration (summarized in ref. 1). Those elements with long terminal repeats (LTRs), the LTR retrotransposable elements, are similar both in structure and in their retrotransposition mechanism to retroviruses (2). Reverse transcription of the RNA templates from these elements is primed by cellular tRNA molecules. Because the reverse transcriptase of these elements is capable of jumping from the terminal repeats at one end of the template to the other end, synthesis of first and second strands results in a complete double-stranded DNA intermediate. This DNA molecule then is integrated into the host chromosome, utilizing an integrase similar to the transposase of DNA-mediated elements (3).

Non-LTR retrotransposable elements, on the other hand, appear to use a simpler mechanism of retrotransposition. Reverse transcription of the RNA template is primed by a 3′ hydroxyl group released by cleavage of the chromosomal target site, a process termed target-primed reverse transcription (TPRT) (4). Synthesis of the cDNA directly onto the chromosome means that integration of non-LTR elements requires only the reverse transcriptase and a "simple" endonuclease. A first clue as to the nature of non-LTR endonucleases was obtained with the discovery of sequence similarity between the amino-terminal end of the second ORF of certain non-LTR elements and cellular apurinic/apyrimidinic (AP) endonucleases (5, 6). Direct evidence now has been obtained that this AP-like domain serves as the endonuclease for non-LTR element integration (6, 7).

We recently have completed a comprehensive phylogenetic analysis of all non-LTR elements that shows these elements can be divided into 11 distinct lineages, each dating back to the pre-Cambrian era (8). Although eight of these lineages contain an AP endonuclease located upstream of the reverse transcriptase (RT) domain, three lineages do not. The latter include R2 elements that insert in the 28S rRNA genes of arthropods (9), R4/Dong elements that insert in the 26S rRNA genes of nematodes or the spacer region of insect rDNA units, respectively (10, 11), and CRE/SLACS-related elements that insert in the spliced leader exons of trypanosomes (12–15). Phylogenetically, these three site-specific clades appear to be the earliest-diverging groups of non-LTR elements (8).

This laboratory has conducted a number of studies of the endonuclease cleavage and RT activity of the R2 element from *Bombyx mori* (4, 16–18). In this report, we show that the sequence-specific endonuclease of R2 is located downstream of the RT domain. Similar motifs are also found in members of the R4/Dong and CRE/SLACS lineages of non-LTR elements. The sequences of these motifs are similar to the active site of certain restriction enzymes.

MATERIALS AND METHODS

Point mutations were generated by QuikChange Site-Directed Mutagenesis (Stratagene) of expression construct pR260 (4). Primers PA..D (5′-GGTCTCCGTAAGCCGGCTATTATCG CCTCCAGGG-3′), PE..D (5′-GGTCTCCGTAAGCCGGAGATTATCGCCTCCAGGG-3′), and YAYD (5′-CGCTCTGGCCTATGCTTACGACCTAGTCCTGC-3′) and their reverse complements were used for *Pfu* polymerase amplification under conditions specified by the manufacturer. Individual transformed products were sequenced to verify the mutations.

Wild-type and mutant R2 proteins were expressed in JM109 and purified as described (18). Mutant proteins PA..D and PE..D were assayed during purification for RT activity by using oligo rA:dT substrates as described below, whereas wild-type

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviations: LTR, long terminal repeat; RT, reverse transcriptase; TPRT, target primed reverse transcription; AP endonuclease, apurinic-apyrimidinic endonuclease.

*Present address: Department of Genetics, Duke University Medical Center, Durham, NC 27710.

†To whom reprint requests should be addressed. e-mail: eick@uhura.cc.rochester.edu.

and YAYD proteins were assayed for specific endonuclease activity (4). Proteins were stored in 50% glycerol/0.4 M NaCl/25 mM Tris·HCl, pH7.5/1 mM DTT at −20°C. Protein concentrations were determined on SDS gels by using the stain Sypro Red (FMC) and the fluoroimaging function of a Storm 860 PhosphorImager (Molecular Dynamics).

Linear 110-bp DNA substrates were prepared by PCR amplification as described previously (18) except that the PCR primers in the reaction were end-labeled with T4 polynucleotide kinase by using 30 μCi [γ-^{32}P]ATP (3,000 Ci/mmol). The prenicked circular DNA template was generated by the incubation of 10 μg of the plasmid pB109 with 1 μg R2 protein in the absence of RNA for 30 min. The products were extracted with phenol/chloroform, ethanol-precipitated, and separated on a 1% agarose gel. The open-circle DNA band was excised and ethanol-precipitated.

RNA in the TPRT reaction was obtained by run-off transcription with T7 RNA polymerase of the construct pBMR2–249A (18). DNA cleavage and TPRT reactions were performed in 20-μl volumes containing 50 mM Tris·HCl, pH 8/200 mM NaCl/10 mM $MgCl_2$/1 mM DTT/0.5 μg R2 RNA/10 μM each of dNTPs. Reactions were stopped by the addition of 3 μl 0.5 M EDTA, and the products were separated on 17-cm 8% denaturing polyacrylamide gels (18). Quantitations were performed by using a Storm 860 Phosphoimager after drying the gel. The standard RT extension assays were performed in 20-μl reactions with 0.3 μg oligo rA:dT/2.5 μM dTTP/2 μCi [α-^{32}P]dTTP (3,000 μCi/mmol)/50 mM Tris·HCl, pH 8/10 mM $MgCl_2$/200 mM NaCl/4 ng of the R2 protein. After 15 min at 37°C, the reactions were spotted onto DE-81 filters, dried, washed three times in 0.3 M NaCl/0.03 M sodium citrate and once in 70% ethanol, dried again, and counted.

RESULTS

As shown in Fig. 1, R2 elements encode a single ORF of approximately 1,100 aa with a centrally located RT domain (19). Sequence comparisons of this R2 ORF from species representing the diversity of arthropods have revealed highly conserved regions both upstream and downstream of the RT domain (19). Near the amino-terminal end of the R2 ORF are two short, conserved domains. The first domain is an exact match to the consensus zinc-finger motif Phe-X-Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-5}-His (CCHH), originally identified in the transcription factor TFIIIa. This motif is perhaps the most prevalent DNA-binding motif found in eukaryotic proteins (20). Immediately downstream of the R2 CCHH motif is a domain with similarity to a DNA-binding motif identified first in the protooncogene, c-myb (21). These conserved domains suggest that the amino-terminal region of the R2 protein is likely to be a DNA-binding domain containing both

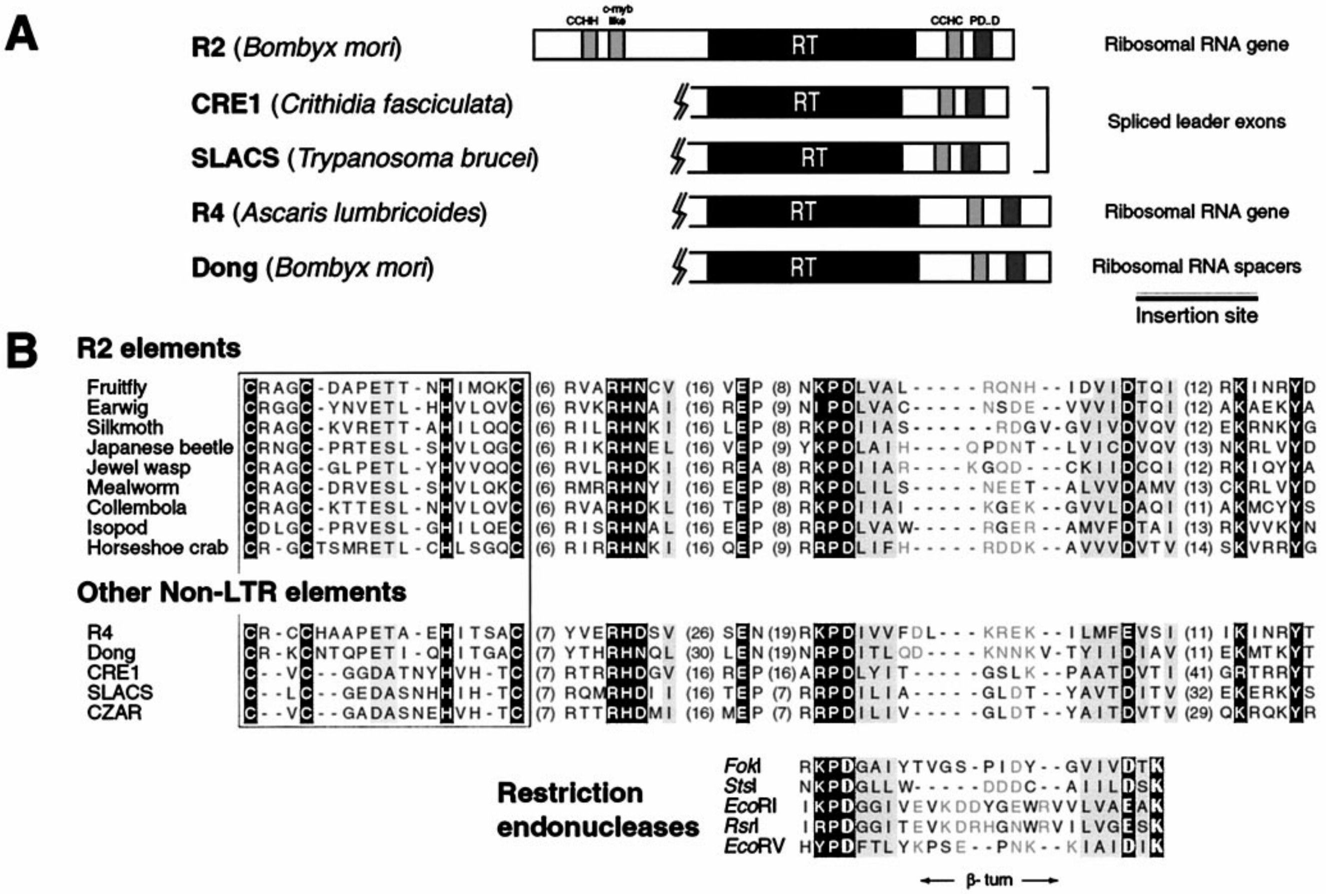
A
CCHH
c-myb like
CCHC
PD..D
R2 (Bombyx mori)
RT
Ribosomal RNA gene
CRE1 (Crithidia fasciculata)
RT
SLACS (Trypanosoma brucei)
RT
Spliced leader exons
R4 (Ascaris lumbricoides)
RT
Ribosomal RNA gene
Dong (Bombyx mori)
RT
Ribosomal RNA spacers
Insertion site
B
R2 elements
Fruitfly CRAGC-DAPETT-NHIMQKC (6) RVARHNCV (16) VEP (8) NKPDLVAL------RQNH--IDVIDTQI (12) RKINRYD
Earwig CRGGC-YNVETL-HHVLQVC (6) RVKRHNAI (16) REP (9) NIPDLVAC------NSDE--VVVIDTQI (12) AKAEKYA
Silkmoth CRAGC-KVRETT-AHILQQC (6) RILRHNKI (16) LEP (8) RKPDIIAS-------RDGV-GVIVDVQV (12) EKRNKYG
Japanese beetle CRNGC-PRTESL-SHVLQGC (6) RIKRHNEL (16) VEP (9) YKPDLAIH----QPDNT--LVICDVQV (13) NKRLVYD
Jewel wasp CRAGC-GLPETL-YHVVQQC (6) RVLRHDKI (16) REA (8) RKPDIIAR----KGQD---CKIIDCQI (12) RKIQYYA
Mealworm CRAGC-DRVESL-SHVLQKC (6) RMRRHNYI (16) EEP (8) RKPDLILS------NEET--ALVVDAMV (13) CKRLVYD
Collembola CRAGC-KTTESL-NHVLQVC (6) RVARHDKL (16) TEP (8) RKPDIIAI------KGEK--GVVLDAQI (11) AKMCYYS
Isopod CDLGC-PRVESL-GHILQEC (6) RISRHNAL (16) EEP (8) RRPDLVAW------RGER--AMVFDTAI (13) RKVVKYN
Horseshoe crab CR-GCTSMRETL-CHLSGQC (6) RIRRHNKI (16) QEP (9) RRPDLIFH------RDDK--AVVVDVTV (14) SKVRRYG
Other Non-LTR elements
R4 CR-CCHAAPETA-EHITSAC (7) YVERHDSV (26) SEN (19) RKPDIVVFDL---KREK--ILMFEVSI (11) IKINRYT
Dong CR-KCNTQPETI-QHITGAC (7) YTHRHNQL (30) LEN (19) NRPDITLQD-----KNNKV-TYIIDIAV (11) EKMTKYT
CRE1 C--VC--GGDATNYHVH-TC (7) RTRRHDGV (16) REP (16) ARPDLYIT------GSLK--PAATDVTI (41) GRTRRYT
SLACS C--LC--GEDASNHHIH-TC (7) RQMRHDII (16) TEP (7) RRPDILIA------GLDT--YAVTDITV (32) EKERKYS
CZAR C--VC--GADASNEHVH-TC (7) RTTRHDMI (16) MEP (7) RRPDILIV------GLDT--YAITDVTV (29) QKRQKYR
Restriction endonucleases
FokI RKPDGAIYTVGS-PIDY--GVIVDTK
StsI NKPDGLLW-----DDDC--AIILDSK
EcoRI IKPDGGIVEVKDDYGEWRVVLVAEAK
RsrI IRPDGGITEVKDRHGNWRVILVGESK
EcoRV HYPDFTLYKPSE--PNK--KIAIDIK
← β-turn →

FIG. 1. Identification of the putative endonuclease domain in R2 and other site-specific non-LTR elements. (*A*) Schematic diagram of the R2 ORF from *B. mori* and its comparison to the C-terminal ends of other site-specific non-LTR retrotransposable elements. Shaded regions in R2 indicate the RT domain, putative zinc finger (CCHH), and c-myb-like DNA-binding motifs. Carboxyl terminal to the RT domain in all elements is a putative zinc finger motif (CCHC) and a motif (PD..D) with similarity to restriction enzymes. (*B*) Sequence comparison of the putative endonuclease domain of nine R2 elements from diverse arthropods (9) with those of other sequence-specific non-LTR elements and with some restriction endonucleases. The number of amino acid residues between the conserved motifs is given in parentheses. Highly conserved residues are shown in shaded boxes, with the three charged residues that are part of the active site of the restriction enzymes also bolded. The region between these conserved residues assumes a β-turn in restriction enzymes (24, 25). The large number of charged residues in these β-turns are indicated as light gray.

zinc-finger and c-myb-like binding motifs. Site-directed mutagenesis experiments are underway to directly test the role of these motifs.

The entire 250-aa region downstream of the RT domain also was found to be highly conserved across all R2 elements (19). This domain includes another putative nucleic acid-binding motif of the type Cys-X$_{2–3}$-Cys-X$_{7–8}$-His-X$_4$-Cys (CCHC). Although variants of this motif have been identified previously in a number of other site-specific and nonspecific non-LTR elements and some similarity to retroviral gag proteins was noted (10), this particular spacing of Cys and His residues is not a good match to any characterized RNA- or DNA-binding protein (20). In addition to the CCHC motif, an extensive region of similarity was found between R2 and members of the R4/Dong and CRE/SLACS lineages of non-LTR elements (Fig. 1) by using the multiple-alignment features of CLUSTAL X (22). Even more revealing, literature searches and iterative BLAST searches (23) revealed matches within this domain to a conserved motif in a variety of restriction endonucleases. Restriction enzymes are known to have little sequence homology but conserved structural motifs (24, 25). Significantly, these non-LTR elements and restriction enzymes have in common the motif Lys/Arg-Pro-$\underline{\text{Asp}}$-X$_{12–19\text{aa}}$-$\underline{\text{Asp/Glu}}$, with the two acidic residues (underlined) separated by a sequence likely to fold into a β-turn. For *Eco*RV, *Eco*RI, and *Fok*I, these acidic residues (abbreviated here as PD..D) have been shown to lie in close proximity to the scissile phosphodiester bonds in the protein–DNA complex (24, 25). Mutations in these residues inactivate the enzymes without affecting their ability to bind their recognition site (26, 27).

To determine whether this motif also is involved in the endonuclease activity of the R2 element, we mutated the first invariant Asp residue in the PD..D sequence of the previously characterized silkmoth R2 protein (4, 17). Two mutations were generated: a conservative Asp-to-Glu change (PE..D) and a nonconservative, Asp-to-Ala change (PA..D). As a control for R2 enzymatic activity, a mutant protein also was generated containing an Asp-to-Tyr change in the first D of the highly conserved RT motif, YXDD (YAYD). Such a mutation has been shown to eliminate RT activity in both LTR and non-LTR retrotransposable elements (28–30). These three mutations did not appear to alter the structure of the R2 protein, because proteins containing each of these mutations were stably expressed in *Escherichia coli* and behaved like wild-type protein in all purification steps. The R2 protein-purification procedure requires tight binding of the protein to both RNA and DNA substrates (4).

Fig. 2*A* shows the assay used to monitor the enzymatic activity of the various R2 mutations. The target DNA is a 5′ labeled, 110-bp DNA segment containing the R2 insertion site (4, 18). Cleavage of the lower (primer) DNA strand occurs first in the TPRT reaction and generates a 60-nt-labeled product on a denaturing gel. In the presence of RNA, primer-strand cleavage is followed by a slower cleavage of the upper (nonprimer) strand, generating a labeled, 48-nt fragment (18). The assay also contains dNTPs and a 283-nt R2 RNA sequence representing the 3′ untranslated region of the silkmoth R2 element, the only RNA sequence required for protein recognition by the R2 reverse transcriptase (17). Utilization of this RNA as a template in the TPRT reaction results in a labeled, 343-nt DNA fragment. Fig. 2*B* shows a denaturing gel of the products from this assay by using each of the mutant proteins. In the presence of wild-type protein, each of the labeled products of the TPRT reaction can be seen. A PhosphorImager was used to quantify cleavage of both the primer and nonprimer strands as well as the extent of the TPRT reaction for each protein (Table 1). The activity of each protein in a standard RT extension assay by using poly(A) as the template and oligo(dT) as the primer is shown in Table 1.

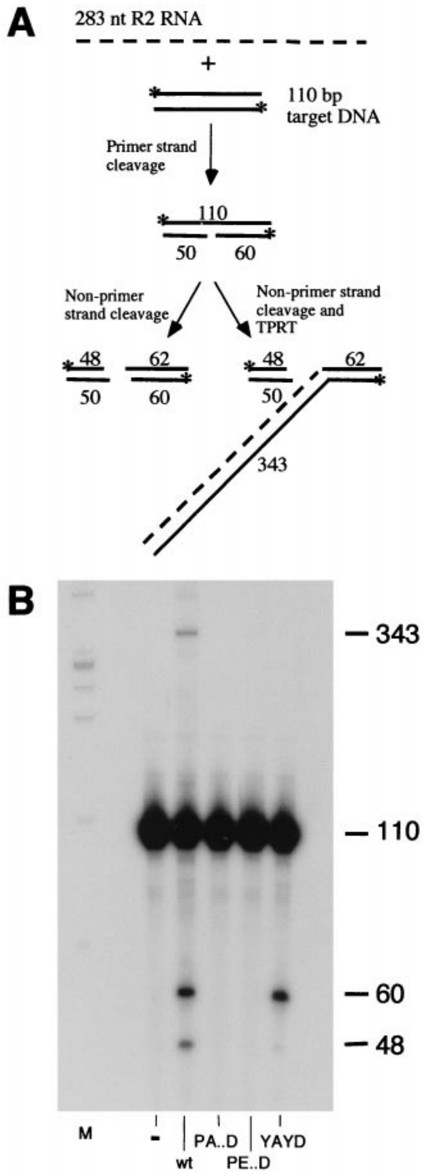

FIG. 2. Enzymatic activity of the R2 protein mutations. (*A*) TPRT assay by using a 110-bp 5′ end-labeled target. The two strands of the target DNA are represented by the straight lines, with the labeled 5′ ends noted with an asterisk. The cDNA made in the reaction is indicated by a straight line, and the RNA template is indicated by a dashed line. The 283-nt RNA template contains the sequence of the 3′ untranslated region of the silkmoth R2 element. In the absence of TPRT, 60- and 48-nt labeled DNA fragments are detected on a denaturing gel. If TPRT occurs, a 343-nt fragment also is generated. (*B*) Autoradiography of the TPRT reaction run on an 8% denaturing polyacrylamide gel. For each reaction, 15 ng (200 fmol) of end-labeled DNA and 4 ng (30 fmol) of protein were incubated for 30 min. Lanes: 1, no protein; 2, wild-type R2 protein; 3, PA..D mutation; 4, PE..D mutation; and 5, YAYD mutation. Numbers to the right indicate the lengths of the observed DNA products.

The PA..D mutant protein has normal levels of activity in the standard RT assay, but is unable to cleave either the primer or nonprimer strands of the 28S target site (Fig. 2*B*). The PE..D mutation also has normal levels of standard RT activity and no visible endonuclease activity. However, more sensitive PhosphorImager quantitation indicated that the PE..D protein has 2% of the activity of wild-type protein (Table 1). Approximately 50-fold reductions in cleavage activity also have been detected with conservative D-to-E changes in the active site of the restriction enzyme *Eco*RV (26). Finally, the YAYD mutant

Table 1. Endonuclease and RT activities of the mutant R2 proteins normalized to that of wild-type protein

Protein	RT activity	Primer strand cleavage	Nonprimer strand cleavage	TPRT activity
Wild-type	1.00	1.00	1.00	1.00
PA..D	0.82	<0.01	<0.01	<0.01
PE..D	0.78	0.02	<0.01	0.06
YAYD	<0.01	0.92	0.10	<0.01

All values represent the average of three experiments.

protein readily cleaves the primer strand of the target DNA, but it has no activity in either the standard RT or TPRT assays. The absence of TPRT activity explains the reduced level of nonprimer strand cleavage by the YAYD mutation. We have shown previously with the wild-type protein that if the TPRT reaction is prevented, because of the absence of dNTPs, for example, then cleavage of the nonprimer strand is inhibited (4, 18). The amount of nonprimer strand cleavage by the YAYD mutant varies from experiment to experiment (see Fig. 4).

To determine whether the mutant PE..D and PA..D proteins are capable of the TPRT reaction when provided with a cleaved target site, we conducted two sets of experiments. As diagrammed in Fig. 3*A*, in the first set of experiments plasmid DNA containing the R2 insertion site was either supercoiled or prenicked at the target site with wild-type R2 protein (see *Material and Methods*). The supercoiled and prenicked plasmids were incubated with the mutant proteins in the presence of the 283-nt R2 RNA template and ^{32}P-labeled dNTPs. As seen in Fig. 3*B* (lanes 1–4), if the plasmid DNA is not prenicked, the PA..D mutant cannot conduct the TPRT reaction, whereas the PE..D mutant does support a very low level of cleavage, which then is used for TPRT. With the prenicked DNA substrate (Fig. 3*B*, lanes 5–8), both mutations are able to conduct TPRT at levels similar to those of the wild-type protein (Fig. 3*B*, lane 9). However, unlike the wild-type protein, neither mutant protein is capable of converting the open-circle plasmid into a linear form by cleaving the nonprimer strand. This experiment clearly suggests that the PD..D motif of R2 is involved in cleavage of both primer and nonprimer DNA strands.

The normal levels of RT activity seen in the PE..D and PA..D mutants and the normal level of endonuclease activity by the YAYD mutant would suggest that these mutants should complement each other and catalyze a complete TPRT reaction. This possibility was tested in Fig. 4 by using the end-labeled, 110-bp target DNA. As shown in lanes 2 and 4, mutants PE..D and YAYD alone are unable to generate significant amounts of the TPRT product, but a mixture of the two proteins (Fig. 4, lane 3) generates the 343-nt TPRT product at levels typical of wild-type protein (Fig. 2), confirming the ability of RT and endonuclease mutants to complement each other. A similar set of experiments also has been conducted with the PA..D and YAYD protein with similar results (data not shown). Based on these simple complementation assays we do not know whether a heterodimer of PE..D and YAYD protein is formed and is responsible for the complete TPRT reaction or whether there is a sequential binding and cleavage of the DNA by the YAYD mutant followed by binding and reverse transcription by the PE..D mutant. In either event, the experiments in Figs. 3 and 4 clearly indicate that although changes in the first Asp residue of the R2 PD..D motif eliminates DNA cleavage, it does not affect binding of the R2 protein to DNA or its ability to conduct the TPRT reaction.

DISCUSSION

The sequence comparisons and site-directed mutagenesis studies described here provide evidence that the catalytic domain of the endonuclease encoded by the R2 element is located at the carboxyl-terminal end of its ORF. The DNA-recognition sequence of the R2 protein is nonpalindromic, with the region required for protein recognition predominately located upstream of the cleavage site (16). The R2 protein can bind this DNA target as a monomer (18). Therefore, although the Lys/Arg-Pro-Asp-X$_{12-14\ \mathrm{aa}}$–Asp/Glu motif of the R2 endonuclease is similar to a variety of restriction enzymes, perhaps most significant is its relationship to the type IIS enzymes such as *Fok*I. *Fok*I also binds as a monomer to a nonpalindromic sequence and cleaves downstream of this recognition sequence (25). The domain structure of R2 and *Fok*I have similarities as well. The *Fok*I catalytic domain is

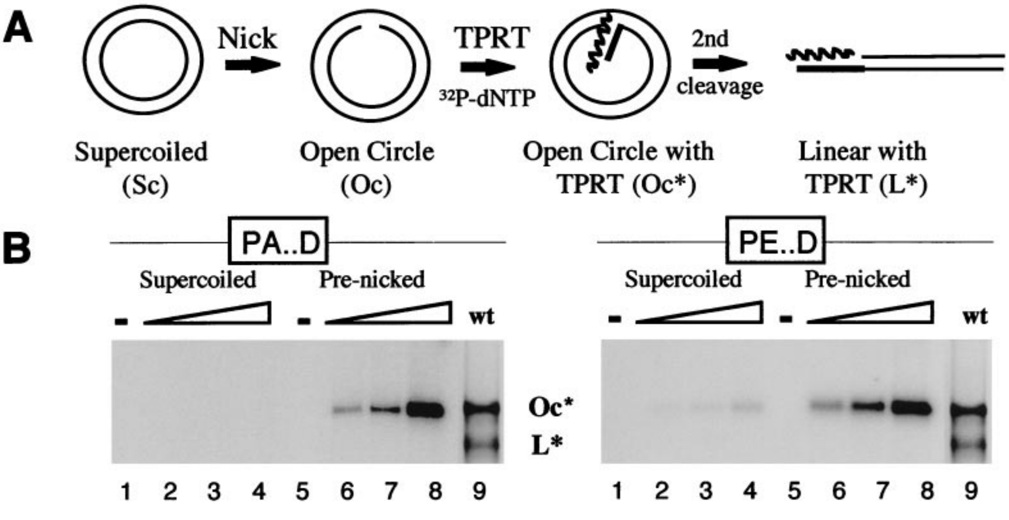

FIG. 3. Endonuclease mutants can conduct the TPRT reaction on prenicked DNA substrates. (*A*) Schematic diagram summarizing the substrates and products of the TPRT reactions. The two strands of the DNA substrate are indicated by thin lines; the cDNA synthesized during the reaction is indicated with a thicker line and the RNA template is indicated with a wavy line. (*B*) Autoradiographs of the reaction products with the PA..D mutation (*Left*) and PE..D mutation (*Right*) separated on 1% agarose gels. Only plasmids that have undergone the TPRT reaction can be seen in these autoradiograms. For each reaction, 0.25 μg supercoiled or prenicked plasmid DNA was incubated with the R2 protein as in Fig. 2, except that 2 μCi [α-^{32}P]dATP was added to each reaction. Lanes: 1 and 5, no protein controls; 2–4, 4, 8, and 16 ng of the mutant protein incubated with the supercoiled target; 6–8, 4, 8, and 16 ng of the mutant protein incubated with the prenicked target; and 9, 16 ng of wild-type protein incubated with supercoiled DNA.

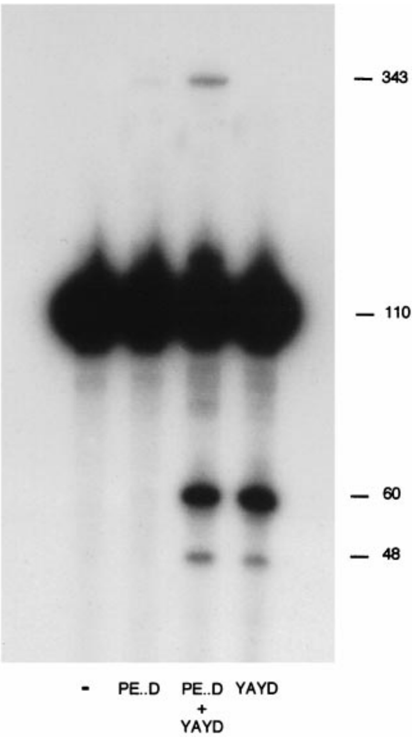

FIG. 4. Complementation of the endonuclease and RT mutations. Assays were performed as diagrammed in Fig. 2*A*, by using the 110-bp end-labeled DNA target. Lanes: 1, no protein; 2, 4 ng PE..D mutation; 3, 4 ng PE..D and 4 ng YAYD mutations; and 4, 4 ng YAYD mutation. Neither the PE..D or YAYD mutant alone can use the DNA target to prime reverse transcription of the 283-nt R2 RNA; however, a mixture of the two proteins is capable. Twice the total level of protein was added in lane 3 compared with the wild-type lane in Fig. 2 to enable the formation of an equivalent amount of an active heterodimer of PE..D and YAYD compared with a wild-type dimer. We have no direct evidence, however, that such a heterodimer is formed.

located at the carboxyl end of the protein and is separated from the DNA-recognition domain by a flexible-hinge domain (25). In the case of R2, highly conserved zinc-finger and c-myb-like DNA-binding motifs are located at the amino-terminal end of the protein (19), separated by the RT domain from its catalytic carboxyl-terminal domain (Fig. 1*A*). Preliminary site-directed mutagenesis of the amino-terminal CCHH motif has provided direct evidence that this domain of the R2 protein is involved in DNA binding (J.Y. and T.H.E., unpublished data).

Although the R2 protein can bind the DNA target site as a monomer, this binding is readily competed by nontarget DNA, and only cleavage of the lower (primer) strand is catalyzed (18). The addition of R2 RNA to the R2 protein induces the formation of a protein dimer that has higher binding specificity for the target site and enables cleavage of both the upper and lower DNA strands (18). Thus, it is possible that each subunit of the R2 protein dimer cleaves one strand of the DNA helix. Alternatively, the dramatic difference in cleavage kinetics of the two DNA strands in the TPRT reaction (cleavage of the first strand is completed before reverse transcription, whereas cleavage of the second strand follows reverse transcription) may reflect a conformational change, enabling that same subunit to be positioned for cleavage of the upper DNA strand 2 bp upstream of its initial nick on the lower strand. There is also an unusual mechanism used for double-stranded cleavage by the *Fok*I protein (25). It recently has been suggested that two *Fok*I monomers, each bound to the DNA-recognition sequences of separate *Fok*I sites, interact *in trans* to form a dimer enabling DNA cleavage (31, 32).

Little is known of the enzymatic activities encoded by the other site-specific non-LTR elements that contain an endonuclease domain like R2. The target sites for these other site-specific elements are also nonpalindromic, and in the case of CRE and SLACS elements, zinc finger motifs can be identified in the amino-terminal region of their ORFs (12–15). The organization of their ORFs and the sequence similarity throughout their carboxyl-terminal domain suggest that these other site-specific elements bind and cleave their target DNA in a manner similar to that of R2. Recent phylogenetic analysis of the non-LTR retrotransposable elements (8) has shown that CRE/SLACS, R4/Dong, and R2 represent three of the oldest lineages of non-LTR elements. Based on the sequence of their RT domains, non-LTR elements are, in turn, closely related to those group II introns of mitochondria and bacteria that encode an RT domain (33). Consistent with this phylogenetic relationship, group II intron mobility also is based on a TPRT mechanism of retrotransposition (34, 35). Cleavage of the DNA strand used as primer is brought about by a protein-encoded activity located downstream of the RT domain. This protein domain has been shown to have sequence similarity to endonucleases with conserved H-N-H motifs (36, 37).

As summarized in Fig. 5, group II introns and the oldest lineages of non-LTR elements have similarity in both the organization of their ORFs and in the location of their endonucleases. These similarities clearly add support to models in which the group II introns and the non-LTR elements have a common origin. However, there are two major differences between the TPRT reaction of group II introns and that of site-specific non-LTR elements, as represented by R2. First, group II elements use reverse splicing of the intron RNA for cleavage of the upper DNA strand, whereas R2 cleaves both DNA strands via the PD..D catalytic domain. Second, DNA cleavage specificity by group II introns is accomplished by their

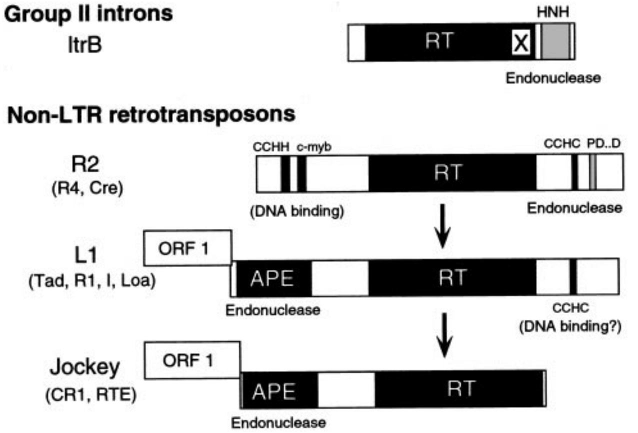

FIG. 5. Comparison of the ORFs encoded by group II introns and non-LTR retrotransposable elements. The group II intron shown is the ltrB intron of *Lactococcus lactis* (37). The protein domains shared by other group II introns are shaded and are similar to those identified previously (41), except that the domains referred to as Z and X in that study are shown here as part of the RT domain (see ref. 8). The putative endonuclease domain of the group II introns is identified as HNH (38, 39). In the case of the non-LTR retrotransposons, schematic diagrams of the R2, L1, and Jockey elements are shown as representatives of the major non-LTR structures found to date. Other major lineages of non-LTR elements with these basic structures are listed within the parentheses (8). The CCHH, c-myb, CCHC, and PD..D domains of the R2 elements are described in Fig. 1. The AP-like endonuclease domain identified at the amino-terminal end of L1 and Jockey elements is labeled APE. Elements with structures similar to L1 contain a CCHC domain downstream of their RT domain; thus, this region is likely to be involved in DNA binding. Arrows represent the likely path of non-LTR evolution in eukaryotes based on the phylogeny of their RT domains (8).

RNA sequences annealing to the target DNA in the reverse splicing reaction (38), whereas R2 DNA recognition is accomplished by protein domains probably located at both ends of its ORF.

Based on what is now known about non-LTR element phylogeny and the different types of endonucleases encoded by these elements, we can propose a correlation between the structure, site-specificity, and age of the non-LTR retrotransposable elements. The compact genomes of mitochondria and bacteria may have required the group II introns to retain site-specificity, whereas the increasing size of eukaryotic nuclear genomes may have allowed the original site-specific non-LTR elements to exploit the greater opportunity for random insertion without deleterious consequences. Only a few clades of the original site-specific non-LTR elements, those residing in conserved multigene families, have survived. The diversification of non-LTR elements into lineages lacking site specificity is correlated with the acquisition of a relatively nonspecific endonuclease, the AP-like endonuclease, upstream of the RT domain. Although all non-LTR lineages containing an AP-like domain have lost the PD..D motif of the carboxyl-terminal endonuclease domain, many of these elements still retain a domain with a highly conserved carboxyl-terminal CCHC motif (see Fig. 5). This CCHC motif has been shown to be essential for L1 activity in mammals (30). Thus, it is possible that this carboxyl domain, although no longer catalytic, plays an essential role in DNA binding. Only in the more recently evolved lineages of the non-LTR elements (e.g., Jockey, CR1, and RTE) (8) is this carboxyl domain completely eliminated.

It is interesting to note that a few of the non-LTR elements with the AP domain have retained target specificity: R1 elements also insert in the 28S rRNA genes of insects (9), Tx1 elements specifically insert into another mobile element of *Xenopus* (39), and Zepp elements insert into preexisting copies of themselves in *Chlorella* (40). It can be suggested that insertion into preexisting copies of a mobile element may represent an intermediate between extreme target-site preference and a random mode of insertion. It is easy to imagine how the evolutionary trend away from site specificity would accelerate as the activity of non-LTR elements both contributed to, and benefited from, an increase in low cost insertion sites.

We thank W. D. Burke for generating the site-directed mutations and W. D. Burke and D. G. Eickbush for comments on the manuscript. This work was supported by National Institutes of Health Public Health Services Grant GM42790.

1. Eickbush, T. H. (1994) in *The Evolutionary Biology of Viruses*, ed. Morse, S. S. (Raven, New York), pp. 121–157.
2. Whitcomb, J. M. & Hughes, S. H. (1992) *Annu. Rev. Cell Biol.* **8,** 275–306.
3. Mizuuchi, K. (1992) *Annu. Rev. Biochem.* **61,** 1011–1051.
4. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. (1993) *Cell* **72,** 595–605.
5. Martin, F., Maranon, C., Olivares, M., Alonso, C. & Lopez, M. C. (1995) *J. Mol. Biol.* **247,** 49–59.
6. Feng, Q., Moran, J. V., Kazazian, H. H. & Boeke, J. D. (1996) *Cell* **87,** 905–916.
7. Feng, Q., Schumann, G. & Boeke, J. D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 2083–2088.
8. Malik, H. S., Burke, W. D. & Eickbush, T. H. (1999) *Mol. Biol. Evol.* **16,** 793–805.
9. Burke, W. D., Malik, H. S., Lathe, W. C. & Eickbush, T. H. (1998) *Nature (London)* **239,** 141–142.
10. Burke, W. D., Müller, F. & Eickbush, T. H. (1995) *Nucleic Acids Res.* **23,** 4628–4634.
11. Xiong, Y. & Eickbush, T. H. (1993) *Nucleic Acids Res.* **21,** 1318.
12. Aksoy, S., Williams, S., Chang, S. & Richards, F. F. (1990) *Nucleic Acids Res.* **18,** 785–792.
13. Gabriel, A., Yen, T. J., Schwartz, D. C., Smith, C. L., Boeke, J. D., Sollner-Webb, B. & Cleveland, D. W. (1990) *Mol. Cell. Biol.* **10,** 615–624.
14. Villanueva, M. S., Williams, S. P., Beard, C. B., Richards, F. F. & Aksoy, S. (1991) *Mol. Cell. Biol.* **11,** 6139–6148.
15. Teng, S.-H., Wang, S. X. & Gabriel, M. (1995) *Nucleic Acids Res.* **23,** 2929–2936.
16. Xiong, Y. & Eickbush T. H. (1988) *Cell* **55,** 235–246.
17. Luan, D. D. & Eickbush, T. H. (1995) *Mol. Cell. Biol.* **15,** 3882–3891.
18. Yang, J. & Eickbush, T. H. (1998) *Mol. Cell. Biol.* **18,** 3455–3465.
19. Burke, W. D., Malik, H. S., Jones, J. P. & Eickbush. T. H. (1999) *Mol. Biol. Evol.* **16,** 502–511.
20. Berg, J. M. & Shi, Y. (1996) *Science* **271,** 1081–1085.
21. Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. & Nishimura, Y. (1994) *Cell* **79,** 639–648.
22. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
24. Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P., Bryan, R. K., Martin, P. D., Petratos, K. & Wilson, K. S. (1993) *EMBO J.* **12,** 1781–1795.
25. Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I. & Aggarwal, A. K. (1997) *Nature (London)* **388,** 97–100.
26. Selent, U., Rüter, T., Kohler, E., Liedtke, M., Thielking, V., Alves, J., Oelgeschlager, T., Wolfes, H., Peters, F. & Pingoud, A. (1992) *Biochemistry* **31,** 4808–4815.
27. Waugh, D. S. & Sauer, R. T. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 9596–9600.
28. Larder, B. A., Purifoy, D. J. M., Powell, K. L. & Darby, G. (1987) *Nature (London)* **327,** 716–717.
29. Mathias, S. L., Scott, A. F., Kazazian, H. H., Boeke, J. D. & Gabriel, A. (1991) *Science* **254,** 1808–1810.
30. Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D. & Kazazian, H. H. (1996) *Cell* **87,** 917–927.
31. Wah, D. A., Bitinaite, J., Schildkraut, I. & Aggarwal, A. K. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 10564–10569.
32. Bitinaite, J., Wah, D. A., Aggarwal, A. K. & Schildkraut, I. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 10570–10575.
33. Xiong, Y. & Eickbush, T. H. (1990) *EMBO J.* **9,** 3353–3362.
34. Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. (1995) *Cell* **82,** 1–10.
35. Cousineau, B., Smith, D., Lawrence-Cavanagh, S., Mueller, J. E., Yang, J., Mills, D., Manias, D., Dunny, G., Lambowitz, A. M. & Belfort, M. (1998) *Cell* **94,** 451–462.
36. Shub, D. A., Goodrich-Blair, H. & Eddy, S. R. (1994) *Trends Biochem. Sci.* **19,** 402–406.
37. Gorbalenya, A. E. (1994) *Protein Sci.* **3,** 1117–1120.
38. Eskes, R., Yang, J., Lambowitz, A. M. & Perlman, P. S. (1997) *Cell* **88,** 865–874.
39. Garrett, J. E., Knutzon, D. S. & Carroll, D. (1989) *Mol. Cell. Biol.* **9,** 3018–3027.
40. Higashiyama, T., Noutoshi, Y., Fujie, M. & Yamada, T. (1997) *EMBO J.* **16,** 3715–3723.
41. Mohr, G., Perlman, P. S. & Lambowitz, A. M. (1993) *Nucleic Acids Res.* **21,** 4991–4997.