Analysis of the 5′ Junctions of *R2* Insertions With the 28S Gene: Implications for Non-LTR Retrotransposition

Janet A. George, William D. Burke and Thomas H. Eickbush

Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received September 14, 1995
Accepted for publication December 6, 1995

ABSTRACT

R2 elements are non-long terminal repeat retrotransposable elements that insert into 28S rRNA genes of most insect species. The single open reading frame of *R2* encodes a protein with both endonuclease activity, which cleaves the target site, and reverse transcriptase activity, which uses this cleavage to prime reverse transcription. This target-primed reverse transcription mechanism is also used by group II introns. Little is known of the mechanism by which the 5′ end of *R2* is integrated after reverse transcription. We have determined the 5′ junction sequence of 94 *R2* elements from 14 different species of Drosophila. Only 37% of the full-length elements contained precise 5′ junctions; the remainder contained deletions of the 28S gene and/or insertions of additional sequences. Because the 5′ junctions of truncated copies were similar to full-length elements, no sequences at the 5′ end of *R2* appear to be required for element integration. A model in which the *R2* reverse transcriptase is capable of switching templates from the *R2* RNA transcript to the upstream 28S gene can best explain the observed 5′ junction sequences. This template jumping is analogous to the template switching of retroviral reverse transcriptases during formation of the double-stranded integration products.

NON-LONG terminal repeat (non-LTR) retrotransposable elements (also referred to as LINE-like elements) are widely distributed, highly abundant mobile elements of eukaryotes. They differ from the LTR-containing retrotransposable elements by the absence of terminal repeats, and their encoded open reading frames (ORFs) lack any identifiable integrase domain (EICKBUSH 1994). Molecular phylogenetic studies using the sequence of their reverse transcriptase domain indicate that non-LTR retrotransposable elements form a monophyletic group, distinct from the LTR-retrotransposable elements (XIONG and EICKBUSH 1988a, 1990; MCCLURE 1993). In fact, the non-LTR elements appear to be more similar to the group II introns of fungi mitochondria (LAMBOWITZ and BELFORT 1993) and the retrons of bacteria (INOUYE and INOUYE 1993), than to the LTR-retrotransposable elements.

The mechanism used by the non-LTR retrotransposable elements to insert within the genome is still poorly understood. Integration of these elements has been shown to result from the reverse transcription of an RNA intermediate by the demonstration that intron sequences placed within an element are precisely removed during the generation of new copies *in vivo* (EVANS and PALMITER 1991; JENSEN and HEIDMANN 1991; PELISSON *et al.* 1991). This retrotransposition mechanism must differ in several important steps from that of retroviruses and LTR retrotransposons, in that the non-LTR elements do not have the structural components (terminal repeats, tRNA primer binding sites) or encode the proteins (RNase H and integrase) required by the retroviral mechanism (MARTIN 1991; EICKBUSH 1994). Because the 3′ ends of non-LTR elements are usually intact while the 5′ ends are often truncated, it has been suggested that the reverse transcriptase encoded by each non-LTR element might be able to directly polymerize its reverse transcript (cDNA) onto a chromosome at a nick or break on the chromosome (SCHWARZ-SOMMER *et al.* 1987; FINNEGAN 1989; HUTCHISON *et al.* 1989; BUCHETON 1990). This model has been greatly supported by studies of the sequence specific *R2* element of *Bombyx mori* (BURKE *et al.* 1987; XIONG and EICKBUSH 1988b). The 120-kD polypeptide encoded by the single ORF of *R2* has been shown to be able to nick the target DNA and use the 3′ OH group exposed by this nick to prime reverse transcription of an R2 transcript (LUAN *et al.* 1993). The *R2* enzyme is also capable of adding nontemplated nucleotides before engaging the RNA in the reverse transcription reaction, thus sometimes generating short repeat sequences at the 3′ end of the element, a characteristic feature of non-LTR retrotransposable elements (LUAN and EICKBUSH 1995). Following reverse transcription of the RNA template, the second strand of the target DNA is cleaved.

The target DNA-primed reverse transcription mechanism of *R2* is probably used by many other non-LTR retrotransposable elements. Whether these elements

Corresponding author: Thomas H. Eickbush, Department of Biology, University of Rochester, Rochester, NY 14627.
E-mail: teickbush@the.biology.rochester.edu

encode their own endonuclease or utilize preexisting nicks or breaks on the chromosome for their insertion is not known. Evidence for the former has been suggested by the presence in some non-LTR retrotransposable elements of a protein domain with similarity to the the major human apurinic endonuclease (MARTIN *et al.* 1995). Direct support for the generality of the *R2* mechanism is the recent finding that the mobility of the group II intron, *al2*, also depends upon the priming of reverse transcription by the target DNA after specific endonuclease cleavage (ZIMMERLY *et al.* 1995).

While the *in vitro* experiments have revealed the initial steps in the integration reaction, they have not helped to resolve the mechanism by which the cDNA of the non-LTR element is attached to the upstream target DNA, nor have they indicated how the RNA template is removed from the DNA and the second DNA strand synthesized. In this report we have taken a different approach to study how the *R2* element attaches to the upstream target sequences. *R2* elements specifically insert at a unique site in the 28S rRNA genes of their host (Figure 1A). Just as the sequence specificity of *R2* for this site has greatly aided the biochemical analysis, this specificity has made it possible to score a large number of naturally occurring integrated products all resulting from insertions into identical target sequences. A striking amount of sequence variation was seen at the 5′ end of the *R2* elements providing insights into the mechanism used by *R2* to complete its integration reaction.

MATERIALS AND METHODS

Strains and DNA isolation: The geographical strains of *D. melanogaster* used have been previously described (JAKUBCZAK *et al.* 1992). Stocks of species from the *melanogaster* and *obscura* species groups were obtained from the National Drosophila Species Resource Center (Bowling Green State University). Species from the *testacea* and *quinaria* groups were obtained from JOHN JAENIKE (University of Rochester). Genomic DNA was isolated from 10 to 50 adults of each stock as previously described (EICKBUSH and EICKBUSH 1995).

PCR amplification and sequence determination: The 5′ ends of *R2* elements were obtained by PCR amplification of genomic DNA using primers complementary to the *R2* and 28S gene positions shown in Figure 1A. Primer 1, 5′-TCNCKCCARTANGGNACCAT-3′ (N, any nucleotide; Y, T or C; R, A or G; W, A or T; K, G or T), encoded the reverse complement of the amino acid sequence, MVPYWRE, starting at position 1108 of the *D. melanogaster R2* element (JAKUBCZAK *et al.* 1990). Primer 2, 5′-CATRTGNACNCCNARNCC-3′, encoded the reverse complement of the amino acid sequence GLGVHM starting at position 403 of the *R2* element. Primers 1 and 2 were used in combination with primer 4, 5′-CTAAGTCGACTGCCCAGT-3′, complementary to the 28S gene sequence starting 62 bp upstream of the *R2* insertion site. *Taq* DNA Polymerase (Bethesda Res. Lab.) was used under conditions specified by the supplier. Approximately 0.2 μg of genomic DNA was amplified in 30 cycles of 94° for 1 min, 55° for 1 min and 72° for 3 min. Primer 1 worked well for species within the *melanogaster* species group and was used for the subsequent sequencing studies. Primer 2 gave the most reproducible results for species from the *obscura, testecea* and *quinaria* groups and was therefore used in the sequencing studies. For determination of the nucleotide sequence of the junctions, the PCR-amplified products were purified from primer sequences by extraction of the DNA bands from agarose gels and cloned into a modified M13mp18 vector, which after digestion with *Xcm*I, is suitable for direct cloning of PCR products (BURKE *et al.* 1995). Multiple clones of both orientations from each species were sequenced by the single-stranded dideoxy chain termination method (SANGER *et al.* 1977).

For direct determination of the length variation associated with the *R2* 5′ junctions primer 3, 5′-CCTCTGCTCTCAAATAC-3′, which was complementary to a region within the 5′ untranslated region of the *D. melanogaster R2* element starting 12 bp from its 5′ junction with the 28S gene, was used in combination with primer 5, 5′-TCAGAACTGGCACGGAC-3′, which was complementary to the 28S gene sequence starting 168 bp upstream of the *R2* insertion site. To visualize the products on sequencing gels, the PCR amplifications were conducted as above with the addition of 60 nM P^{32}-dATP. After amplification, the products were denatured by adding an equal volume of 95% formamide, 20 mM EDTA, 0.05% bromophenol blue and 0.05% xylene cyanole, incubated at 72° for 5 min, and subjected to electrophoresis next to a sequence ladder on a standard 6% urea-acrylamide sequencing gel.

RESULTS

Early characterization of cloned rDNA repeats from *D. melanogaster* indicated that all *R2* insertions (originally termed type II insertions) contained similar 3′ junctions with the 28S gene, while their 5′ ends were subject to large truncations (LONG *et al.* 1980; DAWID and REBBERT 1981; ROIHA *et al.* 1981). A more extensive PCR analysis of the 3′ end of *R2* elements from six species of the *melanogaster* species subgroup confirmed that the 3′ junctions of *R2* elements in each species were identical except for the length of a poly (A) tract defining the junction with the 28S gene (EICKBUSH and EICKBUSH 1995). Similar results have been obtained in a study of the 3′ ends of *R2* elements from 19 additional species of the Drosophila genus (W. C. LATHE and T. H. EICKBUSH, unpublished observations). As shown in Figure 1B these 3′ junctions are consistent with the initial cleavage generated in the bottom (coding) strand of the 28S gene by the *R2* endonuclease *in vitro* (LUAN *et al.* 1993).

Unlike the extensive analysis of the 3′ junctions of *R2* elements, relatively few 5′ junctions have been characterized (BURKE *et al.* 1987; JAKUBCZAK *et al.* 1990). To clone multiple 5′ *R2*/28S gene junctions, PCR amplification of genomic DNA was conducted using the primer locations shown in Figure 1A. One oligonucleotide, primer 4, was complementary to the 28S gene upstream of the insertion site, and a second oligonucleotide was complementary to one of two regions of the *R2* element encoding conserved amino acid sequences of the ORF, primers 1 or 2 (see MATERIALS AND METHODS).

A total of 94 *R2* 5′ junctions were sequenced from

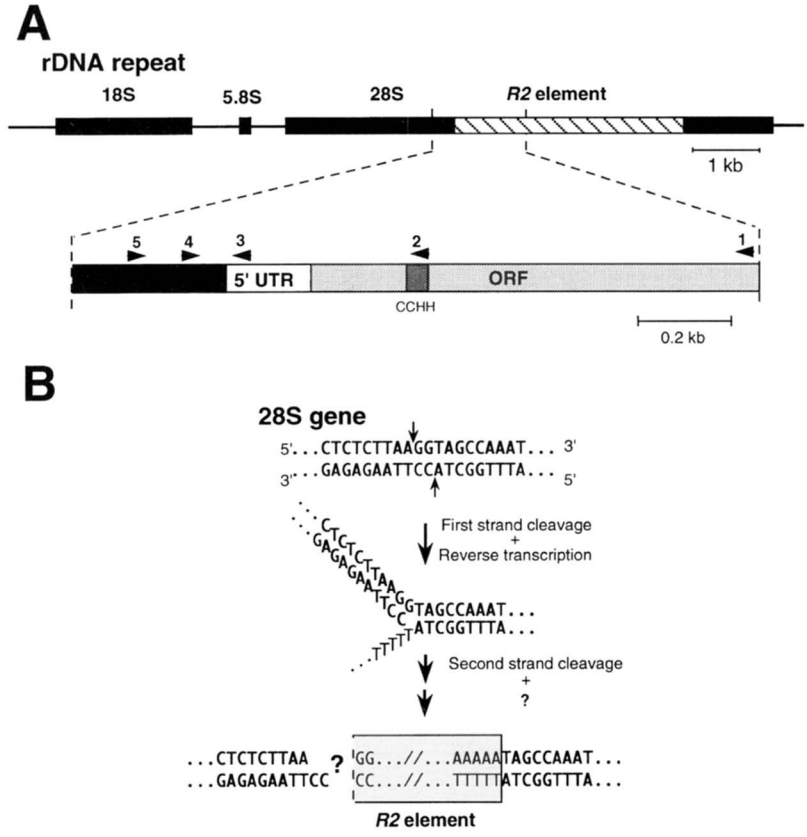

FIGURE 1.—Location of the *R2* element in the rDNA unit. (A) Diagram of an *R2* inserted rDNA unit of *D. melanogaster*. The location of the 18S, 5.8S and 28S genes are indicated by ■, the spacer regions by —, and the *R2* element by ▧. The 5′ end of the *R2* element has been expanded to show the location of the PCR primers used in this study. Various regions of the *R2* element have been indicated as follows: □, 5′ untranslated region (5′ UTR); stippled box, location of the region encoding the ORF; darker stippled box, location of the sequences encoding a putative cysteine, histidine nucleic acid-binding motif. PCR primers and their orientation are indicated by the arrowheads. (B) Cleavage site of the *R2* element and the known steps of the integration reaction. Location of the *R2* endonuclease cleavage sites on the two strands of the 28S gene are indicated by the arrows. The first step of the integration reaction is the cleavage of the bottom strand and the use of the released 3′ OH group to prime reverse transcription. In Drosophila cDNA synthesis always begins with a run of 13–25 T nucleotides, thus the final 3′ junction of the *R2* element is the same in all Drosophila species. The second step of the integration reaction is a cleavage of the upper target DNA strands 2 bp upstream of the first cleavage. The mechanism for the attachment of the *R2* sequences to the upstream 28s gene sequences is not known, and the junction itself is extremely variable. The only common feature is that most elements begin with two G nucleotides (see Figures 2 and 4).

14 species of Drosophila. This included an initial 27 junctions from *D. melanogaster*, to obtain a thorough sampling of junctions from one species, followed by the recovery of two to eight 5′ junctions from 13 other species of Drosophila to determine if the variation seen in *D. melanogaster* was characteristic of *R2* elements in other species. To sample representative species near *D. melanogaster* eight of these other species were chosen from the *melanogaster* species group: *D. simulans, D. sechellia, D. mauritiana* and *D. yakuba* of the *melanogaster* subgroup; *D. takahashii* of the *takahashii* subgroup; and *D. ananassae, D. bipectinata* and *D. varians* of the *ananassae* subgroup. To sample species throughout the genus, the remaining five species were selected from three other species groups: *D. persimilis* and *D. pseudoobscura* from the *obscura* species group; *D. testacea* from the *testacea* species group; and *D. recens* and *D. falleni* from the *quinaria* species group. Comparison of the conserved sequences within the *R2* elements themselves will be discussed in a subsequent report.

Most if not all of the 5′ junctions obtained by this PCR amplification were derived from *R2* elements inserted in typical units of the rDNA locus. *In situ* hybridizations in *D. melanogaster* have indicated that all copies of *R2* are located in the nucleolar organizer (PEACOCK *et al.* 1981). Southern blots using a probe from the 3′ end of the *R2* element have previously shown that most, if not all, *R2* elements in the *melanogaster* species subgroup contained the *Hin*dIII restriction site of the 28S gene located 280 bp downstream of the insertion site (EICKBUSH and EICKBUSH 1995). A similar analysis using probes from the 5′ end of the *R2* element has demonstrated that the *R2* elements of these species also contained the *Cla*I restriction site of the 28S gene located

28S gene		*R2* element	No. clones
D. melanogaster			
TCAACGGCGGGAGTAACTATGACTCTCTTAA		GGGGA---TCATGGGGTAT	4
TCAACGGCGGGAGTAACTATGACTCTCTTAA		GGGGA--GTCATGGGGTAT	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	G	GGGGATCATCATGGGGTAT	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	TT	GGGGA---TCATGGGGTAT	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	GGGAGTAT	GGGGA---TCATGGGGTAT	1
TCAACGGCGGGAGTAACTAAGACTCTCTT--		GGGGA---TCATGGGGTAT	1
TCAACGGCGGGAGTAACTATGACTCTCT---	CTTT	GGGGA---TCATGGGGTAT	1
TCAACGGCGGGAGTAACTAAGACTCTCT---		GGGGA---TCATGGGGTAT	1
TCAACGGCGGGAGTAACTA----------	CCCGGGCGGAACACCA	-GGGA---TCATGGGGTAT	1
TCAAC----------		GGGGA---TCATGGGGTAT	3
TCAACGGCGGGAGTAACTATGACTCTCTTAA	GGG	(28 bp) GAGGGGGAG	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA		(132 bp) ACCTCCTCG	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	G	(511 bp) GAAAGGAGG	1
TCAACGGCGGGAGTAACTATGACTCTCT---	GTGATAACTAGTC	(1931 bp) ACTATTTGC	3
TCAACGACGGGAG----------	GCACAGA	(2070 bp) GTGTTAGAG	1
TCAACGGCGGGAGTAACTATGACTCTCTT--		(2258 bp) CGCCCTTCG	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	G	(2271 bp) GTCCTTATC	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA	GAGA	(2508 bp) CTCACGCAA	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA		(2522 bp) GGTAGCCAG	1
TCAACGGCGGGAGTAACTATGACTCTCTTAA		(2557 bp) GGGCCCGAG	1

FIGURE 2.—Variation in the nucleotide sequences at the 5′ junction of *R2* elements in *D. melanogaster*. The number of PCR clones that were identical to the sequence shown is indicated at the far right of the figure. Nucleotides to the left of the first vertical line are 28S gene sequences. Nucleotides to the right of the second vertical line are *R2* sequences. Nucleotides between the two vertical lines are additional inserted sequences. Segments of these insertions that are at least five nucleotides in length and that are identical to 28S gene sequences are underlined. For the 5′ truncated elements, the number of nucleotides deleted is given in the parenthesis [based on the numbering in JAKUBCZAK *et al.* (1990)].

1.6 kb upstream of the insertion site (data not shown). Finally, sequence variation between *R2* copies from the same species was <1% (data not shown), indicating none of the copies are degenerate in sequence. Therefore, based on their uniform sequence and the conservation of flanking restriction sites within the 28S gene, the cloned 5′ junctions reported here are derived from the uniform family of *R2* elements present in the rDNA locus of each species.

***D. melanogaster* 5′ junctions:** The sequences of the 27 5′ junctions obtained from the standard Oregon R laboratory strain of *D. melanogaster* are shown in Figure 2. Fifteen of the 5′ junctions were derived from full-length *R2* elements, while 12 represented the junctions of 5′-truncated elements. Because of the PCR approach used, the 12 *R2* truncations shown in Figure 2 should not be considered representative of either the abundance or the size of *R2* truncations in *D. melanogaster*. Such estimates have been previously made based on Southern blots (see JAKUBCZAK *et al.* 1992, Figure 6). Most of the truncated copies recovered by our PCR amplifications contained 1.9- to 2.5-kb deletions. The recovery of such large deletions was not expected, because they would not contain the regions bound by PCR primers 1 and 2 (see Figure 1). The recovery of such clones resulted from the cross-hybridization of the degenerate PCR primers to sequences located 2.8 (primer 1) and 3.0 kb (primer 2) from the 5′ end of the *R2* element.

The 5′ junctions of the *D. melanogaster R2* elements were highly variable in sequence. While variation associated with the 5′ truncations was expected because such truncations presumably represent aberrant integration events, it was surprising to find that the 15 full-length elements contained 10 different 5′ ends. Indeed, defining the precise 5′ boundary of the full-length *R2* elements with the 28S gene was somewhat arbitrary. The sequence feature most associated with the end of the full-length *R2* elements was GGGGA. Based on the sequence of only a few *R2* junctions, we and others have previously suggested that the first two G residues in the GGGGA sequence are derived from the 28S gene (DAWID and REBBERT 1981; ROIHA *et al.* 1981; JAKUBCZAK *et al.* 1990). In this interpretation, the 2 bp that are part of the staggered cut in the target DNA (Figure 1B) are not lost during *R2* integration. The expanded data set of 5′ junctions shown in Figure 2 clearly argues against this suggestion. Two-thirds of the full-length junctions contained either deleted 28S sequences and/or the presence of additional nucleotides upstream of these two G residues, strongly suggesting that these G residues are part of the incoming *R2* sequences. Therefore in Figure 2, all sequences to the right of the second vertical line are defined as *R2* sequences, while all sequences to the left of the first vertical line are defined as being derived from the 28S target DNA. Those sequences between the two vertical lines represent additional sequences (insertions). While the origin of these additional sequences is not known, in several cases, segments of these insertions (underlined nucleotides) were identical to segments of the 28S gene 13–27 bp upstream of the cleavage site.

To insure that our sequenced junctions did not represent a biased sampling of the 5′ end sequences in this

strain, as well as to compare this variation to that in other strains of *D. melanogaster,* a more comprehensive method of assaying the 5′ variation in a strain was used. In this approach, direct estimates of the length variation present at the 5′ end of full-length *R2* elements were made using a primer (primer 3, Figure 1) complementary to the *R2* element only 12 bp from the consensus 5′ end of the element, GGGGA. The 28S gene primer was complementary to sequences starting 168 bp upstream of the target site (primer 5 in Figure 1), resulting in amplified DNA of a convenient length to be scored on a 6% polyacrylamide sequencing gel. Full-length *R2* elements with no insertions at their 5′ junctions and with only the two bases of the staggered cut deleted from the target DNA would give rise to a PCR product 197 bp in length. [Because *Taq* polymerase has the tendency to add an additional nucleotide at the end of the template (CLARK 1988), the actual products generated would be a mixture of 197 and 198 bp.]

The PCR amplification products obtained from laboratory strains Oregon R, Canton S and seven geographical isolates of *D. melanogaster* are shown in Figure 3. The *R2* elements in the Oregon R strain (lane 2) gave rise to products that were predominantly 196–201 bp in length (referred to as −1 to +4). A large number of other length bands were also detected ranging from −50 to +65 in length. The different length PCR products obtained from the Oregon R strain were consistent with the 15 cloned full-length 5′ junctions shown in Figure 2. Of the sequenced full-length products, 60% (9/15) would give rise to PCR lengths from −1 to +4. The six remaining full-length junctions shown in Figure 2 would give rise to relative lengths of +8, −2, −3 and three at −26.

The length variation found in the other eight strains of *D. melanogaster* tested in Figure 3 was similar to the variations found in Oregon R. The predominant length products were −1 to +4 in all strains, and most strains contained variants in both the −25 and the +35 regions of the gel. The pattern of other length variants was unique for each strain with some strains containing more insertions (lane 8) and other strains containing more deletions (lane 9). It is interesting to note that a Japanese strain (QD18) had the simplest profile of length variants (lane 5). We have previously shown that strain QD18 had the lowest total number of *R2* elements in a survey of 37 *D. melanogaster* strains (JAKUBCZAK *et al.* 1992). The length variation seen in Figure 3 should be regarded as an underestimate of the actual variety of 5′ junctions present in each strain because many of the sequenced junctions contain combinations of both insertions and deletions of the 28S gene (Figure 2). For example, the nine sequenced clones from Oregon R containing lengths from −1 to +4 represented six different types of junctions. We conclude that most strains of *D. melanogaster* have a broad range of 28S gene deletions and insertions of additional nucleotides at the 5′ end of their full-length *R2* elements, and that the 15 cloned junctions of full-length elements shown in Figure 2 are typical of those present in the species.

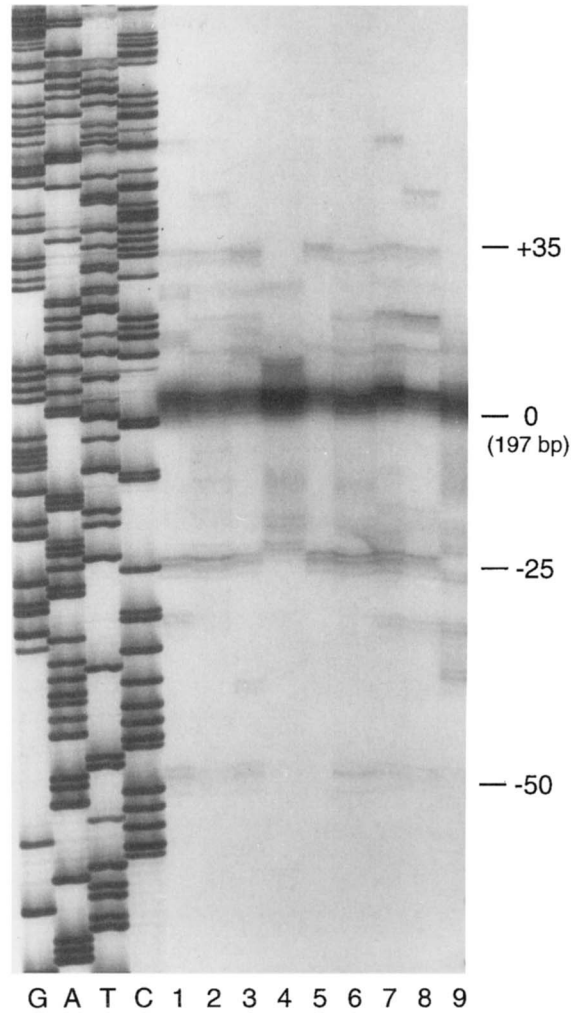

FIGURE 3.—PCR assay to detect sequence heterogeneity at the 5′ end of *R2* elements. PCR amplifications using primers 3 and 5 (Figure 1) were conducted in the presence of ^{32}P-dATP. The products were denatured and run on an 6% polyacrylamide sequencing gel. Most bands appear as doublets presumably because of the tendency for *Taq* DNA polymerase to add an additional A nucleotide at the end of the polymerization (CLARK 1988). Approximate lengths were determined by comparison with the sequence of m13 shown on the left side of the figure. Lane 1, Canton S; 2, Oregon R; 3, Australia (BL-17); 4, California (Lemon Cove 84); 5, Japan (QD18); 6, Netherlands (163); 7, Raleigh NC (CAM 105); 8, Oklahoma (RL); 9, Kenya (5/17/88b#2). Further description of these strains can be found in JAKUBCZAK *et al.* (1992).

Common features of *R2* 5′ junctions throughout Drosophila: The *R2* 5′ junctions obtained from the 13 other species of Drosophila are shown in Figure 4. Nucleotide sequences derived from the 28S gene, the *R2* element, and the insertion of additional nucleotides are separated by vertical lines as described for Figure 2. In most species sufficient numbers of full-length ele-

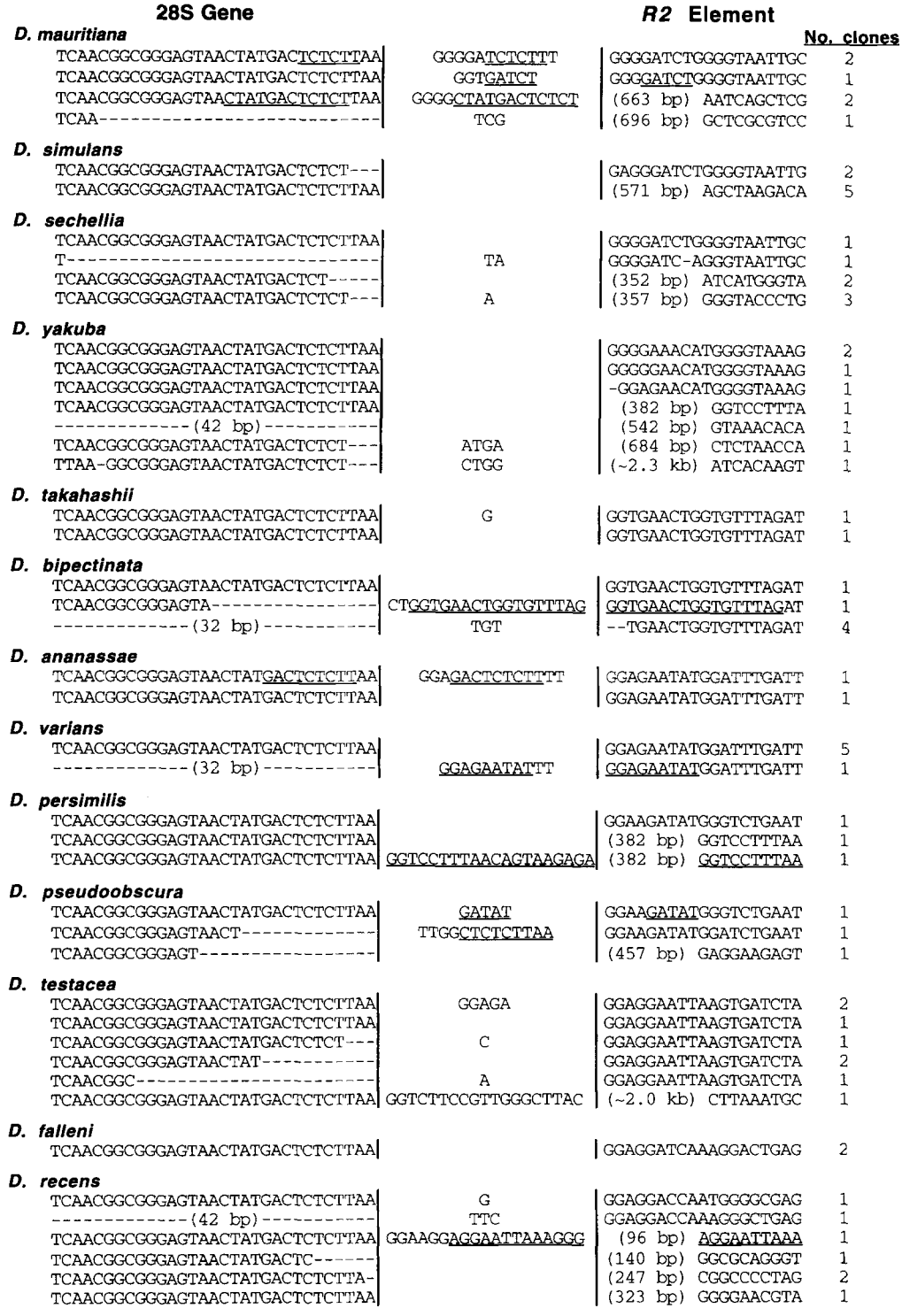

FIGURE 4.—Variation in the nucleotide sequences at the 5′ junction of *R2* elements in 13 species of Drosophila. The number of PCR clones that were identical to the sequence shown is indicated at the far right of the figure. Nucleotides to the left of the first vertical line are 28S gene sequences. Nucleotides to the right of the second vertical line are *R2* sequences. Nucleotides between the two vertical lines are additional inserted sequences. Segments of these insertions, at least five nucleotides in length, that are identical to 28S gene sequences or *R2* sequences are underlined. For the 5′ truncated elements, the number of nucleotides deleted, based on the sequence of full-length elements, is given in the parenthesis. In the case of *D. yakuba* and *D. testacea*, truncations extended beyond the *R2* sequence available from these species, thus an estimated truncation length is given based on homology to the *R2* element of *D. melanogaster*.

ments were recovered that the 5′ ends of the *R2* elements could be defined by sequence comparisons within each species. In the remaining species the 5′ ends were defined by comparing *R2* sequences between closely related species. Only in the case of *D. simulans* was the exact 5′ end somewhat unresolved in that a GA dinucleotide could represent an insertion or the 5′ end of the *R2* element. In Figure 2 this GA dinucleotide was defined as part of the *R2* sequence, because in all other species *R2* elements contained at least five purines in the first six nucleotide positions.

The sequence variants found in all 13 species were of the same general type as found in *D. melanogaster*. Therefore, the following summary of the variation seen at *R2* junctions was based on the combined data from all 14 Drosophila species.

1. Deletion of 28S sequences: The *R2* insertions in all species appear to result in the deletion of 2 bp, consistent with the elimination of the 2 bp within the staggered cut generated by the *R2* endonuclease from *B. mori* (Figure 1B). The elimination of these 2 bp is readily supported by the junction sequences within *D. melanogaster* (Figure 2), *D. pseudoobscura, D. testacea* and *D. recens* (Figure 4). In the case of the other species, this conclusion is inferred by comparisons between species from the same group or subgroup. In addition to this 2-bp deletion, 41% of the junctions (39/94) contained deletions of the 28S gene that extended beyond the staggered cut. Nearly half of these deletions were relatively short, 1–6 bp in length, with 3-bp deletions being most abundant. The other half of the 28S deletions ranged in length from 11 to 42 bp with most of these longer deletions in the range of 11–18 or 26–32 bp.
2. Deletion of *R2* sequences: Unlike the deletions of the target site, only a limited number of junctions contained short deletions of the *R2* sequences (single examples of a 1- and 28-bp deletion in *D. melanogaster,* a 1-bp deletion in *D. yakuba,* and four examples of a 2-bp deletion in *D. bipectinata*). The remaining examples of the *R2* deletions (38 cases) represented deletions of at least 96 bp. No particular sequence appeared to be associated with the locations of these major truncations, although a number did occur at one or two G nucleotides. While these G's could be interpreted as being derived from the 28S gene sequence, most of these junctions had insertions or deletions of the 28S gene beyond the terminal G nucleotides, thus these bases do not appear to be derived from the 28S target site.
3. Insertion of extra nucleotides: 45% (42/94) of the junctions contained from 1 to 20 nucleotides between the 28S gene and *R2* element sequences. In all but one of the longer insertions, a substantial segment of these additional sequences have identity with 28S or *R2* gene sequences near the junction. Instances where this identity was at least five nucleotides in length have been indicated by underlining in Figures 2 and 4. The only long insertion that does not appear to have a counterpart in the *R2* or 28S gene is the 5′ truncated copy in *D. testacea.* However, the exact location of this junction is somewhat ambiguous because an entire *D. testacea R2* element has not been sequenced. The 5′ junction of this *D. testacea* clone was therefore defined by its homology to a region 2.0 kb from the 5′ end of the *R2* elements from *D. melanogaster.*

A limited number of *R2* 5′ junction sequences have also been obtained from *B. mori* and several other widely divergent insects (BURKE *et al.* 1987; W. D. BURKE and T. H. EICKBUSH, unpublished observations). *R2* truncations, 28S gene deletions and duplications similar to those described in Drosophila were seen to a variable degree in all species. Insertions could not be scored in all species because of an insufficient number of junctions to precisely define their 5′ ends.

DISCUSSION

We have previously shown that over 85% of the *R2* elements that have been partially sequenced from species of Drosophila contained neither termination codons nor frameshifts disrupting the ORF (EICKBUSH and EICKBUSH 1995). The rate of nucleotide substitutions at synonymous sites in this ORF averaged 9.6 times the rate at replacement sites, suggesting continuous selection on the encoded proteins of these elements (EICKBUSH *et al.* 1995). These data suggest retrotransposition plays an essential role in maintaining *R2* elements in the 28S genes of its host. Indeed, our evidence to date suggests that *R2* elements turnover rapidly in the rDNA loci. Using a genomic blot assay, the pattern of 5′ truncated *R2* elements was shown to differ for each strain of *D. melanogaster* tested (JAKUBCZAK *et al.* 1992). In this report the pattern of full-length elements was also shown to vary for each strain (Figure 3). Thus individual copies of *R2* do not appear to remain within the rDNA locus for extended periods.

Based on these findings, it is likely that the variation detected at the 5′ end of *R2* elements presented in this report resulted from the process of integration rather than from subsequent recombination or mutation events. Indeed, it was only the 5′ junction of the *R2* element that exhibited such variation in sequence. The 3′ junction of *R2* elements from these same species vary only in the length of the A-rich homopolymer at the precise junction (EICKBUSH and EICKBUSH 1995; W. C. LATHE and T. H. EICKBUSH, unpublished observations). The variation in length of this homopolymer is consistent with our *in vitro* studies of the variation in the number of T residues added to the cDNA strand during

the initiation of reverse transcription (LUAN and EICKBUSH 1995). Truncations of *R2* sequences or deletions of 28S gene sequences have never been detected at the 3′ end of the *R2* elements. This dramatic difference between the 3′ and 5′ junctions suggests that it is the retrotransposition mechanism itself that is generating the many different 5′ junctions. Indeed, the amount of variation we have scored in this report is likely to be an underestimate of the variation generated by the integration mechanism. PCR bias during the amplifications of the junction sequences, and the concerted evolution processes of gene conversion and unequal crossovers acting on the rDNA units would all act to reduce the apparent variation that is scored.

The extensive sampling of the *R2* 5′ junctions and scoring of sequence variants within and between species enables us to evaluate various models for the attachment of the *R2* sequences to the upstream 28S gene. Indeed, over two-thirds of the 5′ junctions obtained in this report were unique, suggesting that there is variability inherent in the mechanism by which *R2* sequences are attached to the upstream 28S gene sequence. Equally important, our results indicate that there are probably no sequences at the 5′ end of the *R2* element that are needed for integration. While it was known that many 5′ truncated *R2* copies could be found in the 28S genes of *D. melanogaster,* it was generally assumed that these insertions represented aberrant integrations. This report has shown for the first time that the 28S-*R2* junctions of these 5′ truncated copies have identical properties to the junctions of full-length *R2* elements. For example, 37% of the full-length elements could be defined as precise, *i.e.,* contain neither deletions of the 28S gene beyond the 2-bp stagger nor insertions of additional nucleotides, compared to 29% of the 5′ truncated elements. Another 39% of the full-length elements contained deletions of the 28S gene compared to 45% of the 5′ truncated elements. Finally, 46% of the full-length elements and 45% of the truncated elements contained additional nucleotides inserted between the *R2* and 28S gene sequences. Whatever mechanism is used to attach the *R2* sequences to the upstream 28S gene in Drosophila, it either does not utilize specific sequences at the 5′ end of the element, or multiple internal sequences can readily substitute for these sequences.

Several models for the attachment of *R2* to the upstream 28S gene sequences are shown in Figure 5. When combined with the known cleavage sites of the *R2* endonuclease (Figure 1B), the 5′ junctions reported here provide evidence for or against each of these models. In model A the 3′ OH exposed by second-strand cleavage serves as the primer for second-strand synthesis of the *R2* element. The signal for this second step could be the reverse transcriptase running off the end of the RNA template, if *R2* transcription begins at the 5′ end of the element. Alternatively, if the *R2* template is derived from a read-through transcript of the 28S gene, the reverse transcriptase could pause when it reaches the secondary structure of the 28S rRNA sequences.

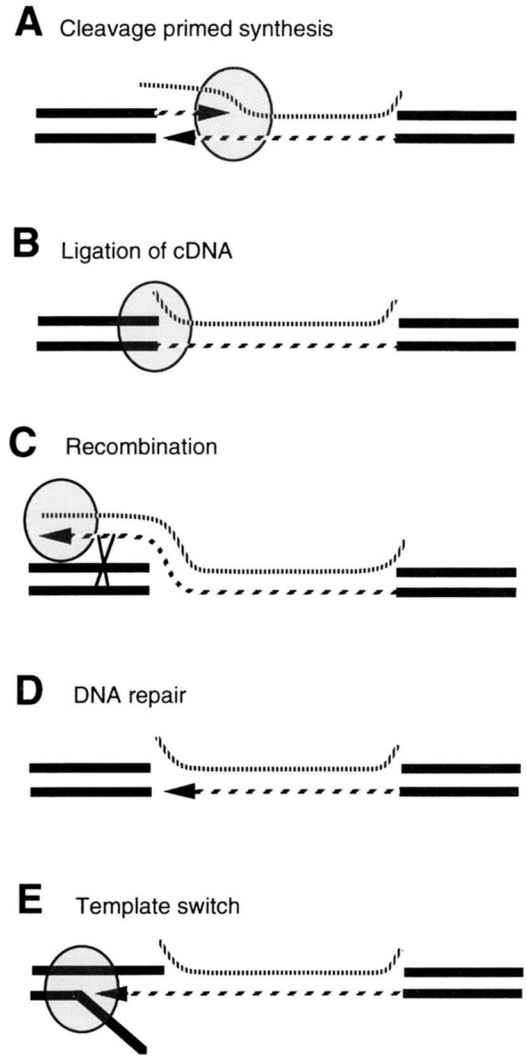

FIGURE 5.—Possible models for the attachment of *R2* sequences to the upstream 28S gene sequences. In all models the first step of the reaction is target DNA-primed reverse transcription of the *R2* template. Thick lines, 28S gene target sequences; dotted line, cDNA and second strand synthesis of *R2* element sequences; stippled line, *R2* RNA template; oval, *R2* protein. (A) Cleavage of the second DNA strand exposes a 3′ OH group to prime synthesis of the second strand of the *R2* element using the cDNA strand as template. (B) Ligation of the 3′ end of the cDNA to the upstream 28S sequences. (C) cDNA synthesis extends to 28S gene sequences upstream of the *R2* sequences. Recombination between this cDNA and the 28S gene attaches the cDNA to the upstream 28S gene. (D) After cDNA synthesis and second strand cleavage no further step are catalyzed by the *R2* protein. Unknown DNA repair processes link the two ends together. (E) At the end of reverse transcription the *R2* polymerase switches from the RNA to the upper strand of the upstream DNA sequence, displacing the lower DNA strand. In all models except model A, second strand synthesis and final polishing of the ends are conducted by DNA repair enzymes.

This model readily explains the deletion of the 2-bp stagger, however it does not explain why in over 40% of the observed junctions additional 28S gene sequences are also eliminated.

A second model for the attachment of *R2* sequences is ligation of the cDNA strand itself to the bottom strand of the upstream 28S gene (model B). This model is similar to that previously proposed for the *cin4* elements of maize (SCHWARZ-SOMMER *et al.* 1987) and the *I* elements of Drosophila (FINNEGAN 1989). This mechanism is unlikely since it would not result in the deletion of either the 2-bp staggered cut or the upstream 28S gene sequences.

A third possible mechanism for 5′ attachment differs from models A and B in that it requires 28S gene sequences be present at the 5′ end of the *R2* template (model C). Reverse transcription of this template beyond the 5′ end of the *R2* elements into the 28S gene would allow homologous recombination of the cDNA strand with the upstream 28S gene sequences. This model is also not supported by the junction sequence data because it would not generate 5′ truncated copies unless the donor RNAs are derived from preexisting 5′ truncated *R2* elements. The very different patterns of 5′ truncated elements seen in different strains of *D. melanogaster* (JAKUBCZAK *et al.* 1992), and the variety of nucleotide additions and deletions seen at the 5′ end of *R2* elements strongly argues against model C.

A fourth model for *R2* integration (model D) is that after reverse transcription and second strand cleavage, the free DNA ends are joined by cellular DNA repair mechanisms. This model predicts that the attachment of the two ends would occur well after cleavage, allowing sufficient time for exonuclease activity to delete part of the 28S gene. Such a repair mechanism also makes the important prediction that full-length and 5′ truncated elements would have similar junctions. Double-stranded break repair studies conducted fungi or mammalian systems support a homologous gap-repair mechanism (SZOSTAK *et al.* 1988; BOLLAG *et al.* 1989). If such a gap repair model were to be applied to *R2* integration, a preexisting *R2*/28S junction already in the rDNA locus would serve as a template for the attachment of the 5′ end of the *R2* element to the upstream 28S gene sequence. This seems unlikely since such a process would give rise to greater uniformity in sequence than what we have observed at the 5′ end of full-length elements and would suggest that truncated copies must rely on alternative mechanisms. Repairs of double-stranded breaks in higher eukaryotes can also occur by nonhomologous means (ROTH and WILSON 1988; THACKER *et al.* 1992; LUKACSOVICH *et al.* 1994). However, such repairs are based on short sequence repeats and usually involve the deletion of nucleotides. The attachment of *R2* sequences at the 5′ end do not involve short repeats and frequently result in nucleotide additions. However, given the many uncertainties in the cellular repair of higher eukaryotes, particularly those involving an RNA:DNA heteroduplex (Figure 5), more general DNA repair models to complete *R2* retrotransposition remain possible.

The model for the attachment of the *R2* sequence to the upstream 28S gene sequences we suggest as most supported by the 5′ junctions is model E. This model requires that the *R2* reverse transcriptase jump from the *R2* RNA template to the top strand of the upstream 28S gene sequence. A similar model has been proposed for the retrotransposition of the *I* element (BUCHETON 1990). One attraction of this model is that it is consistent with the known template switching properties of reverse transcriptases (WHITCOMB and HUGHES 1992). Thus the switch from RNA to a DNA template could be viewed as similar to the template switching that occurs during first- and second-strand DNA synthesis in the retroviral retrotransposition mechanism. Model E is also consistent with the *in vitro* properties of the *R2* reverse transcriptase. We have found that the *R2* reverse transcriptase can readily switch to another RNA molecule when it reaches the end of the *R2* template (D. D. LUAN and T. H. EICKBUSH, unpublished data). Unfortunately we have not been able to demonstrate that this switch can occur efficiently to the upstream 28S target sequence.

The template switch model can explain most features of the 5′ ends of *R2* elements. First, it can explain the addition of a few apparently nontemplated nucleotides at the 5′ junctions. The addition of nontemplated bases has been shown to occur when retroviral reverse transcriptases switch template either *in vitro* or *in vivo* (ZHANG and TEMIN 1993, 1994; WU *et al.* 1995). The addition of nontemplated nucleotides by the *R2* reverse transcriptase has also been seen during the priming of synthesis from the target DNA (LUAN and EICKBUSH 1995) and during template switching to another RNA molecule (D. D. LUAN and T. H. EICKBUSH, unpublished data). Second, the model can explain why full-length and 5′ truncated elements have similar junctions. Retroviral reverse transcriptases undergo template switches at pauses associated with RNA secondary structure or the enzyme reaching the end of the RNA template (HU and TEMIN 1990; DESTEFANO *et al.* 1994). Truncated *R2* elements could be the result of either premature template switching or the use of a degraded *R2* RNA molecule as template (an RNA that does not extend to the 5′ end of the element). Third, the template switch model can explain the deletion of 28S gene sequences at the 5′ end. All junctions would have a 2-bp deletion because the top target DNA strand does not contain these two nucleotides. Larger deletions of the 28S sequence could result from template switchs to more upstream positions on the target DNA. Finally, template switching can explain why some 5′ junctions

have a short sequence of 28S gene repeated. Duplications can be explained if the reverse transcriptase attempts to switch to the target DNA template, falls off, and then must reattach to the target DNA. Model E cannot explain (nor can the other models) the tandem duplication of *R2* sequences at the 5′ junction.

Can the template switching model for the attachment of element sequences to the upstream target site also be applied to other non-LTR elements? Unlike R2, most non-LTR element insertions result in target site duplications. These duplications are typically variable in length in the case of elements that do not have sequence specific integration sites and are usually precise in length for those elements that do have sequence-specific integration sites (EICKBUSH 1992). This difference from *R2* can be explained if cleavage of the second target DNA strand by these elements (*i.e.*, the strand not used to prime reverse transcription) takes place downstream of the first strand cleavage, rather than upstream as it does with the *R2* element. Second-strand cleavage downstream of the first strand releases a single-stranded 3′ overhang onto which template switching can occur. The *R2* endonuclease may be unusual in that it cuts 2 bp upstream from the first cleavage, resulting in a recessed 3′ end that may be more difficult to support a template switch. Uniform template switching to the end of an exposed 3′ overhang would give rise to a target site duplication of precise length. Template switching to variable locations within the 3′ overhang would give rise to variable length target site duplications. While continued analysis of the 5′ and 3′ junctions of non-LTR elements can provide indirect support for or against the template switching model, direct characterization of the template switching properties of the reverse transcriptase from *R2* and other non-LTR elements are ultimately needed.

This work was supported by a National Institutes of Health grant (GM-42790) to T.H.E. We thank DANNA EICKBUSH for comments on the manuscript and ROBERT BAMBARA and CHRIS LAWRENCE for helpful discussions.

LITERATURE CITED

BOLLAG, R. J., A. S. WALDMAN and R. M. LISKAY, 1989 Homologous recombination in mammalian cells. Annu. Rev. Genet. **23:** 199–225.

BUCHETON, A. 1990 I transposable elements and I-R hybrid dysgenesis in *Drosophila*. Trends Genet. **6:** 16–21.

BURKE, W. D., C. C. CALALANG and T. H. EICKBUSH, 1987 The site-specific ribosomal insertion element type II of *Bombyx mori* (*R2*Bm) contains the coding sequence for a reverse transcriptase-like enzyme. Mol. Cell. Biol. **7:** 2221–2230.

BURKE, W. D., F. MÜLLER and T. H. EICKBUSH, 1995 R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. Nucleic Acids Res. **23:** 4628–4634.

CLARK, J. M, 1988 Novel non-templated nucleotide addition reactions catalyzed by prokaryotic and eukaryotic DNA polymerases. Nucleic Acids Res. **16:** 9677–9686.

DAWID, I. B., and M. L. REBBERT, 1981 Nucleotide sequence at the boundaries between gene and insertion regions in the rDNA of *D. melanogaster*. Nucleic Acids Res. **9:** 5011–5020.

DESTEFANO, J. J., R. A. BAMBARA and P. J. FAY, 1994 The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates. J. Biol. Chem. **269:** 161–168.

EICKBUSH, D. G., and T. H. EICKBUSH, 1995 Vertical transmission of the retrotransposable elements *R1* and *R2* during the evolution of the *Drosophila melanogaster* species subgroup. Genetics **139:** 671–684.

EICKBUSH, D. G., W. C. LATHE III, M. P. FRANCINO and T. H. EICKBUSH, 1995 *R1* and *R2* retrotransposable elements of *Drosophila* evolve at rates similar to those of nuclear genes. Genetics **139:** 685–695.

EICKBUSH, T. H., 1992 The non-LTR class of retrotransposable elements. New Biol. **4:** 430–440.

EICKBUSH, T. E, 1994 Origin and evolutionary relationships of retroelements, pp. 121–157 in *The Evolutionary Biology of Viruses*, edited by S. S. MORSE. Raven Press, New York.

EVANS, J. P., and R. D. PALMITER, 1991 Retrotransposition of a mouse L1 element. Proc. Natl. Acad. Sci. USA **88:** 8792–8795.

FINNEGAN, D. J., 1989 The I factor and I-R hybrid dysgenesis in *Drosophila melanogaster*, pp. 503–517 in *Mobile DNA*, edited by D. H. BERG and M. M. HOWE. American Society of Microbiology, Washington, DC.

HU, W.-S., and H. M. TEMIN, 1990 Retroviral recombination and reverse transcription. Science **250:** 1227–1233.

HUTCHISON, C. A., S. C. HARDIES, D. D. LOEB, W. R. SHEHEE and M. H. EDGELL, 1989 LINEs and related retroposons: long interspersed repeated sequences in the eukaryotic gemone, pp. 539–617 in *Mobile DNA*, edited by D. H. BERG and M. M. HOWE. American Society of Microbiology, Washington, DC.

INOUYE, S., and M. INOUYE, 1993 Bacterial reverse transcriptase, pp. 391–410 in *Reverse Transcriptase*, edited by S. GOFF and A. SKALKA. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

JAKUBCZAK, J. L., Y. XIONG and T. H. EICKBUSH, 1990 Type I (*R1*) and Type II (*R2*) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. J. Mol. Biol. **212:** 37–52.

JAKUBCZAK, J. L., M. K. ZENNI, R. C. WOODRUFF and T. H. EICKBUSH, 1992 Turnover of *R1* (Type I) and *R2* (Type II) retrotransposable elements in the ribosomal DNA of *Drosophila melanogaster*. Genetics **131:** 129–142.

JENSEN, S., and T. HEIDMANN, 1991 An indicator gene for detection of germline retrotransposition in transgenic *Drosophila* demonstrates RNA-mediated transposition of the LINE 1 element. EMBO J. **10:** 1927–1937.

MARTIN, S. L., 1991 LINEs. Curr. Opin. Genet. Dev. **1:** 505–508.

MCCLURE, M. A, 1993 Evolutionary history of reverse transcriptase, pp. 425–444 in *Reverse Transcriptase*, edited by S. GOFF and A. SKALKA. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

LAMBOWITZ, A. M., and M. BELFORT, 1993 Introns as mobile genetic elements. Annu. Rev. Biochem. **62:** 587–622.

LEVIS, R. W., R. GANESAN, K. HOUTCHENS, L. A. TOLAR and F. M. SHEEN, 1993 Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell. **75:** 1083–1093.

LONG, E. O., M. L. REBBERT and I. B. DAWID, 1980 Structure and expression of ribosomal RNA genes of *Drosophila melanogaster* interrupted by type-2 insertions. Cold Spring Harbor Symp. Quant. Biol. **45:** 667–672.

LUAN, D. D., and T. H. EICKBUSH, 1995 RNA template requirements for target DNA-primed reverse transcription by the *R2* retrotransposable element. Mol. Cell. Biol. **15:** 3882–3891.

LUAN, D. D., M. H. KORMAN, J. L. JAKUBCZAK and T. H. EICKBUSH, 1993 Reverse transcription of *R2*Bm RNA is primed by a nick at the chromosomal target site—A mechanism for non-LTR retrotransposition. Cell **72:** 595–605.

LUKACSOVICH, T., D. YANG and A. S. WALDMAN, 1994 Repair of a specific double-strand break generated within a mammalian chromosome by yeast endonuclease I-*Sce*I. Nucleic Acids Res. **22:** 5649–5657.

MARTIN, F., C. MARANON, M. OLIVARES, C. ALONSO and M. C. LOPEZ, 1995 Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi:* homology of

the first ORF with the Ape family of DNA repair enzymes. J. Mol. Biol. **247:** 49–59.

PEACOCK, W. J., R. APPELS, S. ENDOW and D. GLOVER, 1981 Chromosomal distribution of the major insert in *D. melanogaster* 28S rRNA gene. Genet. Res. **37:** 209–214.

PELISSON, A., D. J. FINNEGAN and A. BUCHETON, 1991 Evidence for retrotransposition of the I factor, a LINE element of *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **88:** 4907–4910.

ROIHA, H., J. R. MILLER, L. C. WOODS and D. M. GLOVER, 1981 Arrangements and rearrangements of sequence flanking the two types of rDNA insertion in *D. melanogaster.* Nature **290:** 749–753.

ROTH, D. B., and J. H. WILSON, 1986 Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. Mol. Cell. Biol. **6:** 4295–4304.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SCHWARZ-SOMMER, Z., L. LECLERCQ, E. GOBEL and H. SAEDLER, 1987 Cin4, an insert altering the stucture of the A1 gene in *Zea mays,* exhibits properties of nonviral retrotransposons. EMBO J. **6;** 3873–3880.

SZOSTAK, J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN and F. W. STAHL, 1983 The double-strand-break repair model of recombination. Cell **33:** 25–35.

THACKER, J., J. CHALK, A. GANESH and P. NORTH, 1992 A mechanism for deletion formation in DNA by human cell extracts: the involvement of short sequence repeats. Nucleic Acids Res. **20:** 6183–6188.

WHITCOMB, J. M., and S. H. HUGHES, 1992 Retroviral reverse transcription and integration: process and problems. Annu Rev Cell Biol. **8:** 275–306.

WU, W. M., B. M. BLUMBERG, P. J. FAY and R. A. BAMBARA, 1995 Strand transfer mediated by human immunodeficiency virus reverse transcriptase *in vitro* is promoted by pausing and results in misincorporation. J. Biol. Chem. **270:** 325–332.

XIONG, Y., and T. H. EICKBUSH, 1988a Similarity of reverse transcriptase-like sequences of viruses,transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5:** 675–690.

XIONG, Y., and T. H. EICKBUSH, 1988b Functional expression of a sequence-specific endonuclease encoded by the retrotransposon *R2*Bm. Cell **55:** 235–246.

XIONG, Y., and T. H. EICKBUSH, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:** 3353–3362.

ZHANG, J., and H. M. TEMIN, 1993 Rate and mechanism of nonhomologous recombination during a single cycle of retroviral replication. Science **259:** 234–238.

ZHANG, J., and H. M. TEMIN, 1994 Retroviral recombination depends on the length of sequence identity and is not error prone. J. Virol. **68:** 2409–2414.

ZIMMERLY, S., H. GUO, P. S. PERLMAN and A. M. LAMBOWITZ, 1995 Group II intron mobility occurs by target DNA-primed reverse transcription. Cell **82:** 1–10.

Communicating editor: N. L. CRAIG