Completion of RLE LINE Integration Involves an Open "4-Way" Branched DNA Intermediate

By

BRIJESH BIKRAM KHADGI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
May, 2020

Arlington, Texas

Supervising Committee:

Dr. Shawn M. Christensen, Supervising Professor
Dr. Clay Clark
Dr. Esther Betran
Dr. Matthew K. Fujita
Dr. Michael R. Roner

ABSTRACT

Long Interspersed Elements (LINEs), also known as non-LTR retrotransposons, encode a multifunctional protein that reverse transcribes its mRNA into DNA at the site of insertion by target primed reverse transcription. The R2 Long Interspersed Elements (LINEs) specifically integrate in the 28S rRNA genes by a series of DNA binding, DNA cleavage, and DNA synthesis reactions. While the first half of the integration reaction, TPRT, is well understood, the second half of the integration reaction, second-strand DNA cleavage and second-strand DNA synthesis are much less well understood. A hitherto unknown and unexplored branched integration intermediate, an open '4-way' DNA junction which is thoguht to arise by template jumping, was recognized by the element protein and cleaved in a Holliday junction resolvase-like reaction. Cleavage of the branched integration intermediate resulted in a natural primer-template pairing used for second-strand DNA synthesis. In addition, the structure of the branched integration intermediate itself was explored by probing with DNase I footprint and was found to be highly structured. R2 protein binding to the junction was explored by a combination of DNA cleavage assays and DNA footprint studies. The protein appears to bind in a sequence specific manner to the downstream sequence of branched integration intermediates, but less so for the upstream sequence where structure appears to be more important. A new model for RLE LINE integration is presented.

ACKNOWLEDGMENTS

I would like to humbly thank my principal investigator, Dr. Shawn M. Christensen for entrusting me with his important research projects and for giving me guidance, mentorship and motivation throughout my graduate career. I thank Dr. Esther Betran for her consistent support and assistance throughout my graduate career whether it be during the courses that I had taken with her or outside the class. I also thank Dr. Fujita for his advising and motivation for carrying on with my projects and making me understand that failures are just a part of one's success. I thank Dr. Roner for giving me valuable advices and challenging me to excel in my endeavors. I thank Dr. Clark for providing me with valuable information that not only helped me hone my critical thinking but also taught me to be prepared for unforeseen circumstances. Lastly, I would also like to thank Dr. Fondon, who had been one of my committee members early on during my graduate career. I would also like to thank Dr. Kimberly Bowles for supporting me and guiding me not just as an advisor but also as a friend. I am grateful to my peers - past (Aruna Govindaraju, Monika Pradhan, Murshida Mahbub, Jeremy Cortez, Eyad Shihabeddin, Corayma Hernandez, Amy Everett and Thy Tran) and present (Santosh Dhamala, Joydeep Chatterjee, Shalini Rachakonda and Hasan El Monem) lab members. Thank you to each one of you for helping me out throughout; you all shall always be my treasured lab family. I thank the Department of Biology at University of Texas at Arlington and all the staff members for helping and supporting me throughout my stay here as a graduate student. Finally, I am thankful to Phi Sigma Biological Sciences Honor Society at UTA for the funding. Last, but not the least, I would like to thank my parents, my sisters and my dearest wife for always being there, supporting me, guiding me, encouraging me and most of all, loving and believing in me for all these years.

TABLE OF CONTENTS

ABSTRACT……..I

ACKNOWLEGMENTS……II

CHAPTER 1……..1

CHAPTER 2……..21

CHAPTER 3……..58

CHAPTER 4……..86

CHAPTER 1

Non-LTR Retrotransposons: An Overview of LINE Integration Mechanism

Brijesh B. Khadgi and Shawn M. Christensen

Department of Biology, University of Texas at Arlington

Arlington, TX 76019, USA

INTRODUCTION

Transposable elements (TEs) are selfish mobile genetic elements that replicate in host genomes and are thus inherited along with the host chromosomes. Transposable elements are widely distributed among all major taxonomic groups, including fungi, animal, plants, and protozoa and have a profound effect on structure and function of their host genome since the replication and abundancy of these elements often result in insertions, deletions, and recombination events. TEs can also serve as a source of novel genetic material resulting in new genes and regulatory sequences for the host [1] [2] [3]. Retrotransposable elements, also called Class I TEs, transpose through an RNA intermediate. The RNA is reverse transcribed into DNA either before or during integration into the genome. As such, all autonomous retrotransposons encode a reverse transcriptase. The reverse transcriptase has been used to generate cladograms and to assist in the classification retrotransposable elements (Figure 1A). Elements that integrate after reverse transcription encode an integrase (e.g., LTR-retrotransposons and retroviruses) or a DNA recombinase (e.g., DIRS) [4] [5]. These elements have a "copy-out/paste-in" replication mechanism where the element is copied out of the genome by transcription and pasted back in the host genome as double stranded DNA [6]. Elements that integrate during reverse transcription do so by target-primed-reverse-transcription (TPRT) [7]. Target-primed retrotransposons (also called non-LTR retrotransposons) encode a DNA endonuclease which functions as a DNA nickase, cleaving one DNA strand at a time at the site of insertion [7] [8] [9] [10] [4] [5]. The reverse transcriptase uses the free 3'-OH generated by the DNA endonuclease to prime reverse transcription of the element RNA [7]. Target-primed retrotransposons thus use a copy-out/copy-in mechanism of replication.

A major group of target-primed retrotransposons are the long-interspersed-elements (LINEs). LINEs encode either an apurinic-apyrimidinic DNA endonuclease (APE) or a restriction-like DNA endonuclease (RLE) (Figure 1A, 1B). RLE LINEs are considered to be the more ancient of the two (APE vs RLE LINEs) [11] [12]. The RLE LINEs are generally about 3-4 kb in length with a single open reading frame (ORF) (Figure 1B). The single ORF encodes from one to three zinc finger (ZF) motifs, and sometimes a Myb motif, in the N-terminal region, a central reverse transcriptase (RT), a linker region with a IAP/gag-like zinc knuckle, and a PD(D/E)xK-family DNA-endonuclease with its restriction-endonuclease-like fold located after the linker [13] [14] [15] [16] [8]. RLE LINEs are generally site-specific and target multicopy genes of the host (e.g, the ribosomal locus) [15] [11]. There are at least five main groups of RLE LINEs: R2, R4, CRE, NeSL, and HERO [11] [17] [18] [19] [20] [12]. The R2 group is the most studied and include members that specifically insert into the 28S (R2 or R8 site) or into the 18S (R8 site) rRNA genes [21] [22].

R2 group consists of four clades (R2-A, R2-B, R2-C, and R2-D) based on their RT sequences [23] [21] [24]. R2 elements from each clade can have specific number of zinc-fingers, located in their N-terminal region along with a single myb domain [25] [23] [24]. R2-A has three zinc-fingers (CCHH, CCHC, CCHH). R2-B has two zinc-fingers (CCHC, CCHH). R2-C has two zinc-fingers, both CCHH. R2-D has a single CCHH zinc-finger. The number of N-terminal zinc-finger motifs is shown to be consistent across the phylogeny of R2 and therefore, suggests that the common ancestor of R2 had three zinc-finger motifs at the N-terminal end [23]. The R2-A clade is thought to represent the ancestral clade [21, 23 24]. R2s are widely distributed in animal phyla including Arthropoda, Nematoda, Chordata, Echinodermata, Platyhelminthes, and Cnidaria [26] [23] [21] [27] [28]. They have also been reported from hagfishes, cyclostomes, coelacanth, actinopterygian fish, reptiles, and birds [23] [29] [24] [30]. R2 families in Ctenophora, Mollusca, and Hemichordata has also been

recently discovered [23] [30]. R2 group elements have been being vertically transferred since prior to divergence of deuterostomes and protostomes [23].

APE LINEs are about 7 kb in length and encode two ORFs (Figure 1B). The first open reading frame encodes an RNA binding protein that is quite diverse between different clades of APE LINEs. The second open reading frame of APE LINEs is analogous to the single ORF of RLE LINEs as the second ORF of APE LINEs encodes the DNA endonuclease (APE), the reverse transcriptase (RT), and the linker with its IAP/gag knuckle-like motif. APE LINEs are generally not site-specific during integration, although several clades of site-specific APE LINEs are well known, for example Tx1 clade and R1 clades are site-specific [11] [31] [2] [3] [32] [11] [33] [34]. There 20 plus clades of APE LINEs [35] [12].

The replication life cycle and integration mechanism of RLE LINEs and APE LINEs is functionally very similar. The replication of LINEs generally occurs during germline production [12] [36] [37]. An RNA transcript is generated from an element coded promoter or a promotor located upstream of the inserted element (Figure 1B, 1C). The transcript is exported to the cytoplasm and translated. The element encoded protein(s) bind to the transcript from which the protein was translated from, a process termed cis preference, to form an integration-competent ribonucleoprotein (RNP) complex. The RNP enters the nucleus. Integration at a new site occurs by TPRT where the element encoded endonuclease generates a nick at a chromosomal site and the reverse transcriptase uses the free 3'-OH to prime reverse transcription of the element RNA (i.e., TPRT). Completion of integration has remained largely undetermined but is thought to involve second strand DNA cleavage and second strand DNA synthesis events carried out by either element or host factors. This first chapter of my dissertation will briefly review what is known about the replication cycle of LINEs, with special focus on RLE LINEs and in particular, R2

elements. My data chapters, Chapters 2 and 3, will explore the second half of the integration reaction of LINEs, second-strand cleavage and second-strand synthesis, using purified components of the RLE LINE R2 from *Bombyx mori* (R2Bm). R2 and R2Bm has served as a major model system for the biochemistry RLE LINE integration. The mammalian L1 element has served as one of the major elements used to study APE LINE integration. In the remainder of this background chapter, I will focus on what is known about R2 transcription, translation, RNP formation, and integration. I will include brief parallels or tie-ins to APE LINEs, especially L1.

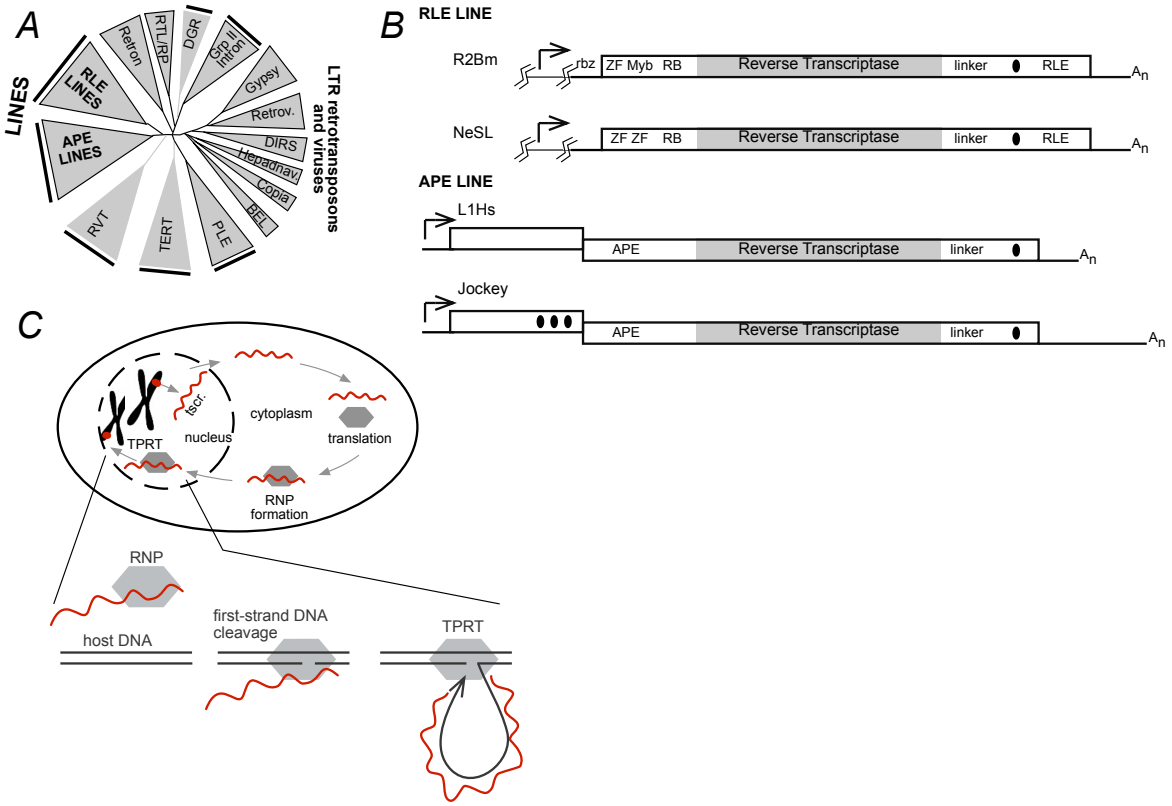


Figure 1. (A) RT tree of life phylogram. The phylogram is adapted from data in [38]. Gray triangles with black perimeters are reverse transcriptases from retrotransposon and viral elements. Gray tringles are host reverse transcriptase genes. Straight black lines are retrotransposons (or host systems) that are either known to, or assumed to, undergo integrate by TPRT. **(B)** RLE and APE LINE ORF structure as rectangles with major motifs indicated. Two representative RLE LINEs and two representative APE LINEs are shown so as to indicate that there is great diversity within these two groups. Black ovals are IAP/gag-like zinc knuckle motifs. Arrows are promoters. Abbreviations: zinc finger (ZF), RNA binding domain (RB), restriction endonuclease-like (RLE), and apurinic-apyrimidinic endonuclease (APE), ribozyme (rbz) **(C)** Replication cycle and basic TPRT mechanism.

Transcription of R2 and processing of the element RNA

RLE LINEs (R2). Most RLE LINEs are site specific and target multi-copy gene sequences and the promotor of the host gene is used to generate a co-transcript. This use of a host promotor to generate a co-transcript has been shown most conclusively shown for the R2 element in Drosophila. R2 elements target the ribosomal locus. The R2 element in *Drosophila*, as in many other species, targets the 28S ribosomal DNA (Figure 2A). Normally the ribosomal promotor generates a long transcript that is processed into the individual ribosomal RNAs by host factors. Ribosomal units with an R2 element inserted into the 28S rDNA, are transcribed, generating a co-transcript that includes the R2 element [39, 40 41].

The R2 element encodes an HDV-like ribozyme located at the 5' end of the element which processes the element away from the majority of the upstream ribosomal sequences (Figure 2A, 2B) [39, 40 42, 43]. *In vitro* studies using synthesized RNAs have shown that the HDV-like ribozyme in several Drosophila species, along with several non Drosophia species, are capable of rapid and efficient self-cleavage of 28S/R2 co-transcript. The HDV ribozyme found in *Drosophila* and other insect R2 elements are around 180 nt in length with a large J1/2 loop [39]. Interestingly, and importantly, the P1 region often includes ribosomal derived RNA and the cleavage site of the ribozyme is within the ribosomal RNA [39, 40]. The self-cleavage site in most R2 elements has been shown to be within the 28S rRNA gene from 9 to 36 nucleotides upstream of the R2 5' junction or the insertion site [39]. The cleavage site of R2 ribozyme from *Bombyx mori*, for instance, is located about 28 bp upstream of the R2 insertion site [40 39]. The site of self-cleavage for the *Drosophila simulans* ribozyme, however, is precisely at the 28S/R2 5' junction [39 15].

R2 elements with ribozymes that are predicted to cleave within the 28S sequences tend to generate 5' junctions with fewer small target deletions and/or nucleotide additions upon insertion than do elements whose ribozymes cleave at the 28S/R2 site [39]. R2 elements with ribozymes that are predicted to cleave within the 28S sequences also sometimes generate 5' junctions that contain tandem duplications of upstream 28S sequences of a length consistent with the location of the predicted ribozyme cleavage site [39]. For this reason, it has been hypothesized that the rRNA sequence not cleaved off by the ribosome, might play a crucial role in second-strand synthesis [39].

Other RLE LINEs appear to encode HDV-like ribozymes (e.g., R4) and might be standard for site-specific elements that do not encode their own promotors [44] [42]. It is not known how the 3' end of the R2 element is determined or processed.

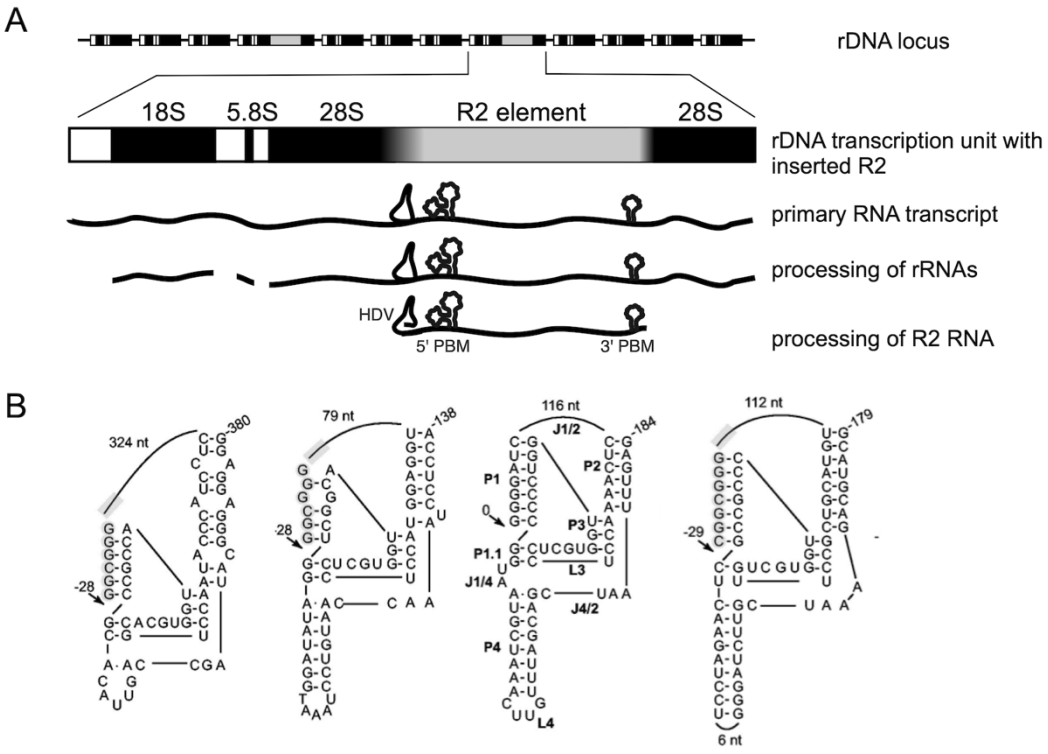


Figure 2. Transcription of ribosomal unit with an R2 insertion. (A) ribosomal rDNA unit with an R2 element insertion it. Primary transcript is initiated at the promotor for the ribosomal unit. The primary transcript is processed into the rRNAs and the R2 RNA. **(B)** R2 HDV-like ribozymes and their cleavage sites, modified from [39].

APE LINEs. APE LINEs generally have internal pol II promoter that are used to transcribe the element RNA [45] [46]. The 5' UTR of human L1 has been shown, or hypothesized, to contain binding sites for a number of transcription factors including Ying Yang-1 (YY1), Runx3, SRY-related (SOX) proteins, and Sp1 [47] [45] [48] [49].

Some APE LINEs appear to encode a HDV-like ribozyme including SART, R6Ag1/3, RTE, and L1Tc, Ingi elements [44] [42] [44]. The ribozymes of L1Tc and Ingi also code for internal promotors [44, 50].

Translation of R2

RLE LINEs (R2). R2 elements lack 5'methyl guanosine cap as the transcript is derived from a pol I transcript and then further processed by the HDV-like ribozyme. R2, therefore, must initiate translation through a cap independent mechanism. Conservation of RNA structure dominates in much of the 5' UTR of R2 because of the constraints of the HDV-like ribozyme. In the ORF, conservation of amino-acid sequences dominate over RNA structure. There is an area of overlap, where RNA structure and the start of the ORF appear to be linked [39, 40, 51, 52]. The complex double pseudoknot structure of self-cleaving HDV ribozyme has been hypothesized to function as an internal-ribosome-entry-site (IRES) similar to the pseudoknot based IRESs found in viruses and a few cellular mRNAs [39, 40, 51, 52]. Indeed, the HDV-like ribozymes of several *Drosophila* R2 elements, when hooked up to a luciferase ORF, appeared to be able to act as an IRES in *in vitro* transcription/translation studies [42].

APE LINEs. APE LINE elements have two ORFs, each of which are required to be translated independently. APE LINEs are generally transcribed Pol II and the mRNA capped. ORF 1 is translated in a cap-dependent manner. ORF 2 translation occurs by either re-initiation or by an IREs. In human L1, two in-frame stop codons at the end of ORF 1 are crucial for termination of ORF1 translation and re-initiation of ORF2 translation. Once the ribosome receives a stop signal, the ORF1 protein is released followed by followed by partial dissociation of the ribosome. The small subunit of ribosome remains associated to L1 RNA transcript and scans the remaining inter-ORF region to reinitiate at the AUG of ORF2 [53]. Both ORFs are translated such that the resulting protein is not ORF1/ORF2 fusion protein [53] [34]. The mouse L1 uses an IRES to initiat translation of ORF2 [53] [54]. In unconventional translation mechanism, ribosome scans the ORF 1 region and keeps translating the ORF1 protein until it gets to the stop codon at the "inter-ORF" region. Other elements such as SART1 element from silkworms exhibit translational coupling mechanism as in some prokaryotes and viruses, where the same ribosome translates both the ORF1 and ORF2 proteins. Also, SART1 element does not require AUG start codon, and rather an overlapping UAAUG stop-start codons and downstream RNA secondary structure are shown to be very important for efficient translation of the ORF2 [55].

R2 RNP formation

RLE LINEs (R2). The R2 protein from *Bombyx mori* (R2Bm) binds to a structured segment of element RNA located in the 5' UTR termed the 5' protein-binding-motif (PBM) [56]. In R2Bm, and other related moths, the 5' PBM is also the hypothesized location of the IRES [52]. The moth 5' UTR is abnormally long, compared to a canonical R2 5' UTR (e.g., *Drosophila*). The

moth 5' UTA consists of a HDV-like ribozyme and the 5' PBM [40, 51, 52]. *Drosophila* R2 elements the 5' UTR is much shorter and consists of just the HDV-like ribozyme. The *Drosophila* HDV-like ribozyme likely functions as the ribozyme, the IRES, and the 5' PBM. It is possible that the moth lineage underwent a duplication and subfunctionalization of the 5' UTR.

The 3' UTR of the R2 transcript also has a conserved secondary structure and binds R2 protein, therefor the R2 3' UTR functions as a PBM and has been called the 3' PBM [7, 51, 57, 58]. Protein bound to the 3' PBM is necessary and sufficient for TPRT (i.e, the first half of the integration reaction) [7, 58]. It is interesting to note that the R2 protein from *Bombyx mori* (R2Bm) is capable of recognizing the 3' UTR RNA from *Drosophila melanogaster* (R2Dm) as well as other distantly related R2 elements, despite having no sequence similarity to the R2Bm 3' PBM [58 57]. Binding of R2 protein and 3' UTR of the given element, is therefore, suggested to be dependent on the secondary or tertiary structures RNA transcript [57]. The 3' UTR RNA from R2Bm and R2Dm are both thought to have secondary structure that include three helical regions with the sequence AAC/UAUC in the loop generated by one of the helices in the structure and this conserved region of the transcript has been shown to be critical in binding of the R2 protein [59].

For full integration, it is thought that the R2 RNP consists of two subunits of R2 protein, one bound to the 3' PBM and one bound to the 5' PBM of the transcript (Figure 3) [56].

APE LINEs. APE LINEs (e.g., mammalian L1) encode two ORFs [33]. The protein encoded by ORF1 has been implicated in element RNA binding [60]. The 3' UTR, especially the poly-A tail is required for RNP formation and integration [61]. The protein encoded by ORF two is thought to bind to the poly-A tail [61].

The R2Bm integration reaction

First strand cleavage and first strand synthesis (TPRT)

Insertion mechanism of R2 protein is thought to involve two protein subunits (Figure 3) [56, 62]. In the presence of 3' PBM RNA, the R2 protein subunit binds upstream of the insertion site. Footprint data has shown R2 protein to be bound between -40 to -20 bp upstream of the target DNA cleavage site [62 63]. Most RLE LINE endonucleases are capable of binding to the target DNA at a distance from the actual cleavage site similar to type IIs restriction endonucleases [8]. R2 is also thought to be capable of making contacts at regions of target DNA, away from actual cleavage site. However, R2 protein domains that happen to be important in DNA binding still remain unidentified. The upstream R2 protein subunit bound to target DNA nicks the first strand (i.e. anti-sense strand of the target DNA) and releases a 3'-hydroxyl which is then used by the reverse transcriptase as a primer to prime reverse transcription of the first-strand cDNA synthesis [62]. R2 protein utilizing 3' PBM RNA as a template and a 3'-OH of the nicked target DNA as a primer to do first-strand synthesis is termed as TPRT. R2 protein is known to bind to target DNA upstream of the insertion site in the presence of any non-specific RNAs, however, only those protein subunits in the presence of 3' PBM RNA is capable of TPRT [7]. Most efficient TPRT occurs when the RNA template for reverse transcription has its 3' end analogous and precise to the boundary of R2Bm element [64]. The templates with polyadenylated at the 3' end (about 8 nt) or with truncated 3' end (about 3-6 nt) do not show much efficiency during integration reaction [58].

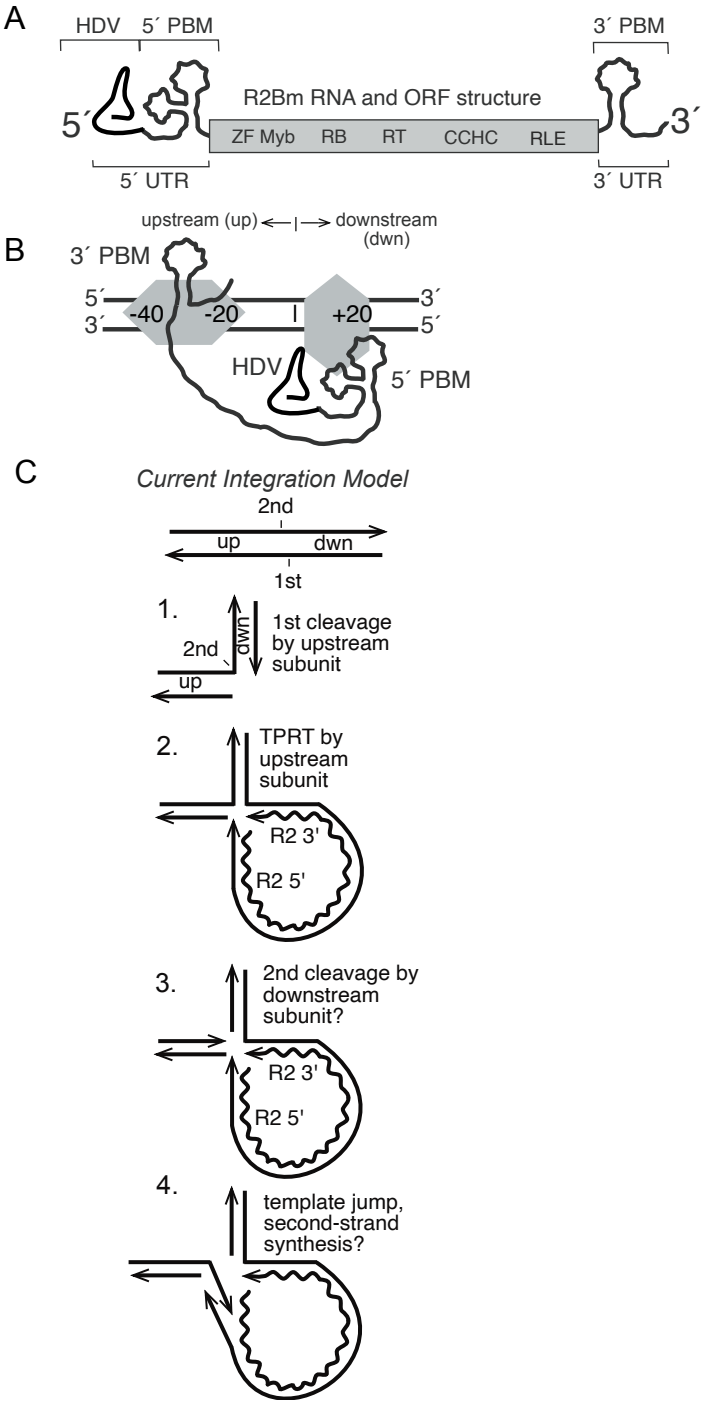


Figure 3. Integration reaction of R2Bm. (A) The R2 RNA and ORF structure. **(B)** R2Bm RNP bound to target 28S rDNA. The black parallel lines represent the segment of 28S rDNA containing the insertion site (black vertical line). An R2 protein subunit (gray horizontal hexagon) is bound upstream of the insertion site (vertical bar), and an R2 protein (gray vertical hexagon) subunit is bound downstream of the insertion site. The upstream subunit is associated with the 3 ′ PBM RNA, and the downstream subunit is associated with the 5 ′ PBM RNA. The footprints of the two protein subunits on the linear target DNA are indicated. The upstream subunit footprints from −40 bp to −20 bp, but it grows to just over the insertion site (vertical line) after first-strand DNA cleavage. The downstream subunit footprints from just prior to the insertion site to +20 bp [65] [56]. **(C)** Diagram of the current integration reaction: (1) cleavage of the 28S rDNA antisense-strand by the upstream bound RNP, (2) TPRT by the upstream bound protein subunit, (3) second strand cleavage, presumably by the downstream subunit, and (4) second strand DNA synthesis by an unknown mechanism which may involve template jump and/or microhomologies.

Second strand cleavage and second strand synthesis

Little is known about the second half of the integration reaction. In the presence of 5' PBM RNA, the R2 protein subunit binds downstream of the insertion site. The endonuclease of downstream subunit makes a second-strand cleavage of the top strand (i.e. sense strand) which occurs 2 bp upstream relative to the first strand cleavage site. Second strand cleavage only occurs after the 5' PBM RNA bound to downstream subunit is removed [56]. The sequence space of the first and second strand cleavage differ. It is unclear how the RLE manages to recognize and cleave two different sites (i.e. first-strand cleavage and second-strand cleavage sites) on the target. In addition, it is unknown as to whether first-strand DNA cleavage is a prerequisite for the second-strand DNA cleavage. The R2 endonuclease can cleave single-stranded DNA (ssDNA) adjacent to duplex DNA, perhaps this activity has something to do with second-strand DNA cleavage. Second-strand cleavage is thought to be "non-site specific" and therefore could possibly make a cleavage on a ssDNA-duplex DNA junction that are usually formed because of local denaturation after first-strand cleavage [66]. After the second-strand cleavage, a 3'- hydroxyl is generated which is hypothesized to be used as a for second-strand synthesis, completing the integration reaction. The reverse transcriptase of downstream subunit is hypothesized to be involved in second-strand synthesis [62] [56].

R2 derived short-internally-deleted-elements (SIDEs)

Several *Drosophila* species have non-autonomous sequences called SIDEs that encode the R2 self-cleaving ribozyme at their 5' end to process themselves from 28S rRNA co-transcript and include sequences with identity to the 3' UTR of R2 that play crucial role in their recognition and

initiation of reverse transcription by R2 machinery [67]. R2 SIDEs and R2/R1 hybrid SIDEs, for example, hijack autonomous R2 retrotranspositon machinery and are facilitated by the high rates of recombination events and yield in a given rDNA locus enabling them to move and therefore, survive within the host genome [67 68 69 70 71].

Integration mechanism of APE LINEs (L1)

APE-bearing LINE-1 contains the endonuclease domain at N-terminal to the RT in ORF2. ORF2p encoded EN is shown to interact with DNA through the Bβ6-Bβ5 loop upon which the EN nicks the target DNA at a degenerate consensus sequence, 5'-TTTT/A-3' or variants of this sequence [9 72]. Following the RNP formation and entry to the nucleus, LINE-1 endonuclease cleaves the first/bottom-strand (i.e. antisense strand), releases a 3' -hydroxyl which is then used by reverse transcriptase as a primer to prime reverse transcription of the element RNA, and thus complete first-strand cDNA synthesis [2]. The first-strand synthesis generally starts within the 3' end poly-A sequence of LINE-1 RNA [73]. The 3' poly-A sequence in LINE-1 is known to be crucial for efficient retrotransposition. While replacing the 3'-end poly A sequence of LINE-1 with non-polyadenylated non-coding RNA does not affect translation, it does result in the RNA being unable to retrotranspose [74]. In addition, retrotransposition of LINE-1 is directly affected by the length of poly-A sequence. Addition of about 20-26 poly-A tract downstream of the LINE-1 element, drastically increases the retrotransposition activity [74]. Unexpectedly, poly-A sequence is known to be not as important for LINE-1 retrotransposition in cultured cells [61].

Also, annealing of LINE-1 cDNA to the second/top-strand (i.e. sense strand) of the target DNA is hypothesized to specify the placement of second-strand DNA cleavage. Second-strand DNA cleavage then generates a 3'-hydroxyl needed for second strand cDNA synthesis [75]. Usually, second/top-strand cleavage occurs at variable distances (i.e. within 15 to 16 bp) downstream of the first/bottom-strand cleavage site. When second/top-strand cleavages either upstream or downstream of the first/bottom-strand cleavage is compared to the cleavages occurring at the same site on both strands (i.e. double stranded nicks), only downstream cleavages are shown to result in target site duplications (TSDs) as observed in human genome reference sequence [76] [75].

Most retrotransposition events are known to initiate at LINE-1 EN consensus cleavage site, 5'-TTTT/A-3'. However, although LINE-1s happen to be flanked by canonical TSDs, it is unclear as to whether there exists a strict consensus cleavage site for the second/top-strand cleavage. Nonetheless, analysis of inversion/deletion and inversion/duplication events in LINE-1s has shown a weak preference to the sequence 5' –TYTN/R [77]. Weak specificity for second/top-strand cleavage is known to occur when retrotransposition events are formed by "twin priming" [75].

Moreover, LINE-1 EN has also been suggested to exhibit sequence preference for second/bottom-strand cleavage activity, but relaxed or no sequence specificity for first/top-strand cleavage [75]. In addition to difference in the sequences for first/bottom and second/top strand DNA cleavages, the locations for either of the cleavages are also hypothesized to be at a distance from each other and therefore obscuring the second part of the integration reaction [75]. Finally, it has also been hypothesized that the LINE-1 may either encode a second nuclease activity that plays an important role in second/top-strand cleavage or the host factors themselves are being involved in the cleavage [73] [75].

REFERENCES

1. Craig,N.L. (2002) Tn7 In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 423-456.
2. Richardson,S.R., Doucet,A.J., Kopera,H.C., Moldovan,J.B., Garcia-Perez,J.L. and Moran,J.V. (2015) The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061.
3. Zingler,N., Weichenrieder,O. and Schumann,G.G. (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res* **110**, 250-268.
4. Bourque,G., Burns,K.H., Gehring,M., Gorbunova,V., Seluanov,A., Hammell,M., Imbeault,M., Izsvák,Z., Levin,H.L., Macfarlan,T.S., Mager,D.L. and Feschotte,C. (2018) Ten things you should know about transposable elements. *Genome Biol* **19**, 199.
5. Arkhipova,I.R. (2017) Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* **8**, 19.
6. Curcio,M.J. and Derbyshire,K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**, 865-877.
7. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605.
8. Yang,J., Malik,H.S. and Eickbush,T.H. (1999) Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A* **96**, 7847-7852.
9. Feng,Q., Moran,J.V., Kazazian,H.H.J. and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916.
10. Guo,H., Zimmerly,S., Perlman,P.S. and Lambowitz,A.M. (1997) Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J* **16**, 6835-6848.
11. Fujiwara,H. (2014) Site-specific non-LTR retrotransposons. *Microbiol Spectrum* **3**, MDNA3-0001.
12. Malik,H.S., Burke,W.D. and Eickbush,T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**, 793-805.
13. Mahbub,M.M., Chowdhury,S.M. and Christensen,S.M. (2017) Globular domain structure and function of restriction-like-endonuclease LINEs: similarities to eukaryotic splicing factor Prp8. *Mob DNA* **8**, 16.
14. Govindaraju,A., Cortez,J.D., Reveal,B. and Christensen,S.M. (2016) Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287.
15. Eickbush,T.H. and Eickbush,D.G. (2015) Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011.

16. Eickbush,T.H. (2002) R2 and Related Site-Specific Non-Long Terminal Repeat Retrotransposons In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 813-835.
17. Eickbush,T.H. and Jamburuthugoda,V.K. (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res*
18. Burke,W.D., Malik,H.S., Rich,S.M. and Eickbush,T.H. (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. *Mol Biol Evol* **19**, 619-630.
19. Kojima,K.K. and Jurka,J. (2015) Ancient Origin of the U2 Small Nuclear RNA Gene-Targeting Non-LTR Retrotransposons Utopia. *PLoS One* **10**, e0140084.
20. Kojima,K.K. (2018) Human transposable elements in Repbase: genomic footprints from fish to humans. *Mob DNA* **9**, 2.
21. Kojima,K.K., Kuma,K., Toh,H. and Fujiwara,H. (2006) Identification of rDNA-specific non-LTR retrotransposons in Cnidaria. *Mol Biol Evol* **23**, 1984-1993.
22. Gladyshev,E.A. and Arkhipova,I.R. (2009) Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* **448**, 145-150.
23. Kojima,K.K. and Fujiwara,H. (2005) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165.
24. Luchetti,A. and Mantovani,B. (2013) Non-LTR R2 element evolutionary patterns: phylogenetic incongruences, rapid radiation and the maintenance of multiple lineages. *PLoS One* **8**, e57076.
25. Burke,W.D., Malik,H.S., Jones,J.P. and Eickbush,T.H. (1999) The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol Biol Evol* **16**, 502-511.
26. Jakubczak,J.L., Burke,W.D. and Eickbush,T.H. (1991) Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci U S A* **88**, 3295-3299.
27. Kapitonov,V.V. and Jurka,J. (2014) A family of HERO non-LTR retrotransposons from the Californian leech genome *Repbase Reports* **14**, 311.
28. Kojima,K.K. and Fujiwara,H. (2004) Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Mol Biol Evol* **21**, 207-217.
29. Kapitonov,V.V. and Jurka,J. (2009) R2 non-LTR retrotransposons in the bird genome *Repbase Rep* **9**, 1329.
30. Kojima,K.K., Seto,Y. and Fujiwara,H. (2016) The Wide Distribution and Change of Target Specificity of R2 Non-LTR Retrotransposons in Animals. *PLoS One* **11**, e0163496.
31. Craig,N.L., Chandler,M., Gellert,M., Lambowitz,A.M., Rice,P.A. and Sandmeyer,S.B. (2015) American Society for Microbiology (ASM).
32. Han,J.S. (2010) Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA* **1**, 15.

33. Moran,J.V. and Gilbert,N. (2002) Mammalian LINE-1 Retrotransposons and Related Elements In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 836-869.
34. Babushok,D.V. and Kazazian,H.H. (2007) Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28**, 527-539.
35. Kapitonov,V.V., Tempel,S. and Jurka,J. (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207-213.
36. Eickbush,T.H. and Malik,H.S. (2002) Origins and Evolution of Retrotransposons In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 1111-1146.
37. Crichton,J.H., Dunican,D.S., Maclennan,M., Meehan,R.R. and Adams,I.R. (2013) Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. *Cell Mol Life Sci*
38. Gladyshev,E.A. and Arkhipova,I.R. (2011) A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* **108**, 20311-20316.
39. Eickbush,D.G., Burke,W.D. and Eickbush,T.H. (2013) Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441.
40. Eickbush,D.G. and Eickbush,T.H. (2010) R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol*
41. Ye,J. and Eickbush,T.H. (2006) Chromatin structure and transcription of the R1- and R2-inserted rRNA genes of Drosophila melanogaster. *Mol Cell Biol* **26**, 8781-8790.
42. Ruminski,D.J., Webb,C.H., Riccitelli,N.J. and Lupták,A. (2011) Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes. *J Biol Chem* **286**, 41286-41295.
43. Webb,C.H., Riccitelli,N.J., Ruminski,D.J. and Luptak,A. (2009) Widespread occurrence of self-cleaving ribozymes. *Science* **326**, 953.
44. Sanchez-Luque,F.J., Lopez,M.C., Macias,F., Alonso,C. and Thomas,M.C. (2011) Identification of an hepatitis delta virus-like ribozyme at the mRNA 5'-end of the L1Tc retrotransposon from Trypanosoma cruzi. *Nucleic Acids Res*
45. Swergold,G.D. (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**, 6718-6729.
46. Dmitriev,S.E., Andreev,D.E., Terenin,I.M., Olovnikov,I.A., Prassolov,V.S., Merrick,W.C. and Shatsky,I.N. (2007) Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. *Mol Cell Biol* **27**, 4685-4697.
47. Athanikar,J.N., Badge,R.M. and Moran,J.V. (2004) A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* **32**, 3846-3855.
48. Tchenio,T., Casella,J.F. and Heidmann,T. (2000) Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* **28**, 411-415.
49. Yang,N., Zhang,L., Zhang,Y. and Kazazian,H.H.J. (2003) An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* **31**, 4929-4940.

50. Yang,N., Zhang,L., Zhang,Y. and Kazazian,H.H.J. (2003) An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* **31**, 4929-4940.
; Heras,S.R., Lopez,M.C., Olivares,M. and Thomas,M.C. (2007) The L1Tc non-LTR retrotransposon of Trypanosoma cruzi contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. *Nucleic Acids Res* **35**, 2199-2214.
51. Moss,W.N., Eickbush,D.G., Lopez,M.J., Eickbush,T.H. and Turner,D.H. (2011) The R2 retrotransposon RNA families. *RNA Biol* **8**,
52. Kierzek,E., Christensen,S.M., Eickbush,T.H., Kierzek,R., Turner,D.H. and Moss,W.N. (2009) Secondary structures for 5' regions of R2 retrotransposon RNAs reveal a novel conserved pseudoknot and regions that evolve under different constraints. *J Mol Biol* **390**, 428-442.
53. Alisch,R.S., Garcia-Perez,J.L., Muotri,A.R., Gage,F.H. and Moran,J.V. (2006) Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* **20**, 210-224.
54. Li,P.W., Li,J., Timmerman,S.L., Krushel,L.A. and Martin,S.L. (2006) The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. *Nucleic Acids Res* **34**, 853-864.
55. Kojima,K.K., Matsumoto,T. and Fujiwara,H. (2005) Eukaryotic translational coupling in UAAUG stop-start codons for the bicistronic RNA translation of the non-long terminal repeat retrotransposon SART1. *Mol Cell Biol* **25**, 7675-7686.
56. Christensen,S.M., Ye,J. and Eickbush,T.H. (2006) RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* **103**, 17602-17607.
57. Ruschak,A.M., Mathews,D.H., Bibillo,A., Spinelli,S.L., Childs,J.L., Eickbush,T.H. and Turner,D.H. (2004) Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* **10**, 978-987.
58. Luan,D.D. and Eickbush,T.H. (1995) RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* **15**, 3882-3891.
59. Mathews,D.H., Banerjee,A.R., Luan,D.D., Eickbush,T.H. and Turner,D.H. (1997) Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3**, 1-16.
60. Martin,S.L. and Bushman,F.D. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**, 467-475.
61. Moran,J.V., Holmes,S.E., Naas,T.P., DeBerardinis,R.J., Boeke,J.D. and Kazazian,H.H.J. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927.
62. Christensen,S.M. and Eickbush,T.H. (2005) R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**, 6617-6628.

63. Christensen,S. and Eickbush,T.H. (2004) Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* **336**, 1035-1045.
64. Luan,D.D. and Eickbush,T.H. (1996) Downstream 28S gene sequences on the RNA template affect the choice of primer and the accuracy of initiation by the R2 reverse transcriptase. *Mol Cell Biol* **16**, 4726-4734.
65. Christensen,S.M., Bibillo,A. and Eickbush,T.H. (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468.
66. Kurzynska-Kokorniak,A., Jamburuthugoda,V.K., Bibillo,A. and Eickbush,T.H. (2007) DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *J Mol Biol* **374**, 322-333.
67. Eickbush,D.G. and Eickbush,T.H. (2012) R2 and R2/R1 hybrid non-autonomous retrotransposons derived by internal deletions of full-length elements. *Mob DNA* **3**, 10.
68. Eickbush,D.G. and Eickbush,T.H. (1995) Vertical transmission of the retrotransposable elements R1 and R2 during the evolution of the Drosophila melanogaster species subgroup. *Genetics* **139**, 671-684.
69. Lathe,W.C. and Eickbush,T.H. (1997) A single lineage of r2 retrotransposable elements is an active, evolutionarily stable component of the Drosophila rDNA locus. *Mol Biol Evol* **14**, 1232-1241.
70. Gentile,K.L., Burke,W.D. and Eickbush,T.H. (2001) Multiple lineages of R1 retrotransposable elements can coexist in the rDNA loci of Drosophila. *Mol Biol Evol* **18**, 235-245.
71. Eickbush,D.G., Ye,J., Zhang,X., Burke,W.D. and Eickbush,T.H. (2008) Epigenetic Regulation of Retrotransposons within the Nucleolus of Drosophila. *Mol Cell Biol*
72. Weichenrieder,O., Repanas,K. and Perrakis,A. (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986.
73. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910.
74. Doucet,A.J., Wilusz,J.E., Miyoshi,T., Liu,Y. and Moran,J.V. (2015) A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell* **60**, 728-741.
75. Gilbert,N., Lutz,S., Morrish,T.A. and Moran,J.V. (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-7795.
76. Gilbert,N., Lutz-Prigge,S. and Moran,J.V. (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315-325.
77. Jurka,J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94**, 1872-1877.

CHAPTER 2

Completion of LINE Integration Involves an open "4-way" Branched DNA Intermediate

Brijesh B. Khadgi, Aruna Govindaraju and Shawn M. Christensen

Department of Biology, University of Texas at Arlington

Arlington, TX 76019, USA

8708–8719 *Nucleic Acids Research, 2019, Vol. 47, No. 16* *Published online 8 August 2019*
doi: 10.1093/nar/gkz673

Completion of LINE integration involves an open '4-way' branched DNA intermediate

Brijesh B. Khadgi, Aruna Govindaraju and Shawn M. Christensen*

Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA

Received December 07, 2018; Revised June 26, 2019; Editorial Decision July 20, 2019; Accepted July 29, 2019

All authors of this published paper has given consent for republishing it in my dissertation.

Individual author roles are as follows:

Shawn M. Christensen: Principal Investigator; responsible for overall conceptual input and discussion of exeprimental data for the paper.

Aruna Govindaraju: Put forth Initial ideas and early experimental data; designed oligos for linear, 3-way and 4-way DNA junctions and helped in establishing the foundation of the paper.

Brijesh B. Khadgi: With the help/input from PI (Dr. Shawn M. Christesen), performed all the major experiments (binding, cleavage assasys and synthesis assays); designed DNA integration intermediates/oligos and responsible for all the major data figures of this paper. Analyzed most data and involved in writing/editing paper. Responsbile for all the supplementary data figures.

ABSTRACT

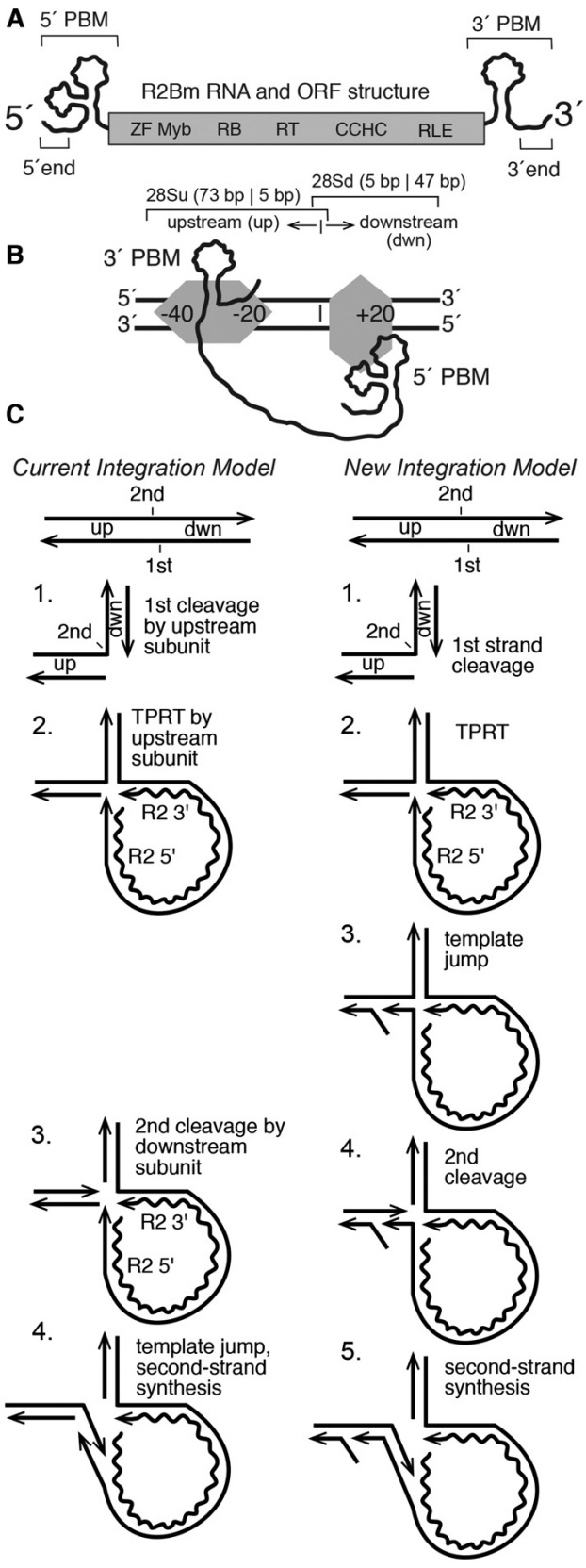
Long Interspersed Elements (LINEs), also known as non-LTR retrotransposons, encode a multifunctional protein that reverse transcribes its mRNA into DNA at the site of insertion by target primed reverse transcription. The second half of the integration reaction remains very poorly understood. Second-strand DNA cleavage and second-strand DNA synthesis were investigated in vitro using purified components from a site-specific restriction-like endonuclease (RLE) bearing LINE. DNA structure was shown to be a critical component of second-strand DNA cleavage. A hitherto unknown and unexplored integration intermediate, an open '4-way' DNA junction, was recognized by the element protein and cleaved in a Holliday junction resolvase-like reaction. Cleavage of the 4-way junction resulted in a natural primer-template pairing used for second-strand DNA synthesis. A new model for RLE LINE integration is presented.

INTRODUCTION

Long interspersed elements (LINEs) are an abundant and diverse group of autonomous transposable elements (TEs) that are found in eukaryotic genomes across the tree of life. LINEs also mobilize the nonautonomous short interspersed elements (SINEs) which appropriate the protein machinery of LINEs to replicate. The movements of LINEs and SINEs have been implicated in genome evolution, modulation of gene expression, genome rearrangements, DNA repair, cancer progression, and as a source of new genes (1,2). LINEs replicate by a process called target primed reverse transcription (TPRT), where the element RNA is reverse transcribed into DNA at the site of insertion using a nick in the target DNA to prime reverse transcription (3–5). LINEs encode protein(s) that are used to perform the critical steps of the insertion reaction. LINE proteins bind their own mRNA, recognize target DNA, perform first-strand target–DNA cleavage

and perform TPRT. The proteins are also hypothesized to perform second-strand target–DNA cleavage and second-strand element-DNA synthesis, although the evidence for these is sparse (3–20). The earlybranching clades of LINEs encode a restriction-like endonuclease (RLE), while the later-branching LINEs encode an apurinic-apyrimidinic DNA endonuclease (APE) (2124). Both types of elements are thought to integrate through a functionally equivalent integration process (5,25–27).

Second-strand DNA cleavage has remained unclear because the cleavage sites generally are not palindromic: the sequence around the second-strand cleavage site is often unrelated to the sequence around the first-strand cleavage site. In addition, blunt or staggered cleavages can occur. The staggered cleavages give rise to target site duplications or target site deletions depending on whether the staggered cut is 3 ′ overhanging or 5 ′ overhanging, respectively. Moreover, the staggered cleavages can be a few bases away (e.g. 2 bp in R2Bm) or quite distant (e.g. 126 bp in R9) (28,29). In APE LINEs, as in RLE LINEs, the cleavages are generally staggered such as to generate a modest 10–20 bp target site duplication upon insertion (26,30–32). The endonuclease from APE-bearing LINEs (APE LINEs) appears to have some specificity for the first DNA cleavage site, but much less so for the second DNA cleavage site on linear target DNA (23,30,31,33,34). The endonuclease from the RLEbearing LINEs (RLE LINEs) is similarly involved in target site recognition (11). In both cases, however, additional specifiers for cleavage have been hypothesized to account for the different specificity of the first and second-strand DNA cleavages including the endonuclease being tethered to the DNA by unidentified DNA binding domains in the protein. Another complicating factor is that the first cleavage event should occur in the presence of element RNA, while the second cleavage event, according to a priori reasoning,



should occur in the absence of element RNA, due to cDNA formation, however cleavage in the absence of RNA has been difficult to demonstrate in vitro (20).

Figure 1. R2Bm structure and integration reaction. (A) R2Bm RNA (wavy line) and open reading frame (ORF) structure (gray box). The ORF encodes conserved domains of known and unknown functions: zinc finger (ZF), Myb (Myb), reverse transcriptase domain (RT), a cysteinehistidine rich motif (CCHC) and a PD-(D/E)XK type RLE. RNA structures present in the 5 ′ and 3 ′ untranslated regions that bind R2 protein are marked as 5 ′ and 3 ′ PBMs, respectively. The small (25 nt) RNA segments from the 5 ′ end and 3 ′ ends of the element RNA used in this study are indicated. (B) The R2 integration complex, as currently understood, is depicted bound to a segment of linear 28S rDNA (black parallel lines). An R2 protein subunit (gray horizontal hexagon) is bound upstream of the insertion site (vertical bar), and an R2 protein (gray vertical hexagon) subunit is bound downstream of the insertion site. The upstream subunit is associated with the 3 ′ PBM RNA, and the downstream subunit is associated with the 5 ′ PBM RNA. The footprints of the two protein subunits on the linear target DNA are indicated. The upstream subunit footprints from −40 bp to −20 bp, but it grows to just over the insertion site (vertical line) after first-strand DNA cleavage. The downstream subunit footprints from just prior to the insertion site to +20 bp (10,20). The overlapping portions of the target DNA, 28Su and 28Sd, used in this study are indicated with brackets. (C) The Current Integration Model and the New Integration Model being proposed in this paper are compared. Straight lines are DNA (28S or R2). The wavy line is the R2 RNA. The four steps of the current model are: (1) DNA cleavage of the bottom/first-strand of the target DNA; (2) TPRT; (3) DNA cleavage of the top/second strand of the target DNA; and (4) second-strand DNA synthesis. The fourth step has not been observed directly in vitro. The five steps of the new model are: (1) DNA cleavage of the bottom/first-strand of the target DNA; (2) TPRT; (3) a template jump/recombination event that generates an open ‘4-way’ DNA junction; (4) second-strand DNA cleavage; and (5) second-strand DNA synthesis. Abbreviations: up (target sequences upstream of the insertion site), dwn (target sequences downstream of the insertion site) and TPRT.

Second-strand DNA synthesis has remained unresolved since TPRT was first described over 20 years ago, and it has never been directly observed in vitro (4,15,25,35,36). Second-strand synthesis (SSS) is hypothesized to be primed off the free 3′-OH generated by the second-strand cleavage event and synthesized by the element-encoded reverse transcriptase (RT). It is unknown how the proposed primer-template association is generated as the ends of the double-stranded cleaved target DNA drift away post cleavage in in vitro reactions (6,20).

The R2 element from Bombyx mori, R2Bm, is one of a number of model systems that has been used to study the insertion reaction of LINEs (27). R2 elements are site specific, targeting the 'R2 site' in the 28S rRNA gene (27). The R2 element encodes a single open reading frame with Nterminal zinc finger(s) (ZF) and Myb domains (Myb), a central reverse transcriptase (RT), an RLE and a C-terminal gag-knuckle-like CCHC motif (Figure 1A). The R2Bm protein has been expressed in Escherichia coli and purified for use in in vitro reactions.

In vitro studies of the R2Bm protein and RNA have contributed to the current model of integration for R2Bm (Figure 1B and C) (20). Two subunits of R2 protein, one bound to the 3′ protein binding motif (PBM) of the R2 RNA and other to the 5′ PBM of the R2 RNA, are thought to be involved in the integration reaction. The 5′ and 3′ PBM RNAs dictate the roles of the two subunits and coordinate a series of DNA cleavage and polymerization steps, resulting in element integration by TPRT (Figure 1A). The protein subunit bound to the element's 3′ PBM interacts with 28S rDNA sequences upstream of the R2 insertion site. The upstream subunit's RLE cleaves the first (bottom/antisense) DNA strand. After first-strand target-DNA cleavage, the subunit's RT performs TPRT using the 3′-OH generated by the cleavage event to prime first-strand cDNA synthesis. The protein subunit bound to the 5′ PBM RNA interacts with 28S rDNA sequences downstream of the R2 insertion site by way of the ZF and Myb domains. The downstream subunit's

RLE cleaves the second (top/sense) DNA strand, but only after the 5′ PBM RNA structure is destroyed by TPRT during cDNA formation, putting the protein in the minus RNA state. Second-strand DNA cleavage, however, is not thought to occur until after the 5′ PBM RNA is pulled from the subunit, presumably by the process of TPRT, putting the protein in a 'no RNA bound' conformation. Confusingly, second-strand DNA cleavage does not readily occur in the absence of RNA in our in vitro reactions. Second-strand cleavage had, until this report, required a narrow range of R2 protein, 5′ PBM RNA and target DNA ratios to be observed in in vitro reactions (20). Additionally, second-strand cleavage had, until this report, disconnected the primer for SSS, the 3′-OH generated by second-strand cleavage, from the cleavage event from the cDNA template, making initiation of second-strand DNA synthesis problematic (6,20).

The DNA endonuclease plays a central role in the integration reaction of LINEs. The RLE found in the earlybranching LINEs is a variant of the PD-(D/E)XK superfamily of endonucleases (11,22). In a previous paper, we reported the similarity of the LINE RLE as having sequence and structural homology to archaeal Holliday junction resolvases (11,37). Our previous paper left open the question as to whether R2 protein could function on branched DNA molecules and what this potential activity tells us about the insertion reaction. Of particular interest is the TPRT product, a pseudo (i.e. open) '3-way' junction, and a proposed open '4-way' junction. The open 3- and 4-way junctions are key substrates that differentiate two models of insertion (Figure 1C): (i) The Current Integration Model, and (ii) a New Integration Model being proposed herein. The two models differ in the timing and substrate of second-strand DNA cleavage. Cleavage of the TPRT product (Current Integration Model) produces a fully cleaved target DNA with no obvious primer-template from which to prime secondstrand DNA synthesis; it is proposed that a template jump occurs post DNA cleavage in order to prime SSS. In the New Integration Model, the proposed

template jump/switch occurs prior to second-strand DNA cleavage and thus forms the 4-way-like junction. The open 4-way junction, upon DNA cleavage, resolves into a natural primer-template that could be used in second-strand DNA synthesis.

MATERIALS AND METHODS

Nucleic acid preparation and R2Bm protein purification

Oligonucleotides (oligos) containing 28S R2 target DNA, non-target (nonspecific) DNA and R2 sequences were ordered from Sigma-Aldrich. The upstream (28Su) and downstream (28Sd) target DNA designations are relative to the R2Bm insertion dyad within the 28S rRNA gene. The DNA constructs were formed by annealing the component oligos: see Supplementary Figure S1 for a list of the oligos used in this study and their sequences. One of the component oligos had been 5 ′ end-labeled (32 P), prior to annealing to the other component oligos. Twenty pmol of the radiolabeled oligo was mixed with 66 pmol of each of the other oligos that make up the construct. The oligos were annealed in 1× TPRT buffer (10 mM Tris–HCl (pH 8.0), 5 mM MgCl 2 , 200 mM NaCl) for 2 min at 95 ◦ C, followed by 10 min at 65 ◦ C, 10 min at 37 ◦ C and at last 10 min at room temperature. The constructs were not further purified post annealing as the procedure of gel purification led to inadvertent formation of partial junctions and gave us less control over DNA concentration. Junctions that shared a common labeled oligo were equalized by radioactive DNA counts; otherwise, equal volumes annealed junctions were used in R2 reactions.

R2Bm protein expression and purification were carried out for wild-type R2 protein, endonuclease mutant (KPD/A or K/ARNKY) and reverse transcriptase mutant (YAD/YD) as previously published (11,22). Briefly, E. coli BL21 cells containing the R2 expression plasmid were grown in LB broth and induced with IPTG. An empty expression vector was used to generate

the protein extract that served as the mock-protein (øprotein) negative control in the functional assays. The induced cells were pelleted by centrifugation, resuspended and gently lysed in a HEPES buffer containing lysozyme and triton X-100. The cellular DNA and debris were spun down, and the supernatant containing the R2Bm protein was purified over Talon resin (Clontech #635501). The R2Bm protein was eluted from the Talon resin column and stored in protein storage buffer containing 50 mM HEPES pH 7.5, 100 mM NaCl, 50% glycerol, 0.1% triton X-100, 0.1 mg/ml bovine serum albumin (BSA) and 2 mM dithiothreitol (DTT) and stored at −20 ◦ C. R2 protein was quantified by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) along with a BSA standard titration and stained by SYPRO Orange (Sigma #S5692) prior to addition of BSA to the R2 protein for storage. All quantitations were done using FIJI software analysis on digital photographs (38).

DNA binding, DNA cleavage and DNA synthesis reactions

R2Bm protein and target DNA binding, DNA cleavage and DNA synthesis reactions were performed largely as previously reported (11). Reactions were 13 l and contained 80 fmol of labeled substrate DNA, 10-fold excess cold competitor DNA (dIdC) by mass, and a dilution series of R2Bm protein, typically ≥ 420 - ≤ 0.40 fmol protein. Each DNA construct was tested for its ability to bind to purified R2Bm protein and to undergo DNA cleavage in the absence of RNA (i.e. in the absence of 5 ′ PBM RNA and 3 ′ PBM RNA). The reactions were analyzed by native 5% polyacrylamide gel electrophoresis (EMSA) to determine fraction bound and denaturing (8 M urea) 8% polyacrylamide gels to determine fraction cleaved. A+G ladders as well as ladders made from different sized DNA oligos were run alongside the reactions in the denaturing urea gels to aid in mapping cleavages. Oligos used to build the constructs were used to also make the end

labeled DNA oligo ladder and the A+G ladder. Only reactions in the linear range on a bound versus cleaved graph were used in determining cleavability, and then only from 20% bound to about 95% bound window as quantitation is problematic below and above that range. An SSS assay was performed by the addition of dNTPs to the DNA cleavage reactions. All gels were dried, exposed to a phosphorimager screen and scanned using a phosphorimager (Molecular dynamics STORM 840). The resulting 16-bit TIFF images were linearly adjusted (levels command) so that the most intense bands were dark gray. Adjusted TIFF files were quantified using FIJI software (38). Gel images presented in the main figures were adjusted (levels command) to visualize the cleaved and/or synthesized products of interest.

RESULTS

First-strand cleavage and TPRT products are poor substrates for second-strand cleavage

R2Bm inserts into a specific site in the 28S rDNA. In previous studies, it was determined that the protein subunit bound to target sequences downstream of the insertion site likely provides the endonuclease involved in second-strand (i.e. top-strand) DNA cleavage (6,10,20). Second-strand cleavage, however, has always been difficult to achieve and study. Previously, second-strand cleavage has required a narrow range of 5 ′ PBM RNA, R2 protein and DNA ratios. The prior data indicated that first-strand DNA cleavage is probably required before the second-strand can be cleaved, that the downstream subunit must be bound to the DNA (which required 5 ′ PBM RNA) and that the 5 ′ PBM RNA must then be dissociated from the downstream subunit for second-strand cleavage to occur (20). In vivo, with a full-length R2 RNA, the process of TPRT would be expected to pull the 5 ′ PBM RNA from the downstream subunit, putting the downstream subunit into the ‘no RNA bound’ state and thus initiating second-strand DNA cleavage.

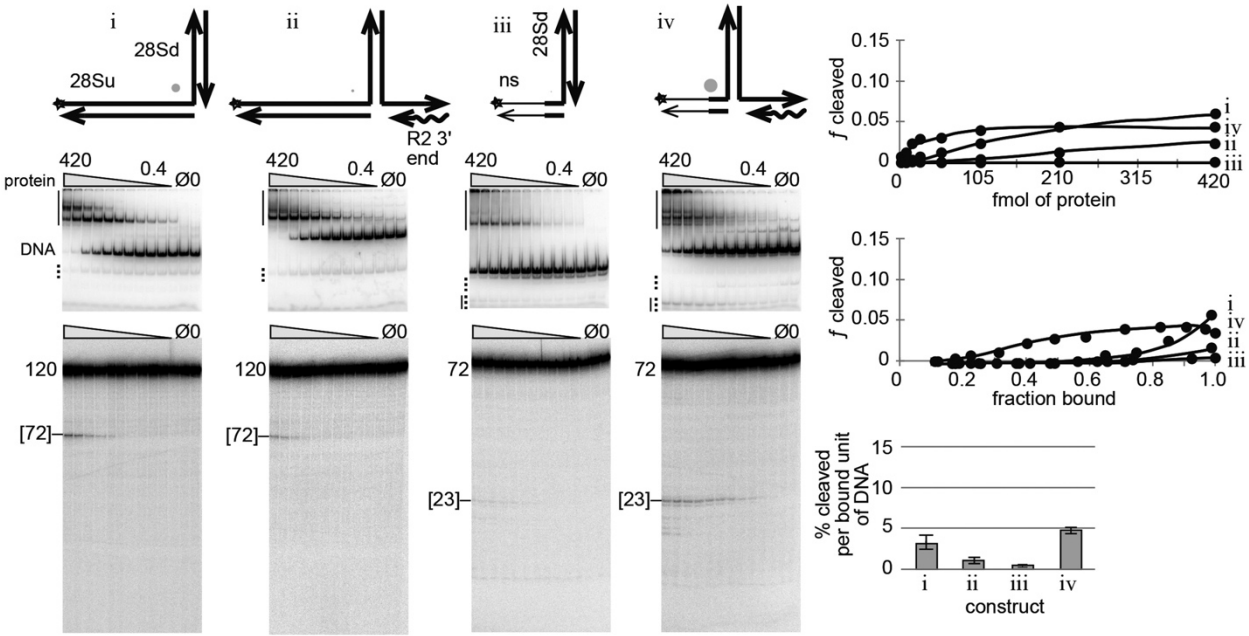


Figure 2. First-strand DNA cleavage and TPRT products are not good substrates for second-strand DNA cleavage. Several bottom/first-strand nicked linear DNAs (i and iii) and TPRT analogs (ii and iv) were tested for cleavability by the R2Bm protein. The 120 bp nicked 28S DNA (i) is diagrammatically bent at a 90 ◦ angle with the downstream (dwn) oriented toward the top of the page (i.e. the North arm). The TPRT product (ii) is similarly drawn; the TPRT (i.e. the East arm) arm is 25 bp. The star indicates that the DNA strand was 5 ′ end-labeled to track DNA binding and cleavage. In constructs iii and iv, the thin lines represent non-specific sequences; the left West arm was 25 bp and only the 5 bp nearest the second-strand cleavage site remained 28S DNA. Below each of the construct cartoons are the native (EMSA) gels and corresponding denaturing gels used to analyze DNA binding (EMSA) and DNA cleavage (denaturing) of the given DNA construct by R2Bm protein. DNA binding and cleavage reactions were 13 l and contained 80 fmol of radiolabeled construct DNA and 420–0.4 fmol of R2Bm protein (gray triangle). All EMSA gels were quantified such that the bands above the full construct DNA in the mock purified protein (Ø) and no protein (0) control lanes were subtracted out of the bound signal in the experimental lanes. Solid vertical lines next to the EMSA gels represent areas of the gel where the bound DNA signal resides. The well, the smear and the gel migrating complexes were all counted as bound DNA. DNA bands located below unbound (free) DNA that increased with protein concentration were counted as bound DNA since these bands were released cleavage products. The released cleavage product co-migrated with partial junctions (dotted line) present in the control lanes. The control lane partial junction signal was subtracted from the experimental lane's co-migrating bound signal. The remaining partial junctions (dotted line) were counted as unbound DNA in the experimental lanes. The main band in the mock purified (Ø; protein purified from an empty expression vector) and the no protein (0) lanes is the location of the unbound junction DNA (DNA). Next to the denaturing gels is the size of the uncleaved radiolabeled oligo. The size and migration of the band resulting from second-strand cleavage is indicated by brackets on the denaturing gel. The DNA binding and DNA cleavage results are plotted on three graphs: (i) fraction (f) cleaved as a function of protein concentration (fmol/reaction); (ii) fraction cleaved as a function of fraction bound for reactions where roughly 20–95% of the DNA was bound; and (iii) a bar graph reporting the average percentage cleaved products per bound unit of DNA (fraction bound) for reactions in the linear part of the second graph. The diameter of the gray dot next to each construct cartoon reflects the relative cleavability of the construct normalized to construct v in the next figure. See Supplementary Figure S2 for a graph of fraction bound as a function of protein concentration and for endonuclease mutant R2 protein controls.

In the first part of this study, the ability of the R2 protein to perform second-strand cleavage in the 'no-RNA-bound' state was investigated on products generated from the first two steps of the insertion reaction: first-strand DNA cleavage and TPRT (Figure 2). The product formed as a result of first-strand cleavage was made by annealing a 120 bp 28S derived target DNA containing 73 nt of 28S sequence upstream of the R2 insertion site and 47 nt of sequence downstream of the insertion site to two oligonucleotides complementary to the upstream and downstream segments of the 120 mer, respectively (10,20). In the diagram of the cleaved linear DNA in Figure 2 (construct i), the DNA has been bent 90 ◦ with the downstream 28S DNA 'arm' being oriented upward ('North') and the upstream 28S DNA arm remaining oriented to the left ('West'). The TPRT product analog, construct ii, was similarly formed by annealing oligonucleotides. The TPRT analog included a 25 bp DNA/RNA heteroduplex arm derived from the 3 ′ end of the R2 element, positioned to the right ('East') in the diagram (3). The 120 nt 'top' (i.e. the sense) strand of the 28S gene was 5 ′ end-labeled with 32 P to facilitate tracking of R2Bm protein induced DNA cleavage events (i.e. second-strand cleavage events). An electrophoretic mobility shift assay (EMSA) was used to measure the ability of the R2 protein to bind to each construct across a range of protein concentrations. Companion denaturing polyacrylamide gels were used to assay for second-strand DNA cleavage. The DNA binding and DNA cleavage data were quantified and are presented in several graphs: fraction cleaved as a function of protein concentration, fraction cleaved as a function of fraction bound, and a bar graph reporting the average percent cleaved per bound unit of DNA (derived from the linear portion of the fraction cleaved as a function of fraction bound graph).

Neither the first-strand cleavage product (construct i) nor the TPRT analog (construct ii) were good substrates for second-strand cleavage (Figure 2). DNA cleavage only occurred at or

near protein excess, and even at these levels only a small percentage of the bound DNA was cleaved. This result is similar to the dynamics previously reported where it was not until the upstream protein binding site, located at −40 to −20 (see Figure 1B), was completely occupied did protein associate with the downstream DNA binding site resulting in cleavage of the sense strand of the 28S gene (6). And in the absence of RNA, as in the presence of 3 ′ RNA, the R2Bm binds to upstream DNA sequences (29).

In an effort to promote efficient second-strand cleavage, the upstream 28S DNA sequences of constructs i and ii were removed, forming constructs iii and iv. The upstream (West) arm of the two new constructs consisted of 25 bp of mostly nonspecific DNA; only the 5 bp prior to the secondstrand cleavage site remained 28S sequence. The shortened first strand cleavage product (iii) failed to undergo secondstrand cleavage. The TPRT product (iv) cleaved better than constructs i–iii and construct iv did not have the need for an excess of protein like i and ii. That said, only about 5% of the protein-bound constructs underwent second-strand DNA cleavage. Construct iv was still a poor substrate for second-strand cleavage.

Specific open ‘4-way’ junctions are cleaved by R2 protein

In the second part of this study, open ‘4-way’ junctions that mimic the template switch hypothesized to occur at the close of TPRT were generated (see construct cartoons in Figure 3; see also New Integration Model, Figure 1C). A template switch is the association between the cDNA and the target DNA and the potential extension of the cDNA using the target DNA as a template. The 5 ′ end of the R2Bm mRNA is believed to contain rRNA sequence corresponding to the upstream target DNA (35,39–41). The reverse transcribed cDNA could then hybridize to the top strand of the target to form the 4-way junction.

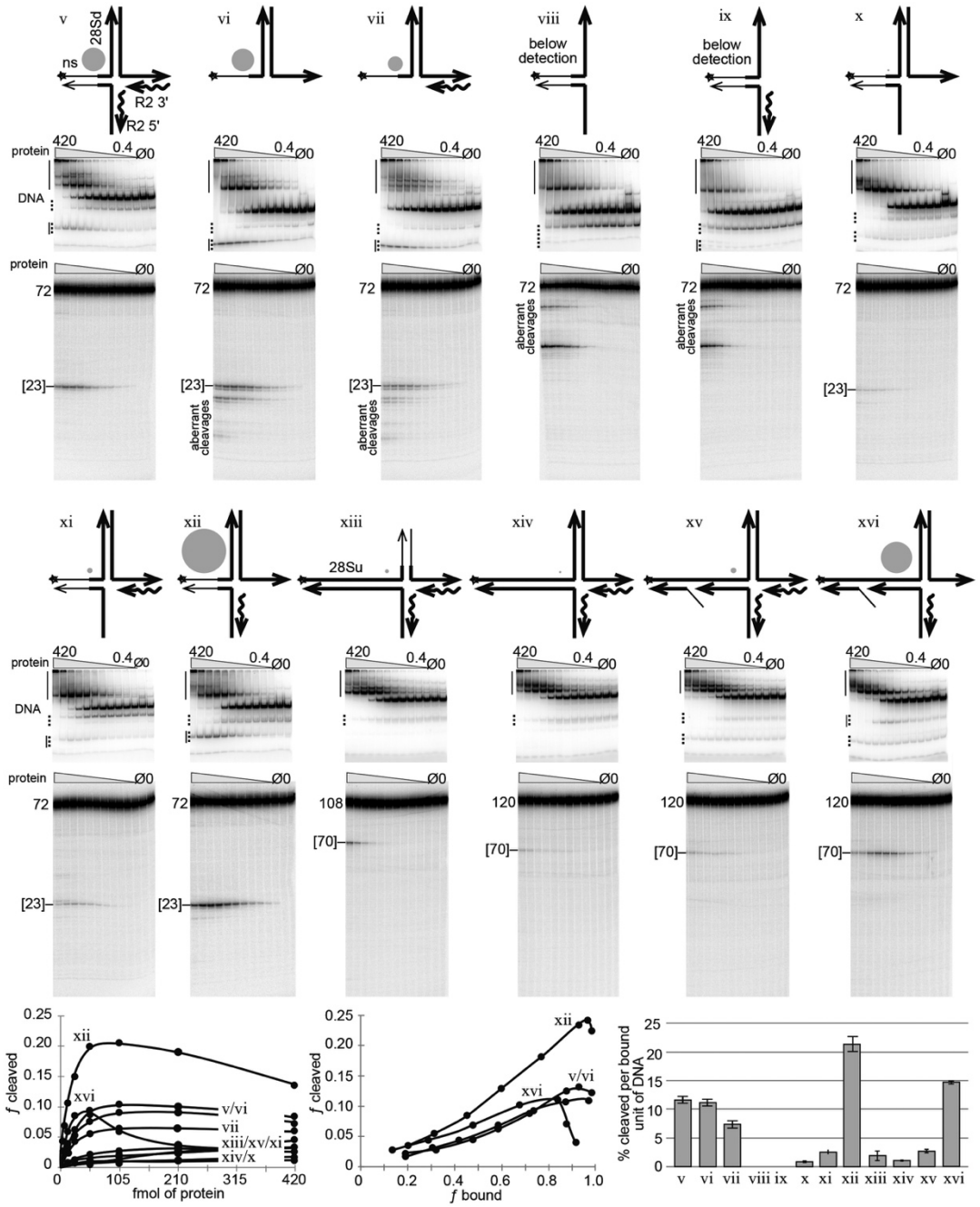


Figure 3. Specific open '4-way' junctions are good substrates for second-strand DNA cleavage. Various R2Bm/28S derived junctions related to the open 4-way junction drawn in step 3 of the New Integration Model (Figure 1) were tested for DNA cleavage. Symbols, conventions, reactions, gels, analysis and graphs are as in Figure 2. The North arms of the constructs contain 47 bp of 28S downstream DNA, which is the same amount of downstream 28S DNA normally used in our linear target DNA (10,20). In construct xiii the 47 bp North arm was replaced with a 35 bp arm of mostly non-specific DNA. The 5 bp nearest the cleavage site, however, remained 28S DNA. The West arm of constructs v–xii were identical to constructs iii and iv (Figure 2), being 25 bp in length and containing mostly nonspecific DNA. The West arms of constructs xiii–xvi were 73 bp of upstream DNA and corresponds to the amount of upstream DNA normally used in our linear target DNA (6,29). East and South arms of all constructs are 25 bp. See Supplementary Figure S3 for mapping of DNA cleavages. See Supplementary Figure S4 for endonuclease mutant protein controls and for a graph of fraction bound as a function of protein concentration. See Supplementary Figure S5 for denaturing gels of specific EMSA bands.

The open 4-way junctions generated in Figure 3 fell into three broad categories designed to analyze the sequence and structure requirements for precise and efficient DNA cleavage: (a) construct v and its derivatives vi–xii which have a short mostly non-specific upstream (West) DNA arm, (b) construct xiv and its derivatives xv–xvi, which have the full upstream and downstream 28S sequence (West and North arms) and (c) construct xiii, which has a medium length, mostly non-specific, downstream (North) DNA arm (see also the derivative construct xvii in Figure 4). The first and third groups of constructs limit the potential conformational space of the resulting protein–nucleic acid complexes due to the fact that known protein-binding-sequences are being removed. The second group of constructs retains the full upstream and downstream 28S sequences and the protein binding sites contained therein. All three groups of constructs retained 28S sequences proximal to the second strand DNA cleavage site (5 bp on either side, West and North arms). Multiple construct variations within each category were explored in order to more precisely define the DNA structure and sequence parameters required for second-strand cleavage. An R2 protein titration series was run on each labeled construct depicted in Figure 3. The labeled strand is marked with an asterisk (*) in the construct cartoons. The reactions were analyzed on native (EMSA) and denaturing polyacrylamide gels as in Figure 2. The gels for each construct are shown below the corresponding construct cartoon. The data that led to the mapping of the R2 cleavages, as well an endonuclease deficient R2 protein control for each construct, are located in the supplementary material (Supplementary Figures S3 and 4, respectively).

There are several parameters to consider in determining cleavability: (i) the amount of protein required to bind to the DNA, (ii) the amount of cleavage per protein-bound unit of DNA and (iii) the precision of the DNA cleavage. The second and third parameters were the most useful ones for comparing the cleavability between constructs. The first parameter was less informative

because protein binding sites were being strategically removed and because of inherent issues with DNA quantitation when making the constructs and inherent issues with pipetting small volumes and protein (stored in glycerol) accurately.

Construct v was picked as a starting point in the analysis as construct v is the template-jump version of construct iv. The template-jump portion of construct v is the R2 5′ end (South) arm covalently attached to, and base paired with, the West arm. The South arm consists of a 25 bp cDNA/RNA duplex originating from the 5′ end of the R2 element RNA that would have been generated by TPRT. Construct v is a substantially better substrate for second strand cleavage than construct iv. Construct v cleaved about 11% of the protein-bound substrate. Interestingly, construct vi, which consisted only of the intact duplexed North arm and single stranded West and East arms (no South arm) cleaved just as well as construct v. The cleavage, however, was less precise. Only cleavages within few bases of the canonical R2Bm cleavage site were counted as second-strand cleavage for all constructs. Aberrant cleavages are marked as such on the denaturing gels and were not counted. Construct vii was structurally identical to construct vi, except it retained the original East arm heteroduplex. Construct vii, like construct vi resulted in imprecise cleavage at the R2 site and additional aberrant cleavages upstream of the R2 site on the single stranded West arm. A gray circle next to each construct cartoon represents the relative cleavability of the construct when normalized to construct v.

Both constructs viii and ix lack a duplexed North arm, resulting in aberrant cleavages at single stranded North arm and none at the R2 cleavage site. Because of the lack of cleavage at the R2 site, constructs viii and ix are noted in the figure as lacking detectable DNA cleavage.

Constructs x–xii test the result of having single-stranded East and/or South arms. The best substrate in terms of precision of cleavage and cleavage per bound unit of DNA was construct xii.

Indeed construct xii was the best substrate out of the v-xii group of constructs, even better that construct v. The cleavage observed for constructs v and xii was primarily due to the R2 protein acting on the full construct and not on the present, but minor, partial junctions (dotted line on EMSAs); otherwise, constructs vi–ix, themselves being partial junctions of construct v, would have been cleaved better than they were. It is unfortunate that the (labeled) cleaved product in constructs v–vii and xii migrates at or very near a naturally occurring partial junction (solid line next to dotted line).

Construct xiii switched the non-specific DNA from the West arm to the North arm. The North arm was made only 35 bp long and retained 5 bp of 28S DNA located near the second-strand cleavage site. Construct xiv returns both North and West arms to the 28S derived DNA sequence containing the full R2 integration site. Both constructs xiii and xiv struggled to be cleaved. Construct xii showed dynamics similar to constructs i and ii (Figure 2), indicating that the protein is binding to the West arm and not the North arm at the lower protein concentrations. For complex iv, this dynamic was less so.

Constructs xv and xvi are direct analogs to the integration intermediate presented in step 3 of the New Integration Model (Figure 1). The West arm of these two constructs contained a 'gap and a flap' as a result of the of the template jump displacing the original DNA strand. The recombined cDNA/target DNA duplex portion of the West arm was 27 bp so as to match the amount of target sequence retained in the R2Bm transcript after processing by the R2 ribozyme (35). The 27 bp template jump/recombination places the gap and flap well into the upstream binding site (DNase footprint) for the R2Bm protein (29,35). The bifurcation of the West arm is thought to impart flexibility to the arm. Construct xv was not very cleavable, presumably because it had a heteroduplexed East arm. Construct xvi, however, with its single-stranded East arm

cleaved well (15% cleavage per bound unit of DNA), second only to construct xii, which also happened to have a single stranded East arm. For example, a single stranded TPRT (East) arm would be expected to occur upon removal of the RNA from the cDNA/RNA heteroduplex by cellular RNase H activity. DNA cleavage decreased sooner on construct xvi than it did on either constructs v or xii in this data set (Figure 3). The early drop in cleavage for construct xvi is less pronounced in the data set presented in Figure 4 which was designed to test for SSS (see next section of the paper). The reverse transcriptase mutant (RT-) protein used in Figure 4 yields the same type of cleavage information as the wild-type protein (WT) used in Figure 3. Not only was the early drop in cleavage not as pronounced for construct xvi in Figure 4, but also the amount of cleavage per bound unit of DNA was nearly double for each of the constructs (v, xii, xvi) tested in Figure 4. The R2 protein used in Figure 4 was more active because it was fresher (1-day-old) than that used in Figures 2 and 3 which were age matched to 7 days old. DNA binding is long lived, but the amount of DNA cleavage per bound unit is age dependent. We also made a minor adjustment to the pH of our protein purification buffers. Datasets 2 and 3 were purified at pH 8, while the dataset in Figure 4 was returned to our traditional pH 7.5. The relative cleavage (gray dots) between constructs remained constant. Figure 4 also introduces a final construct. construct xvii. Construct xvii was similar to xvi with respect to the bifurcated West arm and single stranded East arm, but differs in that xvii had the non-specific North arm that construct xiii had. This construct surprisingly cleaved but was less precise; it cleaved the bases before and after (70–72) the cleavage site in addition to the canonical cleavage site at 71. Counting the cluster of cleavages as proper cleavage, the cleavability of construct xvii was significantly lower than that of construct xvi, but it otherwise had a similar fraction cleavage as a function of fraction bound profile to the other good substrates.

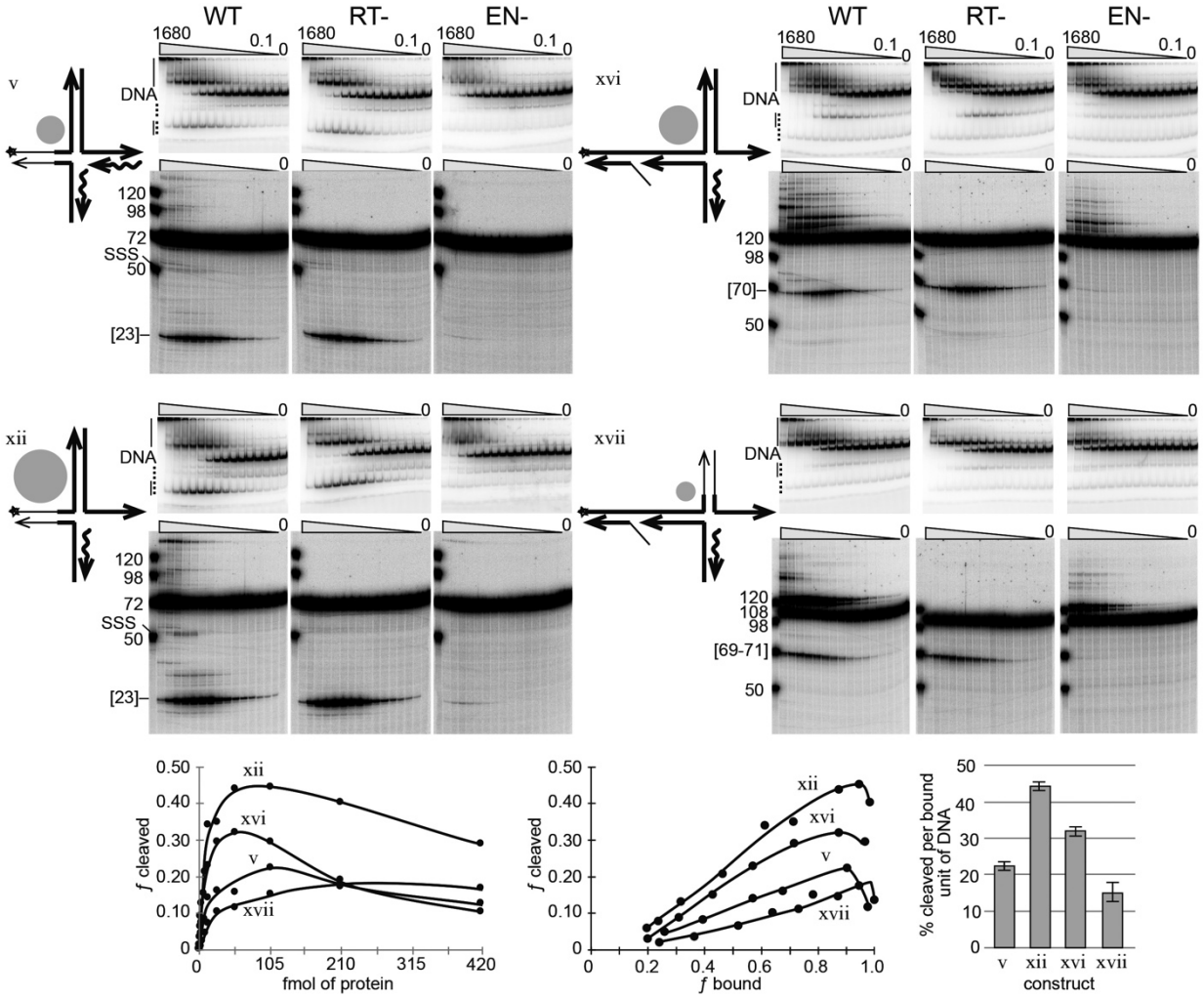


Figure 4. Second-strand DNA cleavage followed by second-strand DNA synthesis. Symbols, conventions, reactions, gels, analysis and graphs are as previous figures, except that dNTPs were added to the reactions. The reactions were carried out using wild-type R2 protein (WT), reverse transcriptase mutant protein (RT-) and endonuclease mutant protein (EN-). In addition, a broader protein titration range was used, 1680–0.1 fmol. The graphs include only 420–0.4 fmol range of the RT- cleavage data. A new construct, xvii, was included in this dataset in addition to constructs v, xii, xvi. The amount of SSS was not quantified as the signal is too low for reliable numbers. See Supplementary Figure S4 for a graph of fraction bound as a function of protein concentration for RT-dataset.

The relative ranking of the best substrates for secondstrand DNA cleavage, normalized to the cleavability of construct v, was xii > xvi > v ≥ vi > vii ≥ xvii. Substrates with a single-stranded East arm were more cleavable by the R2 protein than substrates with a duplexed east arm. The data

further indicated that the template-jump-derived West arm must be within a fairly narrow window of stability and that being too stable and rigid (xiii, xiv) seems to be inhibitory. Too low of a melting temperature leads to dissociation and concomitant loss of cleavage fidelity if the area remains single stranded (constructs vi,vii) or loss of cleavage if the structure returns to a TPRT-like 3-way junction (constructs ii, iv). A duplexed South arm, as opposed to a single-stranded South arm is required (e.g. compare x–xii). The bifurcated West arm appears to reduce/inhibit protein binding to only the upstream-28S R2-binding-site, although additional experiments will be needed to confirm, and DNA cleavage appears to be strongly associated with the North arm (i.e. downstream 28S DNA sequences, e.g. construct vi).

Protein–DNA complexes of constructs capable of being cleaved at the second-strand cleavage site by R2 protein form stable complexes that migrate within the EMSA gel, as opposed to being stuck in the well. Upon cleavage, the 4-way junction is resolved into two linear DNAs: one DNA containing the downstream (North) and R2 3 ′ (East) arms and one DNA containing the 'upstream' (West) and R2 5′ (South) arms. The West plus South DNA appears to be largely released by the R2 protein after cleavage, at least in the case of constructs iv–xii (see the lower solid vertical line in the EMSAs; see also Supplementary Figure S5). Some cleaved products for construct xvi can be found in the upper shifted region in the EMSA gel (i.e. still bound by protein) (Supplementary Figure S5). It is the West plus South DNA cleavage product that is expected to be the primer-template for second-strand DNA synthesis; as such, we would not expect it to be released by R2 protein in vivo. The fate of the North plus East half of the junction was not tracked post DNA cleavage.

Second-strand cleavage leads to second-strand synthesis in the presence of dNTPs

The third and final part of this study was to explore SSS. To test if second-strand cleavage could progress to SSS, dNTPs were added to the DNA cleavage reaction. In addition, the reactions were carried out using wild-type R2 protein (WT), reverse transcriptase mutant protein (RT-) and endonuclease mutant protein (EN-). The WT protein cleaves and synthesizes DNA. The RT- protein cleaves but does not synthesize DNA. The data for the RT-protein is therefore analogous to the binding and cleavage reactions in Figure 3. The EN-protein does not cleave but still has an active reverse transcriptase. The mutation in the reverse transcriptase was YAD/YD and the endonuclease mutation was K/ARNKY (11,22). The EN-protein retained a low-level residual junction-cleavage-activity.

The best cleaving constructs, v, xii and xvi, along with the new construct, xvii, were used in reactions containing dNTPs (Figure 4) to test for SSS. The reactions were analyzed by denaturing polyacrylamide gel electrophoresis. The labeled strand of constructs v and xii was 72 nt uncleaved and 23 nt in length upon second-strand DNA cleavage. SSS, i.e. extension of the labeled strand post-DNA cleavage, would generate a 50 nt product when analyzed on a denaturing gel. The radiolabeled strand of construct xvi was 120 nt long uncleaved, 70 nt cleaved and 98 nt upon SSS. The labeled strand of construct xvii was 108 nt long uncleaved, 69–71 nt cleaved and 98 nt upon SSS. A larger range of R2 protein concentrations was used than in the previous figures. Second-strand DNA synthesis was observed on the denaturing gels for constructs v and xii at the higher end of the protein titration series. When the R2 RT gets to the end of the template, it adds on several untemplated nucleotides (42). The signal above the full-length oligo on the denaturing gels is the result of the original full-length oligo being extended by the R2 protein. The R2 protein can take almost any 3 ′ end and extend it, given a template in cis or in trans (42,43). The reason

why full-length SSS was only prominent for constructs v and xii under conditions of protein excess was because the synthesis appears to be occurring primarily on the released primer template (lower vertical line on the EMSA) generated by second-strand cleavage and released from the protein/DNA complex. In vivo, it is not expected that the cleaved product would be released.

Partial SSS products were also detected, particularly in the case of construct xii. Several strong stops exist above the second-strand cleavage signal. These strong stops appear to occur as a direct result of synthesis being primed off the 3′-OH of the second-strand cleavage event; they are not present in either the RT- or the EN- datasets. The same stoppages were observed when construct xvi was used, indicating that priming of second strand synthesis also occurs on construct xvi. The strong stops may be the result of a structural constraint or required protein-DNA conformation change to switch from priming to elongation. The presence of the strong synthesis stops tracks strongly with the DNA cleavage profile.

DISCUSSION

A new model for R2Bm integration

The deeper understanding of the second half of the insertion reaction for R2Bm derived from the above experiments has allowed for an improved R2Bm integration model to be put forth (Figure 1C). The first half of the integration reaction is identical to steps 1 and 2 in the old ('current') integration model. After TPRT, however, the new integration model proposes a template jump or recombination event from the 5′ end of the R2 RNA to the top-strand of the 28S rDNA, upstream of the R2 insertion site, forming a 4-way junction (Figure 1C, step 3). It is this step that, to date, has not been shown to occur in vitro and may require host factors to form. An association of the cDNA to the upstream target DNA is consistent; however, with previous data,

and the 4-way junction intermediate leads to a simple unified mechanism for 5′ junction formation and completion of integration.

Indeed, the new integration model makes sense of earlier in vivo experiments in which the 'upstream' ribosomal RNA sequence attached to the 5′ end of the R2Bm element RNA had been noted as a requirement for full-length element insertion (40,41). Studies have also determined that the R2 RNA is co-transcribed with ribosomal RNAs as part of the same large transcript (35,44). The R2 RNA is then processed from the bulk of the ribosomal RNA by a hepatitis delta virus (HDV)-like ribozyme found near the 5′ end of the R2 RNA (35,44). For a number of R2 elements, the processed R2 RNA retains some ribosomal RNA on the 5′ end, 27 nt of ribosomal RNA in the case of R2Bm (35). For elements that retain this much of the ribosomal RNA, the 'template jump' may be more of a strand invasion or recombination event than an actual template jump (40,41). For other R2 elements, however, the ribozyme leaves no ribosomal sequence on the processed R2 RNA (e.g. Drosophila simulans R2), and a template jump, as diagrammed in Figure 1C step 3 (of the new integration model), is envisioned to occur (16,35,39,43). The RT of both APE LINEs and RLE LINEs has been shown to have the ability to jump from the end of one template to the beginning of another without any homology (43). Template jumps have long been hypothesized to be involved in 5′ junction formation for both types of elements (16,35,39,43). In addition to template jumping, the reverse transcriptase of LINEs is able to use both DNA and RNA as templates during DNA synthesis and to displace a duplexed strand while polymerizing (16)

Recently, the R2 RLE's reported similarity to Archaeal Holliday junction resolvases raised the question as to whether R2 binds and cleaves branched DNAs during integration (11,37). It turns out that the binding and cleavage of branched DNA is fundamental to the integration process itself. However, despite the formation and resolution of a 'Holliday junction-like' integration

intermediate, with nearly symmetrical DNA cleavages, the R2 protein is not a Holliday junction resolvase. In fact, the cleavages are separated in time and arise via an activity much closer to that of a monomeric, single-stranded, DNA-endonuclease activity. Second strand cleavage appears to be the result of the endonuclease, and/or the R2 protein, associating with a double-stranded region and cleaving a nearby singlestranded region. This activity is exemplified by the cleavage data for constructs vi and vii in Figure 3. The other constructs that cleaved well, presumably, have a single stranded attribute to the cleavage site. Indeed, the cleavage site migrated between constructs as if in response to small local changes to the single-strandedness in the cleavage region (Figures 2–4; Supplementary Figure S6).

Similarly, there are good reasons to believe that firststrand cleavage and second-strand DNA cleavages are, at a fundamental level, identical with respect to how they arise since both instances are brought about by the same RLE. Indeed, DNase footprints of R2 protein bound to target linear DNA, prior to first-strand cleavage, show R2 protein induced DNase hypersensitive sites near the R2 cleavage/insertion site: local unwinding of a double helix would lead to DNase hypesensitive sites (6,29). Thus firststrand DNA cleavage may also be the result the endonuclease associating with a double stranded region and cutting a nearby single-stranded region.

Further, it is not known which part of the R2 protein binds the 4-way DNA junction. It may or may not be the endonuclease (45). It remains to be investigated whether the jump/recombination event precludes protein binding to the upstream (−40 bp to −20 bp) binding site, as our results suggest.

Cleavage of the 4-way junction generated a natural primer-template used for second-strand DNA synthesis. In our in vitro reactions, however, much of the primer-template is released after cleavage. As such, it remains an open question as to wether or not R2 provides this function in

vivo. It is encouraging, however, that construct xvi yielded SSS products in the form of constrained synthesis. It is possible that the priming occurred on cleaved substrates still bound by protein.

One protein subunit or two? Is the integration reaction performed by one protein subunit or two? It is an open and unresolved question. The two subunit model presented in Figure 1B and C (current integration model), still fits all the data. That said, the one subunit model also fits most, if not all, of the data. The new data has the R2 protein recognizing a sequential set of complicated branched DNA structure(s). Each arm of the branched structure(s) appear to have their own sequence and local structure requirements that must be met for integration to occur. Our new data intellectually fits well with a one subunit 'rock and roll' model. In the rock and roll model, the R2 protein is bound to the 3 ′ PBM RNA and thus binds to the upstream 28S DNA (West arm) on the linear DNA. Binding of the upstream R2 protein subunit to the target DNA induces local unwinding at the R2 site. The endonuclease of the upstream bound R2 protein 'rocks' into place and cleaves the single-stranded R2 site. The reverse transcriptase of the upstream bound R2 protein is rocked into place and begins TPRT. The initial stages of TPRT removes the 3 ′ PBM RNA from the protein (due to heteroduplex formation). The 5 ′ PBM RNA associates with the R2 protein and the protein adopts the downstream binding conformation; the protein 'rolls' to the North arm while also making potential contacts with the TPRT (East) and West arms. TPRT finishes, removing the 5 ′ PBM RNA from the protein (due to heteroduplex formation). The R2 protein is now in the minus RNA state. The template jump occurs to form the open 4-way junction the R2 protein rolls to bind the 4-way junction as described in the 'Results' section. The endonuclease rocks into place and cleaves the open 4way junction. The reverse transcriptase is then rocked into place to perform second-strand DNA synthesis. More experiments are needed to determine which model, one or two subunit, is correct and to more fully understand the integration reaction.

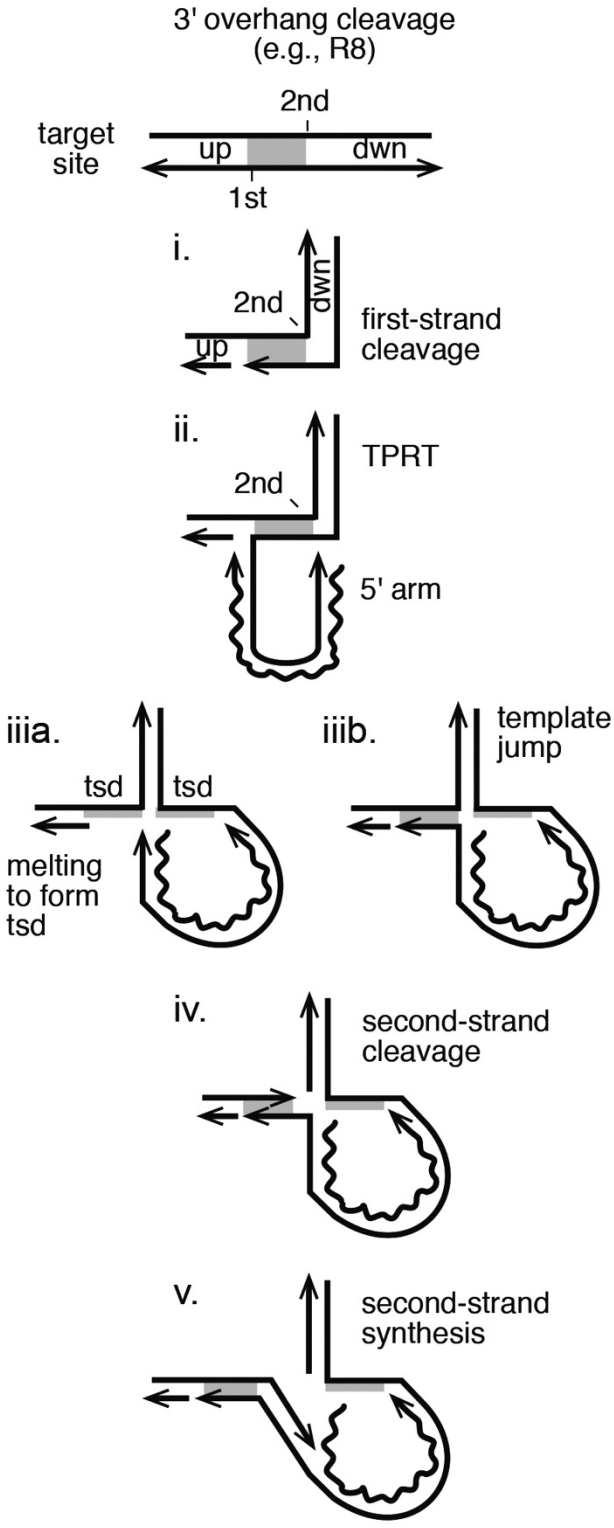


Figure 5. Extrapolating the New Integration Model to RLE LINES that generate target site duplications. A target site is diagrammed with the firstand second-strand DNA cleavages staggered such that a target site duplication (tsd) would occur upon element insertion. The steps are as in R2 integration, except that the template jump displaces/melts the DNA between the two cleavages to generate the open 4-way junction and the tsd upon DNA synthesis.

Extrapolating the R2 model to LINEs with different cleavage staggers

The position of the second-strand DNA cleavage site relative to the first-strand cleavage site is variable across species, and even more so across the R2 clade. The stagger of the first and second DNA cleavage events in R2Bm is a small 5 ′ overhang of 2 bp that leads to 2 bp target site deletion upon insertion of the element. In Drosophila melanogaster the R2 endonuclease produces blunt cleavages (39). Other R2 elements produce small 3 ′ overhangs (26). The 3 ′ prime overhanging staggered cuts produce target site duplications instead of deletions. The model presented in Figure 1 works equally well for elements with any form of small staggers. The model easily can be adapted for elements that generate larger target site duplications. The R8 element in Hydra magnipapillata

generates a 9 bp target site duplication upon insertion (46). The R4 element generates a 13 bp target site duplication (46). The model for elements like R8 and R4 is presented in Figure 5. The difference between the model in Figure 1, where the cleavage stagger is small, and that proposed for R8 is that a local melting or displacement of the region between the cleavage sites is hypothesized to occur along with the template switch, generating the 4-way junction.

APE LINEs also tend to produce a 3 ′ overhanging stagger in the range of 10–20 bp. It remains to be determined if APE LINEs use a 4-way junction structure to drive secondstrand DNA cleavage and synthesis. Bioinformatic analysis of 5 ′ junctions of full-length L1 and Alu elements is suggestive of template jumping to the upstream target sequence and that DNA repair might be an alternative path to 5 ′ junction formation for abortive insertion events (1,15,17,47,48). Twin priming in L1 might be a related, albeit aberrant, phenomenon to SSS (49). An association between the cDNA and the upstream target DNA has been hypothesized for some R1 elements (39). Ribosomal sequences are also important for element–RNA/target–DNA interactions during first-strand synthesis for R1Bm as well as several other site-specific LINEs, but they do not appear to be as important for R2Bm (26,50,51).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

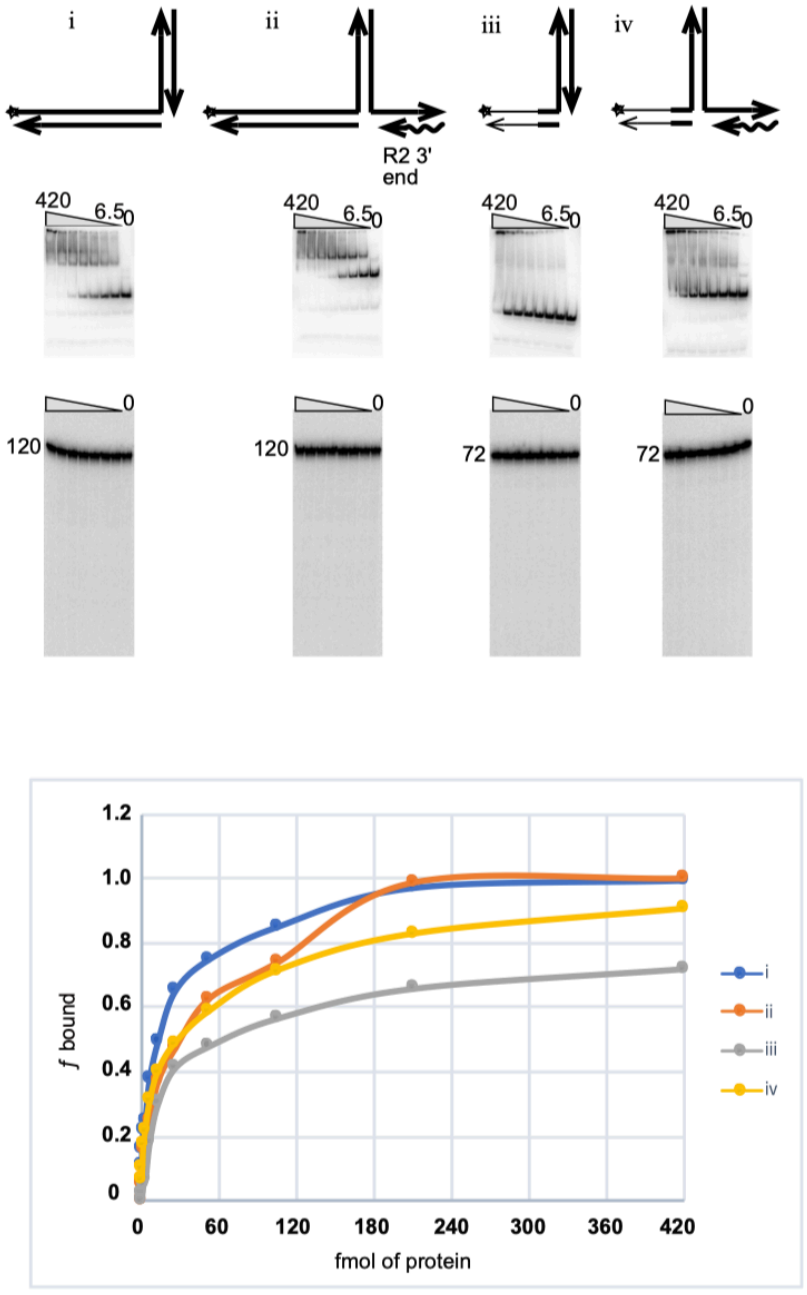
We thank Monika Pradhan, Santosh Dhamala and Murshida Mahbub for helpful discussions along the way and for critical reading of the manuscript. We thank Micki Christensen for final copy editing.

FUNDING

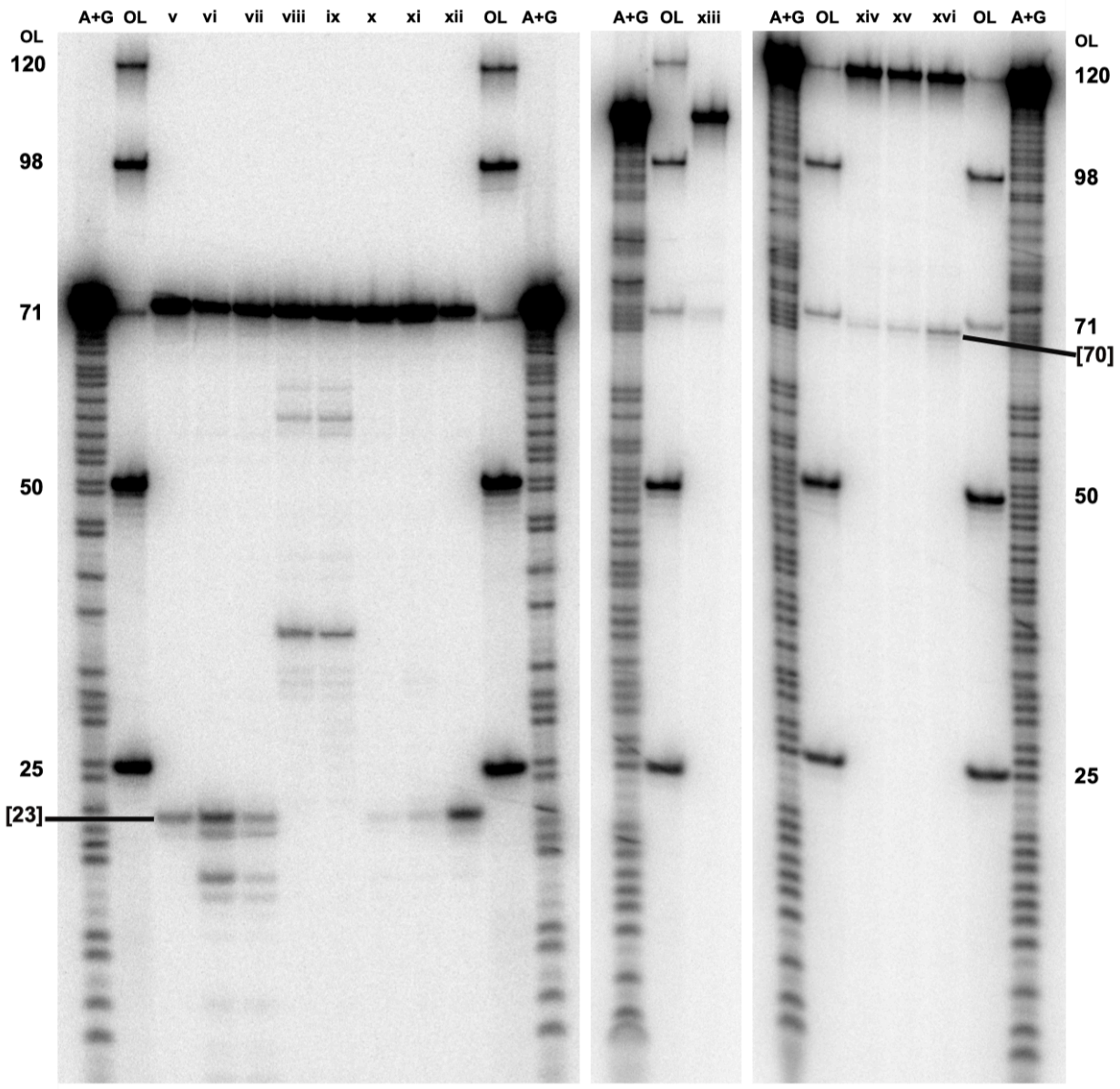
Phi Sigma Graduate Student Biology Society (to B.B.K., A.G.) (in part); University of Texas Arlington Research Enhancement (REP) Grant Program (to S.C.). Funding for open access charge: University Funds; Personal Funds. Conflict of interest statement. None declared.

Supplemental S1: Oligonucleotides used in the study.

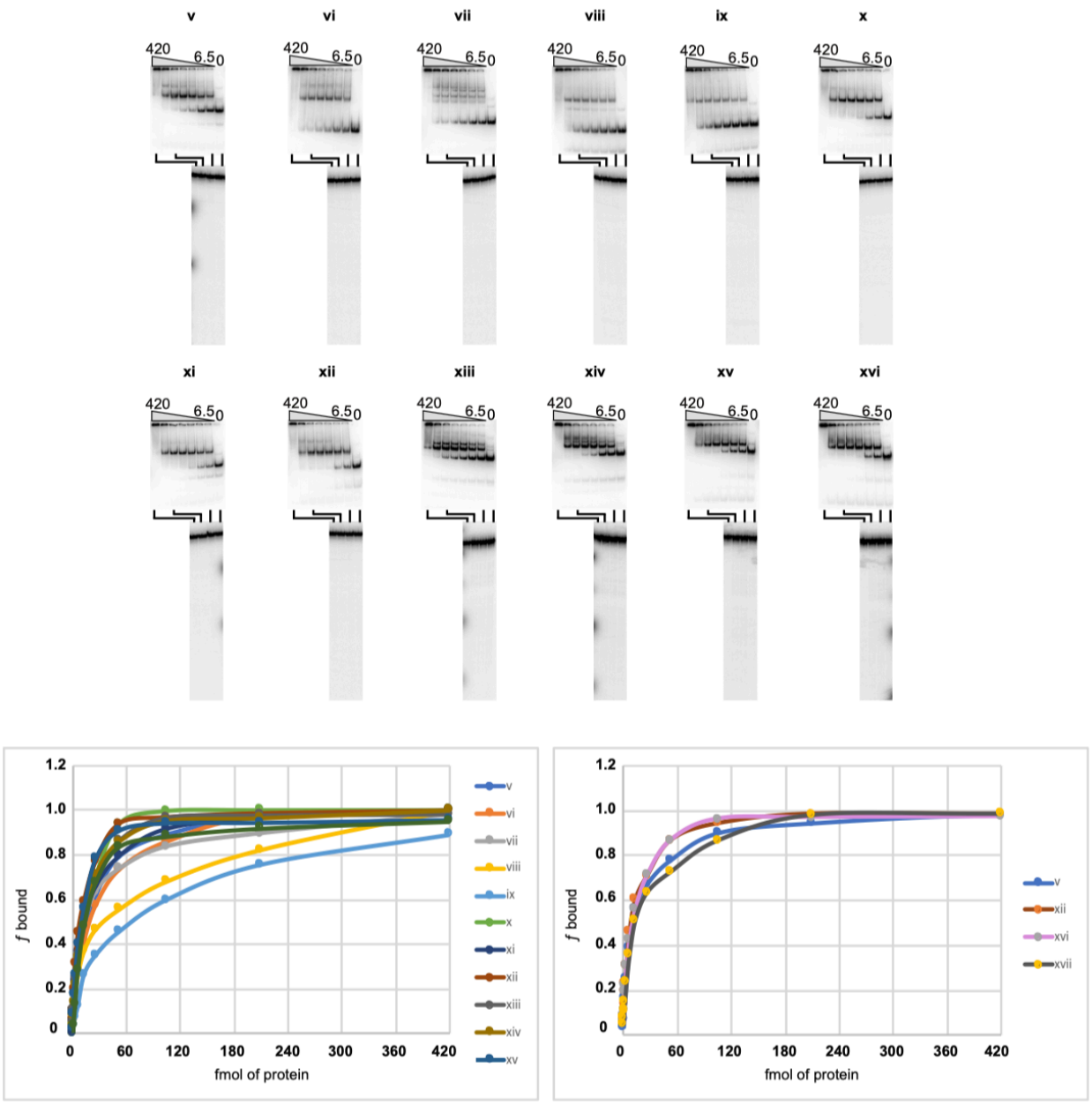
Name	Oligonucleotide Sequence (5'---3')
NS / 28Sd 25bp	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTCATCTAATT
Comp 28Sd 25bp / NS	AATTAGATGACGAGGCATTTGGCTACCTTAAGCTACCGGAAGCTTCTGGA
Pre-cleaved (1) Comp 28Sd 25bp; (2) NS	(1) AATTAGATGACGAGGCATTTGGCTA (2) CCTTAAGCTACCGGAAGCTTCTGGA
R2 3' DNA 25bp / NS	TGGCATGATGATCCGGCGATGAAAACCTTAAGCTACCGGAAGCTTCTGGA
R2 3' DNA 25bp / R2 5' DNA 25bp	TGGCATGATGATCCGGCGATGAAAAGGGGCGATACGCATAATTTTAATTT
R2 3' DNA 25bp	TGGCATGATGATCCGGCGATGAAAA
R2 3' RNA 25bp	UGGCAUGAUGAUCCGGCGAUGAAAA
R2 5' DNA 25bp	GGGGCGATACGCATAATTTTAATTT
R2 5' RNA 25bp	GGGGCGAUACGCAUAAUUUUAAUUU
Comp 28Sd 25bp / R2 3' DNA 25bp	AATTAGATGACGAGGCATTTGGCTATGGCATGATGATCCGGCGATGAAAA
Comp 28Sd 25bp / Comp R2 3' DNA 25bp	AATTAGATGACGAGGCATTTGGCTATTTTCATCGCCGGATCATCATGCCA
Comp R2 3' RNA 25bp / NS	TTTTCATCGCCGGATCATCATGCCACCTTAAGCTACCGGAAGCTTCTGGA
Comp R2 5' RNA 25bp / NS	AAATTAAAATTATGCGTATCGCCCCCCTTAAGCTACCGGAAGCTTCTGGA
NS / 28Sd 47bp	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTA
Comp 28Sd 47bp / Comp R2 3' RNA 25bp	TAATCCATTCATGCGCGTCACTAATTAGATGACGAGGCATTTGGCTATTTTCATCGCCGGATCATCATGCCA
28Su 73bp / NS	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGG GTCCAGAAGCTTCCGGTAGCAGCGAGAGCGG
Comp NS / Comp R2 3' RNA 25bp	CCGCTCTCGCTGCTACCGGAAGCTTCTGGACCCTATTTTCATCGCCGGATCATCATGCCA
Comp R2 5' RNA 25bp / Comp 28Su 73bp	AAATTAAAATTATGCGTATCGCCCCCCTTAAGAGAGTCATAGTTACTCCCGCCGTTTACCCGCGCTTGCTTGAATTTCT TCACGTTGACATTCAGAGC
28Su 73bp / 28Sd 47bp	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGCC AAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTA
Comp R2 5' RNA 25bp / Comp 28Su 27bp	AAATTAAAATTATGCGTATCGCCCCCCTTAAGAGAGTCATAGTTACTCCCG
Flap Comp 28Su	ATATATGTTTACCCGCGCTTGCTTGAATTTCTTCACGTTGACATTCAGAGC



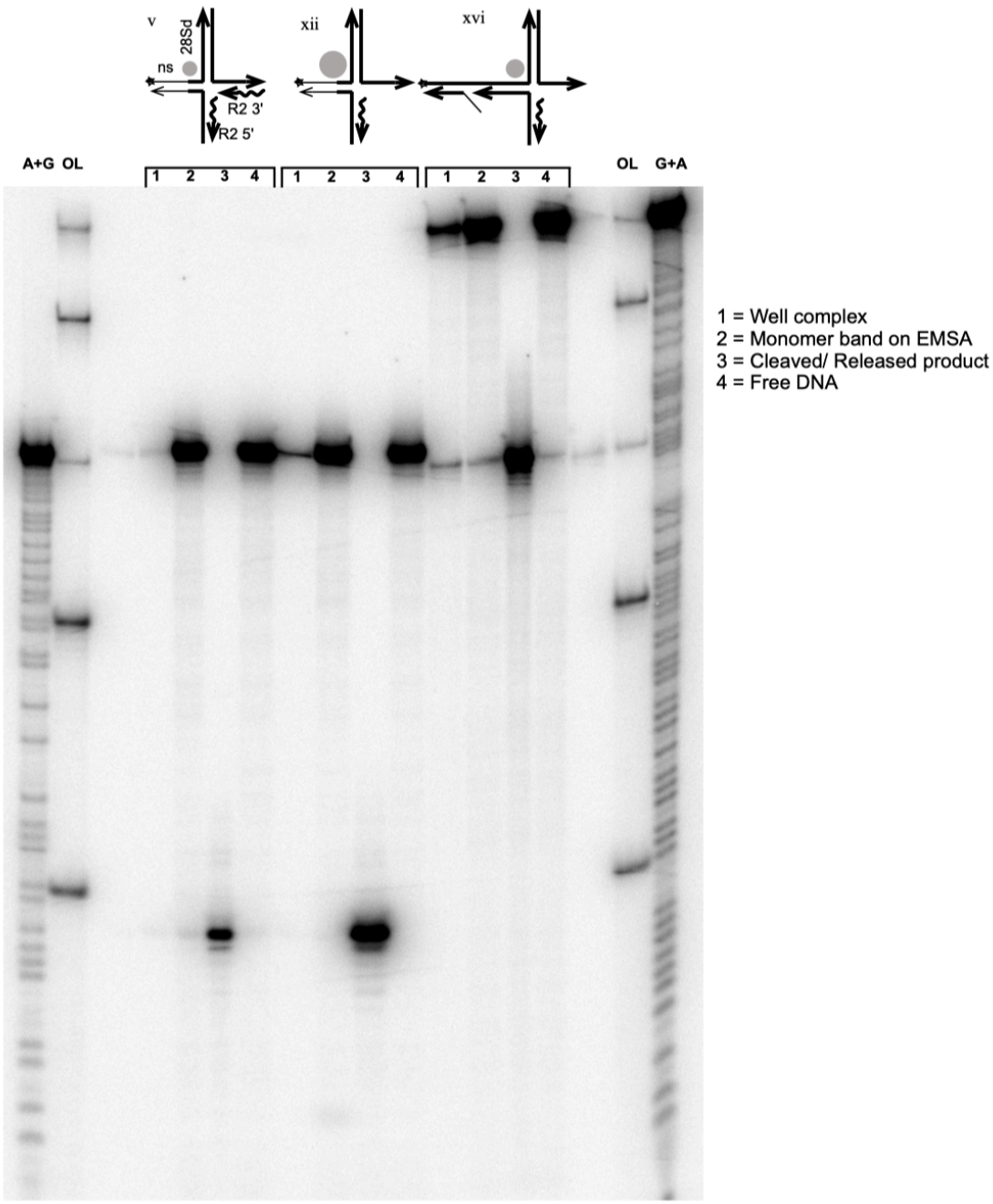
Supplemental S2: Related to Figure 2. EMSA gels of R2Bm protein acting on constructs i-iv across a range of endonuclease mutant (EN-) R2Bm protein concentrations (420-6.5 fmol). A cartoon of each construct is presented. Under the construct cartoon is the native gel (EMSA) analysis and corresponding denaturing gel for each construct. The exposure of the gels in this figure was linearly adjusted such that the bands, if any, would readily be visible. A graph of average DNA bound (*f* bound) as a function of protein concentration (fmol/reaction) is shown for constructs i-iv.



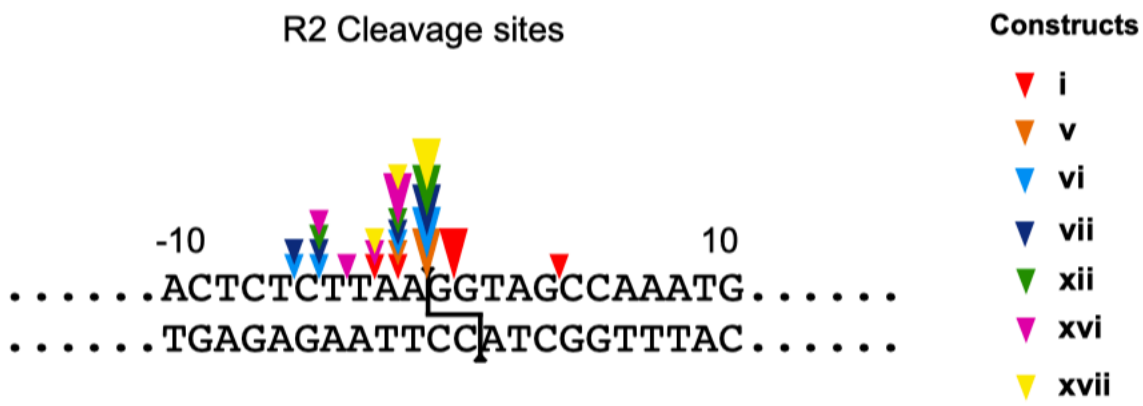
Supplemental S3: Related to Figure 3. Urea-denaturing 8% polyaclylamide gel showing mapping of second-strand DNA cleavages for constructs v-xvi. The in vitro reactions were carried out using wild type R2Bm protein (WT) for each construct listed at the top of the gel. A+G ladder as well as ladders made from different sized DNA oligos (as indicated on the figure) were run alongside the reactions to aid in mapping of second-strand DNA cleavage. The exposure of the gels in this figure was linearly adjusted such that bands resulting from DNA cleavage were readily visible.



Supplemental S4: Related to Figure 3. EMSA gels and urea-denaturing 8% polyacrylamide gels showing data for endonuclease mutant protein (EN-) acting on constructs v-xvi across a range protein concentration (420-6.5 fmol). Native gel (EMSA) analysis is presented. Under each native gel (EMSA) are the corresponding denaturing gel for each construct. Each lane on denaturing gel represents a specific protein concentration of the titration series on the corresponding EMSA gel, as indicated by straight lines. The exposure of the gels in this figure was linearly adjusted such that the bands, if any, would readily be visible. A graph of average DNA bound (*f* bound) as a function of protein concentration (fmol/reaction) are shown for constructs v-xii. Also, a graph of average DNA bound (*f* bound) as a function of protein concentration (fmol/reaction) are shown for constructs v-xviii (Related to Figure 4).



Supplemental S5: Related to Figure 3. Urea-denaturing 8% polyaclylamide gel showing mapping of cleaved and released products for constructs i, viii and xii. A cartoon of each construct is shown above each corresponding set [1-4; described each next to denaturing gel]. The in vitro reactions were carried out using wild type R2Bm protein (WT) for each construct. A+G ladder as well as ladders made from different sized DNA oligos (as indicated on the figure) were run alongside the reactions to aid in mapping cleaved and released products. The exposure of the gels in this figure was linearly adjusted such that bands resulting from DNA cleavage and product released were readily visible.



Supplemental S6: Diagram of the second-strand cleavages on the 28S target. Colored triangles indicate second-strand cleavage for designated constructs. Nucleotide positions are numbered with respect to the central dyad. The canonical cleavage site are indicated by black arrows. Each construct shows a major cleavage indicated by a "larger" triangle and other cleavage indicated by "smaller" triangles. The cleavages for construct i were determined by looking at the denaturing gel in Figure 5 from 2004 Christensen and Eickbush paper (Christensen and Eickbush, 2004).

REFERENCES

1. Richardson,S.R., Doucet,A.J., Kopera,H.C., Moldovan,J.B., Garcia-Perez,J.L. and Moran,J.V. (2015) The Influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol. Spectr., 3, MDNA3-0061.

2. Casola,C. and Betr´an,E. (2017) The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses Genome Biol. Evol., 9, 1351–1373.

3. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell, 72, 595–605.

4. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription in vitro. EMBO J., 21, 5899–5910.

5. Moran,J.V. and Gilbert,N. (2002) Mammalian LINE-1 retrotransposons and related elements. In: Craig,NL, Craigie,R, Gellert,M and Lambowitz,AM (eds). Mobile DNA II. ASM Press, Washington, DC, pp. 836–869.

6. Christensen,S.M. and Eickbush,T.H. (2005) R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. Mol. Cell Biol., 25, 6617–6628.

7. Kulpa,D.A. and Moran,J.V. (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat. Struct. Mol. Biol., 13, 655–660.

8. Dewannieux,M. and Heidmann,T. (2005) LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. Cytogenet. Genome Res., 110, 35–48.

9. Doucet,A.J., Wilusz,J.E., Miyoshi,T., Liu,Y. and Moran,J.V. (2015) A 3 ′ Poly(A) tract is required for LINE-1 retrotransposition. Mol. Cell, 60, 728–741.

10. Christensen,S.M., Bibillo,A. and Eickbush,T.H. (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. Nucleic Acids Res., 33, 6461–6468.

11. Govindaraju,A., Cortez,J.D., Reveal,B. and Christensen,S.M. (2016) Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. Nucleic Acids Res., 44, 3276–3287.

12. Martin,S.L. (2010) Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol., 7, 67–72.

13. Martin,S.L. (2006) The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. J. Biomed. Biotechnol., 2006, 45621.

14. Matsumoto,T., Hamada,M., Osanai,M. and Fujiwara,H. (2006) Essential domains for ribonucleoprotein complex formation required for retrotransposition of telomere-specific non-long terminal repeat retrotransposon SART1. Mol. Cell Biol., 26, 5168–5179.

15. Zingler,N., Willhoeft,U., Brose,H.P., Schoder,V., Jahns,T., Hanschmann,K.M., Morrish,T.A., Lower,J. and Schumann,G.G. (2005) Analysis of 5 ′ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5 ′ -end attachment requiring microhomology-mediated end-joining. Genome Res., 15, 780–789.

16. Kurzynska-Kokorniak,A., Jamburuthugoda,V.K., Bibillo,A. and Eickbush,T.H. (2007) DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. J. Mol. Biol., 374, 322–333.

17. Ichiyanagi,K., Nakajima,R., Kajikawa,M. and Okada,N. (2007) Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. Genome Res., 17, 33–41.

18. Gasior,S.L., Wakeman,T.P., Xu,B. and Deininger,P.L. (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol., 357, 1383–1393.

19. Suzuki,J., Yamaguchi,K., Kajikawa,M., Ichiyanagi,K., Adachi,N., Koyama,H., Takeda,S. and Okada,N. (2009) Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLos Genet., 5, e1000461.

20. Christensen,S.M., Ye,J. and Eickbush,T.H. (2006) RNA from the 5 ′ end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. Proc. Natl. Acad. Sci. U.S.A., 103, 17602–17607.

21. Eickbush,T.H. and Malik,H.S. (2002) Origins and evolution of retrotransposons. In: Craig,NL, Craigie,R, Gellert,M and Lambowitz,AM (eds). Mobile DNA II. ASM Press, Washington, DC, pp. 1111–1146.

22. Yang,J., Malik,H.S. and Eickbush,T.H. (1999) Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc. Natl. Acad. Sci. U.S.A., 96, 7847–7852.

23. Feng,Q., Moran,J.V., Kazazian,H.H.J. and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell, 87, 905–916.

24. Weichenrieder,O., Repanas,K. and Perrakis,A. (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure, 12, 975–986.

25. Han,J.S. (2010) Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob. DNA, 1, 15.

26. Fujiwara,H. (2015) Site-specific non-LTR retrotransposons.

Microbiol. Spectr., 3, MDNA3-0001.

27. Eickbush,T.H. and Eickbush,D.G. (2015) Integration, regulation, and long-term stability of R2 retrotransposons. Microbiol. Spectr., 3, MDNA3-0011.

28. Gladyshev,E.A. and Arkhipova,I.R. (2009) Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. Gene 448, 145–150.

29. Christensen,S. and Eickbush,T.H. (2004) Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. J. Mol. Biol., 336, 1035–1045.

30. Zingler,N., Weichenrieder,O. and Schumann,G.G. (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. Cytogenet. Genome Res., 110, 250–268.

31. Christensen,S., Pont-Kingdon,G. and Carroll,D. (2001) Comparative studies of the endonucleases from two related Xenopus laevis retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. Genetica, 110, 245–256.

32. Ostertag,E.M. and Kazazian,H.H.J. (2001) Biology of mammalian L1 retrotransposons. Annu. Rev. Genet., 35, 501–538.

33. Feng,Q., Schumann,G. and Boeke,J.D. (1998) Retrotransposon R1Bm endonuclease cleaves the target sequence. Proc. Natl. Acad. Sci. U.S.A., 95, 2083–2088.

34. Maita,N., Aoyagi,H., Osanai,M., Shirakawa,M. and Fujiwara,H. (2007) Characterization of the sequence specificity of the R1Bm endonuclease domain by structural and biochemical studies. Nucleic Acids Res., 35, 3918–3927.

35. Eickbush,D.G., Burke,W.D. and Eickbush,T.H. (2013) Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. PLoS One, 8, e66441.

36. Kajikawa,M., Yamaguchi,K. and Okada,N. (2012) A new mechanism to ensure integration during LINE retrotransposition: a suggestion from analyses of the 5 ′ extra nucleotides. Gene, 505, 345–351.

37. Mukha,D.V., Pasyukova,E.G., Kapelinskaya,T.V. and Kagramanova,A.S. (2013) Endonuclease domain of the Drosophila melanogaster R2 non-LTR retrotransposon and related retroelements: a new model for transposition. Front Genet., 4, 63.

38. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B. et al. (2012) Fiji: an open-source platform for biological-image analysis. Nat. Methods, 9, 676–682.

39. Stage,D.E. and Eickbush,T.H. (2009) Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. Genome Biol., 10, R49.

40. Fujimoto,H., Hirukawa,Y., Tani,H., Matsuura,Y., Hashido,K., Tsuchida,K., Takada,N., Kobayashi,M. and Maekawa,H. (2004) Integration of the 5 ′ end of the retrotransposon, R2Bm, can be complemented by homologous recombination. Nucleic Acids Res., 32, 1555–1565.

41. Eickbush,D.G., Luan,D.D. and Eickbush,T.H. (2000) Integration of Bombyx mori R2 sequences into the 28S ribosomal RNA genes of Drosophila melanogaster. Mol. Cell Biol., 20, 213–223.

42. Bibillo,A. and Eickbush,T.H. (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J. Biol. Chem., 279, 14945–14953.

43. Bibillo,A. and Eickbush,T.H. (2002) The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. J. Mol. Biol., 316, 459–473.

44. Eickbush,D.G. and Eickbush,T.H. (2010) R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. Mol. Cell Biol., 30, 3142–3150.

45. Wyatt,H.D. and West,S.C. (2014) Holliday junction resolvases. Cold Spring Harb. Perspect. Biol., 6, a023192.

46. Kojima,K.K., Kuma,K., Toh,H. and Fujiwara,H. (2006) Identification of rDNA-specific non-LTR retrotransposons in Cnidaria. Mol. Biol. Evol., 23, 1984–1993.

47. Gasior,S.L., Roy-Engel,A.M. and Deininger,P.L. (2008) ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst.), 7, 983–989.

48. Coufal,N.G., Garcia-Perez,J.L., Peng,G.E., Marchetto,M.C., Muotri,A.R., Mu,Y., Carson,C.T., Macia,A., Moran,J.V. and Gage,F.H. (2011) Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc. Natl. Acad. Sci. U.S.A., 108, 20382–20387.

49. Ostertag,E.M. and Kazazian,H.H.J. (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res., 11, 2059–2065.

50. Anzai,T., Osanai,M., Hamada,M. and Fujiwara,H. (2005) Functional roles of 3 ′ -terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. Nucleic Acids Res., 33, 1993–2002.

51. Luan,D.D. and Eickbush,T.H. (1996) Downstream 28S gene sequences on the RNA template affect the choice of primer and the accuracy of initiation by the R2 reverse transcriptase. Mol. Cell Biol., 16, 4726–4734.

CHAPTER 3

DNA structure, DNA sequence, and protein binding requirements for second-strand cleavage during integration of RLE LINEs

Brijesh B. Khadgi and Shawn M. Christensen

Department of Biology, University of Texas at Arlington

Arlington, TX 76019, USA

ABSTRACT

The R2 Long Interspersed Elements (LINEs) specifically integrate in the 28S rRNA genes by a series of DNA binding, DNA cleavage, and DNA synthesis reactions. The first half of the integration reaction, TPRT, is well understood. The second half of the integration reaction, second-strand DNA cleavage and second-strand DNA synthesis are much less well understood. Recently, a branched DNA integration intermediate was found that appears to finally allow biochemical dissection of the second half of the integration reaction. The branched integration intermediate, thought to arise by template jumping, is further explored using purified components *in vitro*. The structure of the junction itself was explored by probing with DNase I and was found to be highly structured. R2 protein binding to the junction was explored by a combination of DNA cleavage assays and DNA footprint studies. The protein appears to bind in a sequence specific manner to the north arm, but less so for the west arm where structure appears to be more important.

INTRODUCTION

Eukaryotic genomes across the tree of life contain a diverse group of transposable elements (TEs) known as long interspresed nuclear elements (LINEs) or non-long terminal repeats (non-LTRs). LINEs transpose via self-encoded protein machinery and are therefore considered to be autonomous in their mobility. Short interspresed nuclear elements (SINEs), on the other hand, are non-autnomous and utilize the protein machinery of LINEs to replicate. Mobility of both LINEs and SINEs have a profound effect on both the structure and function of their host genome since the replication and abundancy of these elements often result in insertions, deletions, and recombination events that are deleterious. TEs also act as a source of novel genetic material

resulting in new genes and regulatory sequences for the host [1] [2] [3]. LINEs replicate by a process called Target Primed Reverse Transcription (TPRT) within a host genome. TPRT involves a rionucleoptoein complex (RNP) which binds to target DNA near the site of insertion, cleaves and releases a 3'-hydroxyl that is used as a primer to prime reverse transcription of an element RNA into DNA [4] [5] [6]. LINEs encode multifunctional protein that binds to their own mRNA, perform first-strand cleavage and TPRT, and is also thought to involve in second-strand cleavage and second-strand synthesis. LINEs generally encode endonucleases that are either site-specific during integration as in the case of early branching R2 clade elements that encode restriction-like endonuclease (RLE) or non site-specific during integration like recently branched LINE-1elements that encode an apurinic-apyrimidinic family endonuclease (APE) [6] [7] [8] [9]. Both APE-bearing and RLE-bearing LINEs, share a common reverse transcriptase (RT) and contain a IAP/gag-like cysteine-histidine (CCHC) zinc knuckle motifs [10] [11] [12]. In addition, both LINEs encode multi-functional protein(s) with RNA binding, DNA binding, DNA cleavage, and reverse transcriptase activities and are hypothesized to have comparable integration process [6] [9] [10] [13] [14] [15] [16] [17].

Target-site specificity, recognition and subsequent DNA cleavages appear to be similar in both APE-LINEs and RLE LINEs. Unlike blunt cleavages, staggered cleavages generally can occur anywhere from a few bases away (e.g. 2 bp in R2Bm) to quite afar (e.g. 126 bp in R9). Staggered cuts, in particular, give rise to either a target site duplication or a target site deletion depending on whether the cleavage is 3' overhanging or 5' overhanging, respectively. [18] [19]. LINEs often have their target cleavage sites for first-strand and second-strand at different sequences (i.e. non-palindromic) and therefore, determination of second-strand cleavage has remained ambiguous. While the endonuclease for both APE LINEs and RLE LINEs have shown some specificity for first-strand DNA cleavage site, there are other unknown aspects that could possibly

be crucial including unidentified DNA binding domains in the element encoded protein(s) and varied nucleic acids conformations that could be important for target site recognition and second-strand DNA cleavage [20] [21]. Further, question remains as to whether first-strand DNA cleavage is a prerequisite for the second-strand DNA cleavage. R2 clade RLE is comparable to Archaeal Holliday junction resolvases. While RLE-bearing R2 has been shown to bind and cleave branched DNA intermediates during integration, R2 is not a Holliday junction resolvase [22] [20]. Rather, R2 RLE is hypothesized to associate with a double-stranded region and cleave a nearby single-stranded region. Cleavages of both the first-strand DNA and second-strand DNA are believed to be similar and by the same endonuclease [21]. However, how this endonuclease of R2 protein manages to recognize and bind to the DNA target site remains unclear, specifically with the DNA integration intermediates. In addition, it is also not known how the endonuclease recognizes two different sites (i.e. first-strand cleavage and second-strand cleavage sites) on the same target DNA.

RLE-LINE R2 elements from *bombyx mori* (R2Bm) specifically insert into the 28S rRNA genes of their host. R2s encode a single open reading frame (ORF) with N-terminal zinc finger(s) (ZF), a myb domain (Myb) and an RNA binding region (RB), a central RT domain, an RLE-type endonuclease, and a gag-knuckle-like CCHC motif at the C-terminal end (Figure 1) [13]. The RLE is a variant of the PD-(D/E)XK superfamily of endonucleases and has recently been recognized to be comparable to the Archaeal Holliday junction resolvases in both sequence and structure [20]. Holliday junction resolvases are endonucleases that cleave DNA integration intermediates symmetrically, during homologous recombination events, resulting in two resolved double stranded DNAs (dsDNAs) [23] [24]. The R2 element is also flanked by untranslated regions (UTR) on either side of the element and are termed as the 5' and 3' protein binding motifs (PBMs) both of which are implicated in R2 protein binding [13] [14] [15]. The R2 RNA is co-transcribed with rRNA

which is then processed by a hepatitis delta virus (HDV)-like ribozyme found near the 5' end of the R2 RNA [25] [26].

The R2 protein is a multifunctional protein that binds RNA structures located at the 5' and 3' ends of its own RNA to form a ribonucleoprotein complex (RNP). *In vitro* reactions with two protein subunits model in R2 show partial integration reactions: one protein subunit is bound to the 3' PBM of the R2 RNA and another bound to the 5' PBM of the R2 RNA, each hypothesized to be involved in integration reaction. In the presence of 3' PBM RNA, the R2 RNP binds to target DNA sequences upstream of the insertion site. The upstream subunit's RLE cleaves the first (bottom/antisense) DNA strand, releases a free 3' OH which is subsequently used as a primer by the upstream subunit's RT to prime first-strand cDNA synthesis (TPRT). In the presence of the 5' PBM RNA, the R2 protein binds DNA sequences downstream of the insertion site via ZF and Myb domains. The downstream subunit's RLE cleaves the second (top/sense) DNA strand only after the 5' PBM R2 RNA dissociates (i.e. no RNA state of the protein subunit) from the complex by the process of TPRT. Second-strand DNA cleavages, however, has been very difficult to show n *in vitro* reactions. *In vitro* assays are limited by a very narrow space of protein/RNA/DNA ratios. In addition, second-strand cleavage signals are usually very weak possibly due to either improper ratios of the reactants or absence of a suitable intermediate structure of target duplex DNA. Moreover, second-strand DNA synthesis is not observed with either of the isolated 5' PBM RNA and 3' PBM RNA [4] [13] [14] [15].

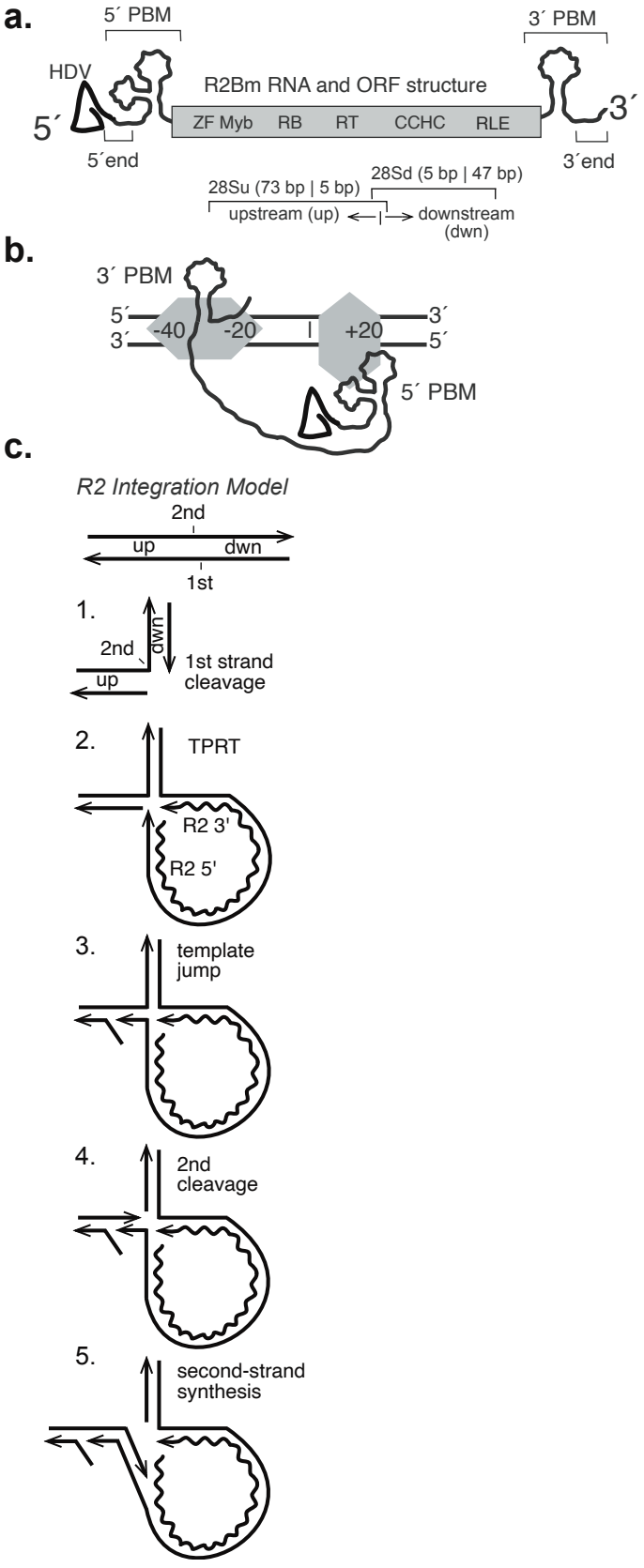


Figure 1. R2Bm ORF structure and integration mechanism. (A) R2Bm RNA (wavy line) and open reading frame (ORF) structure (gray box). The ORF encodes conserved domains of known and unknown functions: zinc finger (ZF), Myb motif (Myb), reverse transcriptase domain (RT), a cysteine histidine rich motif (CCHC) and a PD-(D/E)XK type RLE. RNA structures present in the 5′ and 3′ untranslated regions that bind R2 protein are marked as 5′ and 3′ PBMs, respectively [21]. Hepatitis Delta Virus (HDV) is indicated as triangular structure. **(B)** The R2 integration complex, as currently understood, is depicted bound to a segment of linear 28S rDNA (black parallel lines). An R2 protein subunit (gray horizontal hexagon) is bound upstream of the insertion site (vertical bar), and an R2 protein (gray vertical hexagon) subunit is bound downstream of the insertion site. The upstream subunit is associated with the 3 ′ PBM RNA, and the downstream subunit is associated with the 5 ′ PBM RNA. The footprints of the two protein subunits on the linear target DNA are indicated. The upstream subunit footprints from −40 bp to −20 bp, but it grows to just over the insertion site (vertical line) after first-strand DNA cleavage. The downstream subunit footprints from just prior to the insertion site to +20 bp [14 15]. **(C)** The R2 Integration Model. The straight lines are 28S DNA. The wavy line indicates the PBM R2 RNA. The four steps of the integration model are: (1) DNA cleavage of the bottom/first-strand of the target DNA; (2) TPRT; (3) a template jump/recombination event that generates an open '4-way' DNA junction/DNA integration intermediate; (4) second-strand DNA cleavage; and (5) second-strand DNA synthesis. Abbreviations: up (target sequences upstream of the insertion site), dwn (target sequences downstream of the insertion site) and TPRT [21].

In our previous paper, second-strand DNA cleavage and second-strand DNA synthesis in R2 were shown *in vitro*, where branched DNA integration intermediates mimicking *in vivo* DNA conformations and with association of the cDNA to the upstream target DNA (i.e. template switch/jump) were shown to be involved in full-length element insertion [21]. Template switch/recombination event is hypothesized to hinder protein binding upstream (-40 bp to -20 bp) of the target site [21]. Previous footprint data on linear DNA, based on two protein subunits model for element integration, have shown that R2 protein binds a large region of linear target DNA extending from 40 bp upstream to 10 bp downstream of the target site [18]. It is not known, however, if the protein could still recognize and bind to these regions in the context of branched DNA integration intermediate structure(s). Further, high-resolution footprints of specific DNA sequences in the context of DNA junctions that are important for protein binding remain unidentified. Recognition of target DNA is thought to be entirely controlled by the R2 protein. In addition, R2 protein binds specifically to 3' PBM R2 RNA and *in vitro* reactions have shown only this RNA as being crucial for TPRT reactions [27 28]. Second-strand cleavage does not require PBM RNA nor does the RNA itself gets involved in target site recognition, as any RNA sequences can support cleavage. The target site sequence upstream of the actual cleavage site is recognized by R2 protein-RNA complex followed by cleavage and TPRT. In the presence of 5' PBM RNA, R2 protein is known to bind sequences downstream of the target site on linear DNA, by the Myb domain. This paper provides evidences via DNase footprints on protein binding to Myb region (i.e. downstream sequence) in the context of branched DNA integration intermediate structure. It is possible that the R2 protein recognizes the open branched DNA intermediates and binds sequences both upstream and downstream of the target cleavage site by a single protein subunit and thus, efficiently perform full integration reaction.

MATERIALS AND METHODS

Nucleic acid preparation

Various oligonucleotides (oligos) containing specific target DNA (i.e. 28S R2 target DNA), non-specific DNA (i.e. non-target DNA) and R2 sequences were ordered from Sigma-Aldrich. R2 integration intermediates and analogs of presumptive integration intermediates were specifically engineered such that the resultant nucleic-acid constructs would consist of a combination specific target DNA, non-target DNA, and R2 derived sequences. For component oligos: see Supplementary Figure S1 for a list of oligos and their detailed sequences. The upstream (28Su) and downstream (28Sd) target DNAs are designated relative to the R2 insertion dyad within the 28S rRNA gene. To make each construct (i.e. DNA integration intermediate), twenty pmol of one of the component-oligos had been 5' end-radiolabeled (^{32}P), prior to annealing with 66 pmol of each of the remaining oligos together. Annealing reaction was carried out in a 1× TPRT buffer (10 mM Tris–HCl (pH 8.0), 5 mM $MgCl_2$, 200 mM NaCl) for 2 min at 95° C, followed by 10 min at 65° C, 10 min at 37° C and finally, 10 min at room temperature. The constructs were not subjected to gel purification post annealing in order to avoid any formation of partial junctions and also to prevent variations in DNA concentration. For R2 reactions, the junctions were either equalized by radioactive DNA counts or equal volume of annealed junctions were used.

R2Bm protein purification

His-tagged wild-type R2Bm protein was expressed in E. Coli (BL-21 cells) and affinity purified using Talon (Co++) resin as previously published [7]. Briefly, BL-21 cells containing R2 expression plasmid were grown overnight with antibiotic (kanamycin 50mg/mL) in LB broth and

then induced with IPTG. The induced cells were pelleted, resuspended, and gently lysed in a HEPES buffer containing lysozyme and triton x-100 followed by 20-hr spin to separate cellular nucleic acids and debris. The supernatant containing R2Bm protein was then purified over Talon resin (Clontech #635501) and finally eluted through Talon resin column. The eluted R2Bm protein was stored in protein storage buffer containing 50 mM HEPES pH 7.5, 100 mM NaCl, 50% glycerol, 0.1% triton X-100, 0.1 mg/ml bovine serum albumin (BSA) and 2 mM dithiothreitol (DTT) and stored at −20 ◦C. R2 protein sample and BSA standard titrations were run together on a sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE), followed by SYPRO Orange (Sigma #S5692) staining prior to addition of BSA for protein storage. R2 protein was quantified using FIJI software analysis on digital photographs [29].

DNA binding and DNA cleavage reactions

R2Bm protein binding to target DNA and DNA cleavage reactions were carried out as previously reported [21]. Each 13 ul reactions contained 80 fmol of radiolabeled DNA (i.e. DNA integration intermediates), and a dilution series of R2Bm protein, typically ≥ 420 - ≤ 0.40 fmol protein in TPRT buffer (10 mM Tris–HCl (pH 8.0), 5 mM $MgCl_2$, 200 mM NaCl). Reactions were then incubated at 37° C for 30 mins, allowing reactions to undergo top-strand cleavage followed by chilling the reactions on ice prior to loading onto 5% native polyacrylamide gel. After gel run, gels were dried and exposed to PhosphorImager screen or X-day film (Kodak film). The phophorimager screen was scanned using a phosphorimage (Molecular dynamics STORM 840) and resulting 16-bit TIFF images were linearly adjusted (levels command) so that the most intense bands were dark gray. Quantitation of adjusted TIFF files were carried out using FIJI software [29]. The ability of R2Bm protein to bind to and cleave the DNA construct was tested in the absence of

protein binding motif RNA (i.e. 5' PBM RNA and 3' PBM RNA). Binding assays were analyzed by native polyacrylamide gel electrophoresis (EMSA) and used for determining fraction bound. Companion denaturing (6 M urea) 6% polyacrylamide gels were used to determine fraction cleaved per bound unit of target DNA. A+G ladders were made using end-labeled DNA construct and were run alongside the reactions in the denaturing urea gels to aid in mapping cleavages. To determine cleavability for each construct, only reactions in the linear range on a bound versus cleaved graph were used and also only 20% to about 95% bound window was included as quantitation is problematic below and above that range.

DNase I Footprints

Fivefold scaled-up binding reactions (i.e. five times the amount of DNA and R2Bm protein compared to standard 13 ul reaction) were carried out in 35 ul reactions. After the R2Bm protein was bound to the target DNA (i.e. DNA integration intermediates), one-unit of DNase I (Promega) was added, and the complex incubated for two minutes at room temperature. Each reaction was then stopped by placing on ice. Then DNase I-treated reactions were separated on 5% native polyacrylamide gel (EMSA gel) to separate the bound from the free DNA. Each band (complex) of interest on EMSA gel were excised, and eluted into a crush and soak buffer containing 0.3 M sodium acetate, 1.0 mM EDTA and 0.1 % SDS, ethanol-precipitated, re-dissolved in 95% formamide buffer, and finally separated on a denaturing (6 M urea) 6% polyacrylamide gel. Each loaded sample on denaturing was equalized by radioactive DNA counts. The denaturing gel was dried onto filter-paper and then exposed to a PhosphorImager screen.

RESULTS

DNA sequence and DNA structure modulate second-strand DNA cleavage on integration analogs

R2 protein recognizes and binds to specific target DNA sequences followed by R2 insertion into a specific site in 28S rDNA. Earlier experiments with linear DNA have determined that the protein subunit bound upstream of the insertion site, in the presence of 3' PBM RNA, provides the endonuclease which cleaves the first-strand DNA and does the TPRT. Similarly, the protein subunit bound downstream of the insertion site possibly provides the endonuclease and is involved in the second-strand (i.e. top-strand) DNA cleavage [13 14 15]. In linear target DNA, the association of 5' PBM RNA is shown to be crucial for the protein subunit to bind to the downstream target DNA sequences. Interestingly, experiments have also shown that the second-strand DNA cleavage only ensues after the dissociation of the 5' PBM RNA from the downstream protein subunit [15]. R2 protein has been shown to recognize, bind and cleave DNA integration intermediates [21]. However, specific DNA sequences required by the R2 protein subunit(s), in the presence and absence of PBM RNAs is yet to be determined.

In the first part of this study, various DNA junction intermediates/substrates mimicking the products generated from the first-strand DNA cleavage and TPRT (See Figure 1c) were tested to determine the ability of the R2 protein to perform second-strand DNA cleavage. Each four-way branched DNA integration intermediate was prepared by annealing different oligos that were either derived from the specific 28S sequence or R2 sequences and/or consisted of "non-specific (i.e. non-target)" DNA sequences, either upstream or downstream from the insertion site. Also, branched integration intermediates have structures similar to that of the linear DNA (See Figure

1c (3)) where the downstream sequence is bent 90° and therefore the downstream "arm" is seen oriented upward "North" and the upstream 28S DNA arm is oriented to the left or "West" with heteroduplex (DNA/5' PBM RNA) arm on the "South" and "East" arm remaining a single oligo. For each DNA integration intermediate in Figure 2, there is an equivalent analog with the inclusion of gap and a flap 27 bp upstream of the insertion site. The gap and a flap upstream of the insertion site on DNA integration intermediates are assumed to provide more flexibility to the DNA structure which could then possibly allow R2 protein to gain more access into crucial DNA sequences important for second-strand cleavage.

Construct i (*1) has 120 bp of 28S target sequence while its equivalent analog has gap and a flap as seen in construct ii (*1). The 120 bp target sequence included 73 nt of 28S sequence upstream of the R2 insertion site and 47 nt of sequence downstream of the insertion site. The complementary oligonucleotides for both upstream and downstream sequences were annealed together to form the full branched DNA integration intermediate structures.

Construct iii (*1) and construct iv (*1) with gap and a flap and non-specific North arm (i.e. downstream sequence) consisted of 42 bp of mostly non-specific DNA sequences downstream and 5 bp of 28S DNA sequence just prior to second-strand DNA cleavage site.

Construct v (*1) and construct vi (*1) with gap and a flap and non-specific West arm (i.e. upstream sequence) consisted of 68 bp of mostly non-specific DNA sequences upstream of the insertion site; only the 5 bp prior to second-strand cleavage site remained 28S sequence.

For each construct, the 120 nt "top" (i.e. the sense) strand of the 28S gene was 5' end-labeled with ^{32}P to allow for tracking of R2 protein induced DNA cleavage events (i.e. second-strand cleavage events). Electrophoretic mobility shift assay (EMSA) was utilized to measure the ability of R2 protein to bind to each construct across a range of protein concentrations. Companion

denaturing polyacrylamide gels were used to assay for second-strand DNA cleavage. Quantitation of bound vs cleaved DNA are presented in the graphs: fraction cleaved as a function of fraction bound, and a bar graph representing the average cleaved per bound unit of DNA.

Neither the junction with full 120 bp 28S sequence (constructs i (*1) and construct ii (*1) with gap and a flap) nor the junction with non-specific downstream sequence (construct iii (*1) and construct iv (*1) with gap and a flap) were good substrates for second-strand cleavage (See graphs in Figure 2). While some instances of DNA cleavages were seen at or near protein excess for these constructs, none of them provided cleavages more than 20% of the bound unit of R2 protein. Only the DNA integration intermediates (construct v (*1) and construct vi (*1) with gap and a flap) with 28S derived downstream sequence showed significant amount of second-strand cleavage. Both DNA integration intermediate structures, construct v (*1) and construct vi (*1) with gap and a flap showed at least 30% and in excess of 45% of second-strand cleavage, respectively.

It is possible that the downstream sequence is important and/or the R2 protein is binding downstream of the insertion site, especially in the myb region at position +10 to +20. In order to determine specific sequences recognized and bound by the protein, DNase I footprint was utilized on best substrates for second-strand cleavage.

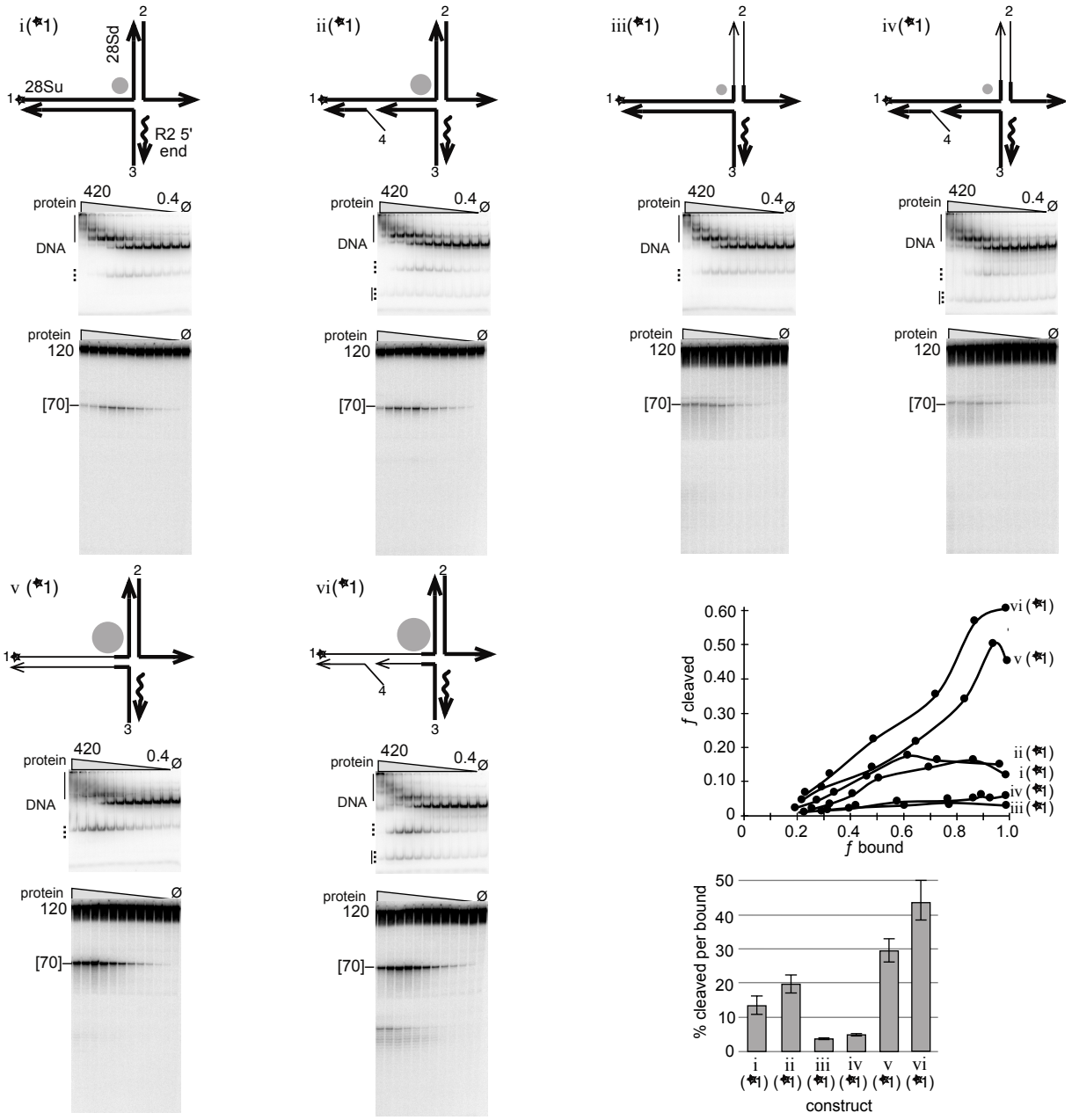


Figure 2. Second-strand DNA cleavage is affected by inherent structure of specific DNA integration intermediate. Various R2/28S derived junctions related to open DNA integration intermediate represented in step 3 of The R2 Integration Model (Figure 1) were tested for second-strand DNA cleavage. Several DNA integration intermediates without gap and a flap (i (*1), iii (*1), v (*1)) and with gap and a flap (ii (*1), iv (*1), vi (*1)) are tested for cleavability by the R2 protein. Each pair of junctions contain full 120 bp 28S target sequence (i (*1), ii (*1)), 73 nt of 28S sequence upstream of the R2 insertion site with non-specific downstream sequence (iii (*1), iv (*1)) and 47 nt of 28S sequence downstream of the R2 insertion site with non-specific upstream sequence (v (*1), vii (*1)). Each construct is diagrammatically bent at a 90◦ angle with the downstream (dwn) oriented toward the top of the page (i.e. the North arm). The star, "*", indicates the DNA oligo that was 5 ′ end-labeled to track DNA binding and cleavage. Each arm/DNA oligo of a construct is labeled with a number. Thick lines on each construct represent 28S DNA; in constructs iii, iv, v and vi, the thin lines represent non-specific sequences. The squiggly line on each construct is a 25

bp 5' PBM RNA derived sequence. Below each of the construct cartoons are the native (EMSA) gels and corresponding denaturing gels used to analyze DNA binding (EMSA) and DNA cleavage (denaturing) of the given DNA construct by R2 protein. Each DNA binding and cleavage assays were carried out in a 13 ul reaction and contained 80 fmol of 5' end-labeled construct DNA and 420–0.4 fmol of R2 protein (gray triangle). All EMSA gels were quantified where the bands above the full construct DNA in the mock purified protein (Ø; protein purified from an empty expression vector) control lane were subtracted out of the bound signal in the experimental lanes. On each EMSA gel, areas of the gel where the bound DNA signal resides are represented by solid vertical lines. Bound DNA signals include the well, the smear and the gel migrating complexes and are counted as bound DNA for quantitation. The unbound (free) DNA is indicated as "DNA" on EMSA gels. DNA bands located below the unbound DNA that is seen as increasing with protein concentration were also counted as bound DNA and are considered as "released cleavage products." In construct ii, iv and vi, the released cleavage products are seen comigrating with partial junctions (dotted lines) that is present in the control lanes. The control lane partial junction signal was subtracted from the experimental lane's co-migrating bound signal. The remaining partial junctions (dotted line) were counted as unbound DNA in the experimental lanes. The main band in the mock purified (Ø) lane represents the unbound junction DNA (DNA). The un-cleaved radiolabeled oligo for each construct is indicated (120 bp) as well as the size and migration of the band resulting from second-strand cleavage by brackets on the denaturing gel. The DNA binding and DNA cleavage results are plotted on three graphs: (i) fraction cleaved as a function of fraction bound for reactions where roughly 20–95% of the DNA was bound; and (ii) a bar graph reporting the average percentage cleaved products per bound unit of DNA (fraction bound) for reactions in the linear part of the second graph. The diameter of the gray dot next to each construct cartoon reflects the relative cleavability.

The branched integration intermediates show inherent structure in DNase footprint

DNA footprint is affected by inherent structure of branched DNA intermediates, as previously reported [23]. R2 protein sees the structure of branched integration intermediates, especially the template jump (West arm) structure and the single stranded East arm. The protein is also found to associate with the downstream sequence from the insertion site and that the upstream sequence is not as important for binding. With these findings, DNase I footprints were carried out for the "best" cleavers (i.e. for second-strand cleavage).

DNase I footpritns were carried out with wild type (WT) R2 protein and endonuclease mutant (KPD mutant) in the presence and absence of PBM RNAs (i.e. 3' PBM RNA (data not shown in this chapter) and 5' PBM RNA). Even in the absence of protein, the branched intermediate structures have a lot of structure to it (i.e. no protein lanes in footprint data, See Figure 3). In fact, these branched intermediates have specific DNA structures and binds the R2 protein over a large region of DNA. The DNA integration intermediates used in this study also seem to

have an inherent structure with visible footprint spanning "x" bases across to either side of the junction. In Figure 3, the junction with 120 bp 28S sequence plus the gap and a flap (construct ii (*1)) appears to have significant footprint regions upstream of the insertion site between -35 to -18 as well as between -11 to +15 (See Ø + w/DNase I lane for construct ii (*1) in Figure 3). Similarly, the junction intermediate with mostly non-specific West arm (construct v (*1)) shows footprint between positions -11 to +15 (See Ø + w/DNase I lane for v (*1) in Figure 3). Since DNA integration intermediates footprint due to their inherent structure, often times, determining specific sequences through footprint data is difficult.

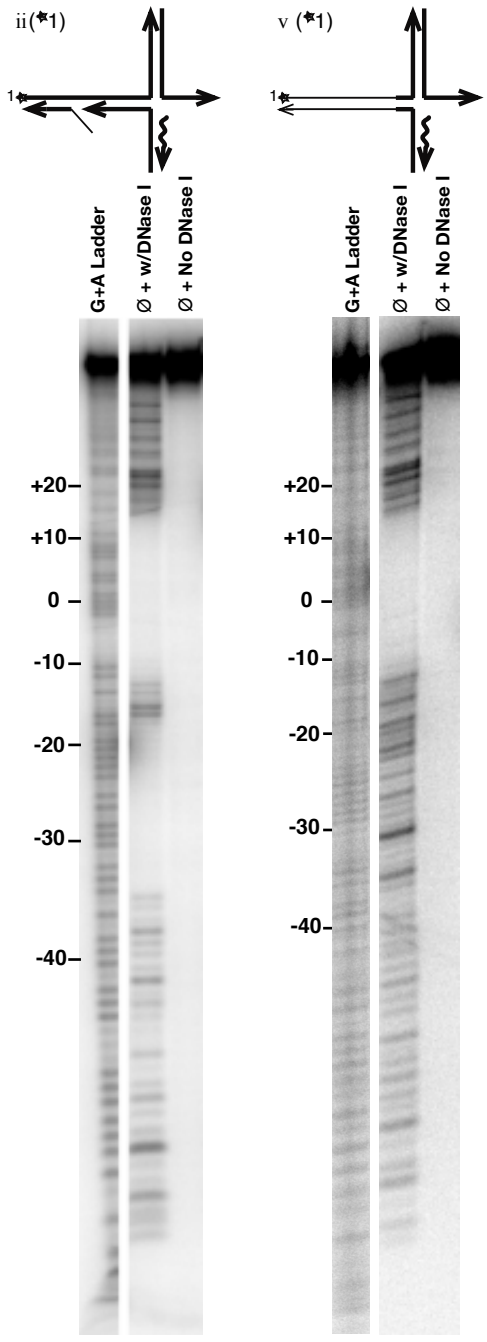


Figure 3. Branched integration intermediates showing structure-derived footprints. Branched DNA junction with full 28S sequence plus gap and a flap (construct ii (*1)) show structure-derived footprint spanning both upstream sequences between -35 to -18 and between positions -11 to +15. Branched DNA junction with mostly non-specific upstream sequence (construct v (*1)) show structure derived footprint between positions -11 to +15. The star, "*", indicates the DNA oligo that was 5 ′ end-labeled to track for DNA binding and cleavage. Each arm/DNA oligo of a construct is labeled with a number. Thick lines on each construct represent 28S DNA. Thin lines represent non-specific sequences. The squiggly line on each construct is a 25 bp 5' PBM RNA derived sequence. Also, G + A ladder is used to determine the exact location on sequence that show footprint. Both upstream and downstream positions are indicated next to G+A ladder. The Ø + w/DNase I lane indicates reaction with no R2 protein but with DNase I treatment. Ø + No DNase I lane indicates reaction with no R2 protein and no DNase I treatment.

DNase I footprint in the absence of R2 PBM RNA

In Figure 4, all constructs v (*1), v *(2), and v (*3) contain 68 bp of mostly non-specific DNA sequences in the West arm upstream of the insertion site; only the 5 bp prior to second-strand cleavage site remained 28S sequence. Complementary oligonucleotides were annealed to form a DNA integration intermediate with only the East arm left as non-duplex. Similarly, in Figure 5, constructs ii (*1), ii (*2), ii (*3) and ii (*4) contain 120 bp of 28S sequence with gap and a flap—73 nt of 28S sequence upstream of the R2 insertion site and 47 nt of 28S sequence downstream of the insertion site. Complementary oligonucleotides for both upstream and downstream segments, except the East arm of the DNA integration intermediate were annealed together to form the DNA integration intermediate structure. DNase I footprint reactions were carried out in the absence of PBM RNA for both panels. However, the footprint experiments in two panels differed in R2 protein that was used: WT R2 protein in Figure 4 and KPD (endonuclease mutant) protein in Figure 5.

DNase I footprint analyses on DNA integration intermediates were carried out with R2 protein (either WT or KPD mutant) titration series. Each panel had reactions where the R2 protein was bound to junction intermediates followed by DNase I treatment (See details on Materials and Methods). Reaction in lanes (Ø + w/DNase I) in both panels were void of R2 protein, nonetheless, structure derived footprints were clearly visible (See Figure 3). Also, the "well complexes" on denaturing gels showed some variant of footprints; however, any extrapolation from this data is difficult since most junction intermediates are bound in excess protein.

In Figure 4, branched integration intermediates with non-specific upstream and specific sequence downstream of the insertion site show protein binding to the myb region at position +15

to +18 in construct v (*1) and at position +10 to +18 in construct v (*2). Both instances of footprint results indicate protein binding to downstream myb region.

Interestingly, there are instances of high molecular weight stuffs located above full-length DNA seen on denaturing gels. We are unclear about the reason and mechanism as to how these high molecular weight stuff result in the presence of R2 protein. See discussion section for potential explanations.

In Figure 5, branched integration intermediates with full 120 bp target sequence plus gap and a flap was used for DNase I footprints. Most footprints seen on construct ii (*1) were at positions -35 to +15 but the footprint itself was largely derived from inherent structure of the integration intermediate. Similarly, construct ii (*2) does seem to indicate binding at the myb region, but is not conclusive since the footprint signals were significantly small. Footprints on construct ii (*3) and construct ii (*4) were not informative enough.

Any indications of protein binding to upstream sequences either in the construct ii or v are not conclusive because these branched intermediate structures exhibit large structure-derived footprints. However, construct ii (*1) show hypersensitive sites between -18 to -20 on the footprints (See Figure 5) which is suggestive of upstream binding.

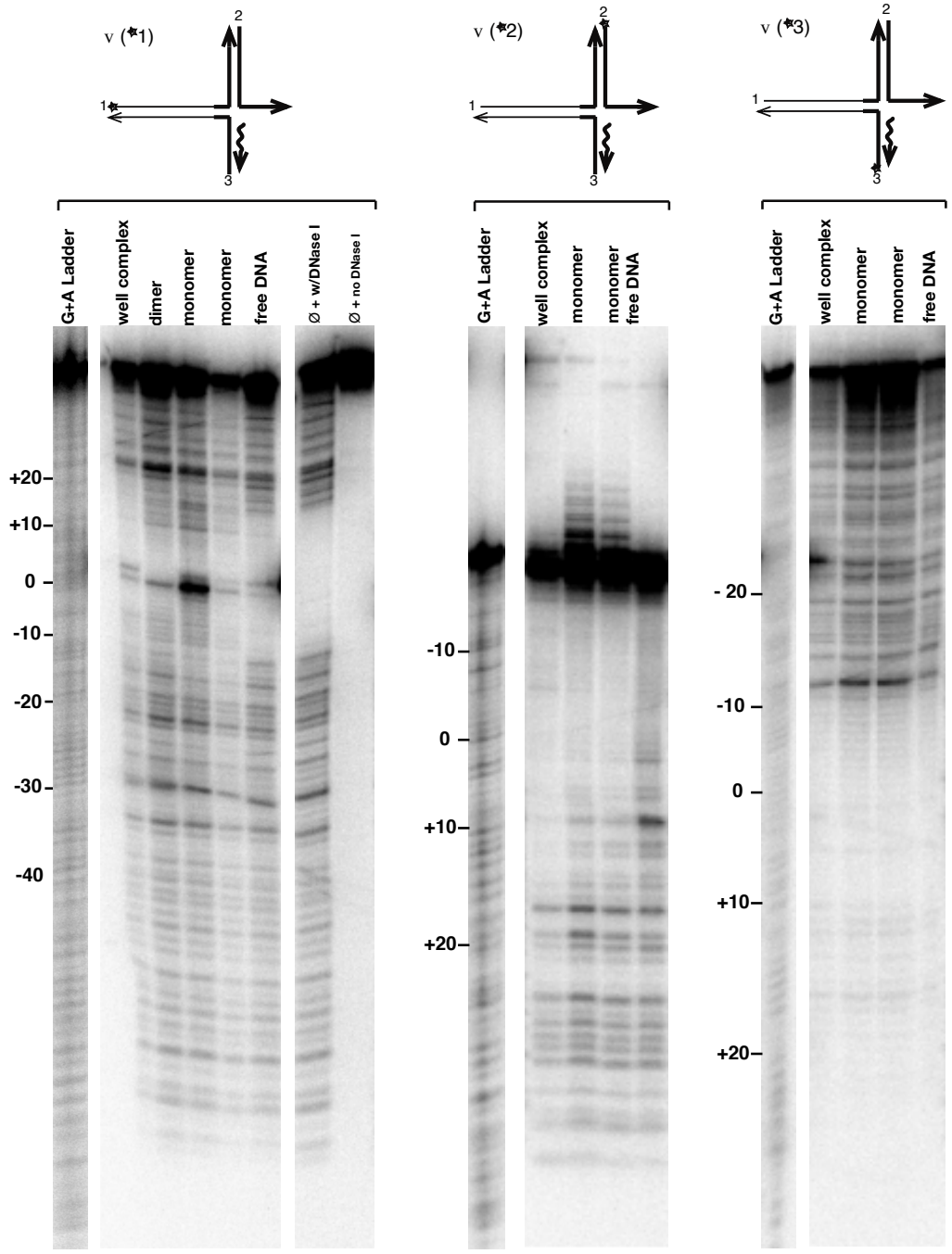


Figure 4. DNase I footprint analysis of WT R2 protein on DNA integration intermediate with non-specific upstream sequence. DNA integration intermediates with mostly non-specific upstream sequence (constructs v (*1), v (*2) and v (*3)) were used in DNase I footprint analyses. m). The star, "*", indicates the DNA oligo that was 5 ′ end-labeled to track DNA binding and cleavage. Each arm/DNA oligo of a construct is labeled with a number. Thick lines on each construct represent 28S DNA. Thin lines represent non-specific sequences. The squiggly line on each construct is a 25 bp 5' PBM RNA derived sequence. Also, G + A ladder is used to determine the exact location on sequence that show footprint. Both upstream and downstream positions are indicated next to G+A ladder. Below each of the construct cartoons are the corresponding footprint data generated on denaturing gel. The analyses were performed on WT R2 protein titration series with DNase I digestion. Lanes for each junction: well complex, dimer, monomer and free DNA are indicated. The Ø + w/DNase I lane indicates reaction with no R2 protein but with DNase I treatment. Ø + No DNase I lane indicates reaction with no R2 protein and no DNase I treatment.

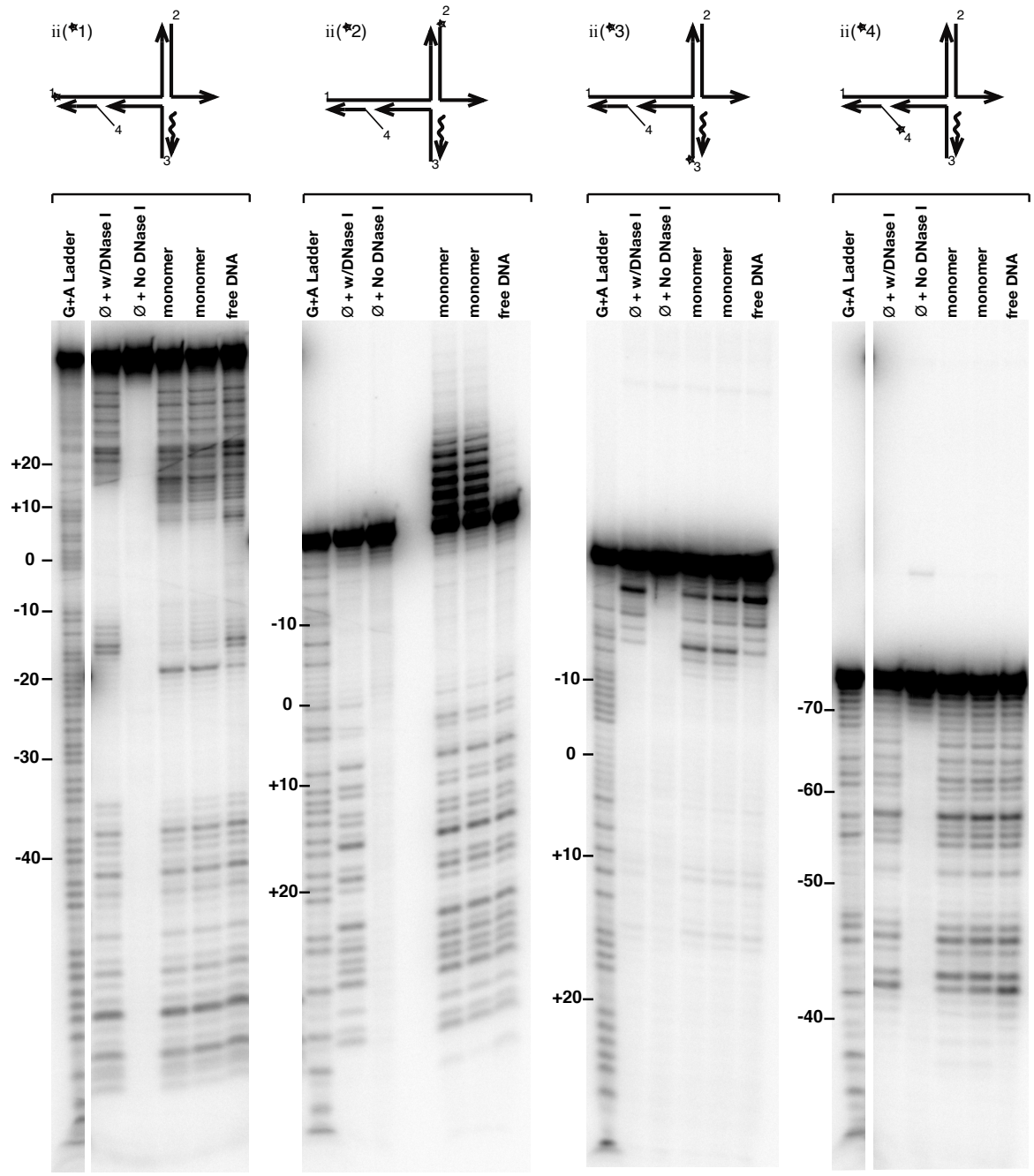


Figure 5. DNase I footprint analysis of KPD mutant R2 protein on DNA integration intermediate with 28S sequence + gap and a flap. DNA integration intermediates with 120 bp of 28S sequence, constructs ii (*1), ii (*2), ii (*3) and ii (*4), were used in DNase I footprint analyses. The star, "*", indicates the DNA oligo that was 5 ′ end-labeled to track DNA binding and cleavage. Each arm/DNA oligo of a construct is labeled with a number. Thick lines on each construct represent 28S DNA. Thin lines represent non-specific sequences. The squiggly line on each construct is a 25 bp 5' PBM RNA derived sequence. Also, G + A ladder is used to determine the exact location on sequence that show footprint. Both upstream and downstream positions are indicated next to G+A ladder. Below each of the construct cartoons are the corresponding footprint data generated on denaturing gel. The analyses were performed on KPD R2 protein titration series with DNase I digestion. The Ø + w/DNase I lane indicates reaction with no R2 protein but with DNase I treatment. Ø + No DNase I lane indicates reaction with no R2 protein and no DNase I treatment.

DNA binding to downstream sequence in the presence of 5' PBM RNA

In Figure 6, all constructs v (*1), v *(2), and v (*3) contain 68 bp of mostly non-specific DNA sequences in the West arm upstream of the insertion site and only retained 5 bp of 28S sequence just prior to second-strand cleavage; complimentary sequences are annealed to form DNA integration intermediate, except on

Figure 6 shows the footprint data for constructs v in the presence of 5' PBM RNA with WT R2 protein. As previously explained, DNase I footprint analyses were carried out on junctions to determine whether the presence of 5' PBM RNA show preferential binding to downstream sequence. Both the non-specific upstream sequence and 5' PBM RNA serves as a "control" for binding reaction as the protein is forced to bind downstream sequence as previously shown on linear DNA [14 15]. Nonetheless, construct v in Figure 6 showed some interesting but expected results where the protein is seen protecting downstream sequence where the myb region resides.

Footprint signals on construct v (*2) is shown at positions +5 to +18 region (downstream sequence) which possibly indicates that binding to the downstream region of the integration intermediate structure is important. In fact, strong footprint signals on the myb region is indicative of this region in protein binding. Higher resolution methylation interference footprinting studies have indicated that the myb region resides on position +12 to +15 (GGAG) on linear DNA [13].

In construct v (*1), while footprint is seen at downstream sequence (i.e. near the myb region at position +15 to +18), much of the footprint signals are occluded by structure-derived footprint. Also, similar to previously determined hypersensitive sites (position -2 to -1) on linear DNA, one instance of hypersensitive site can be seen on construct v (*1) at or near comparable position [18]. Construct v (*3) did not present any instance of footprint, possibly affected by the inherent structure of the construct itself.

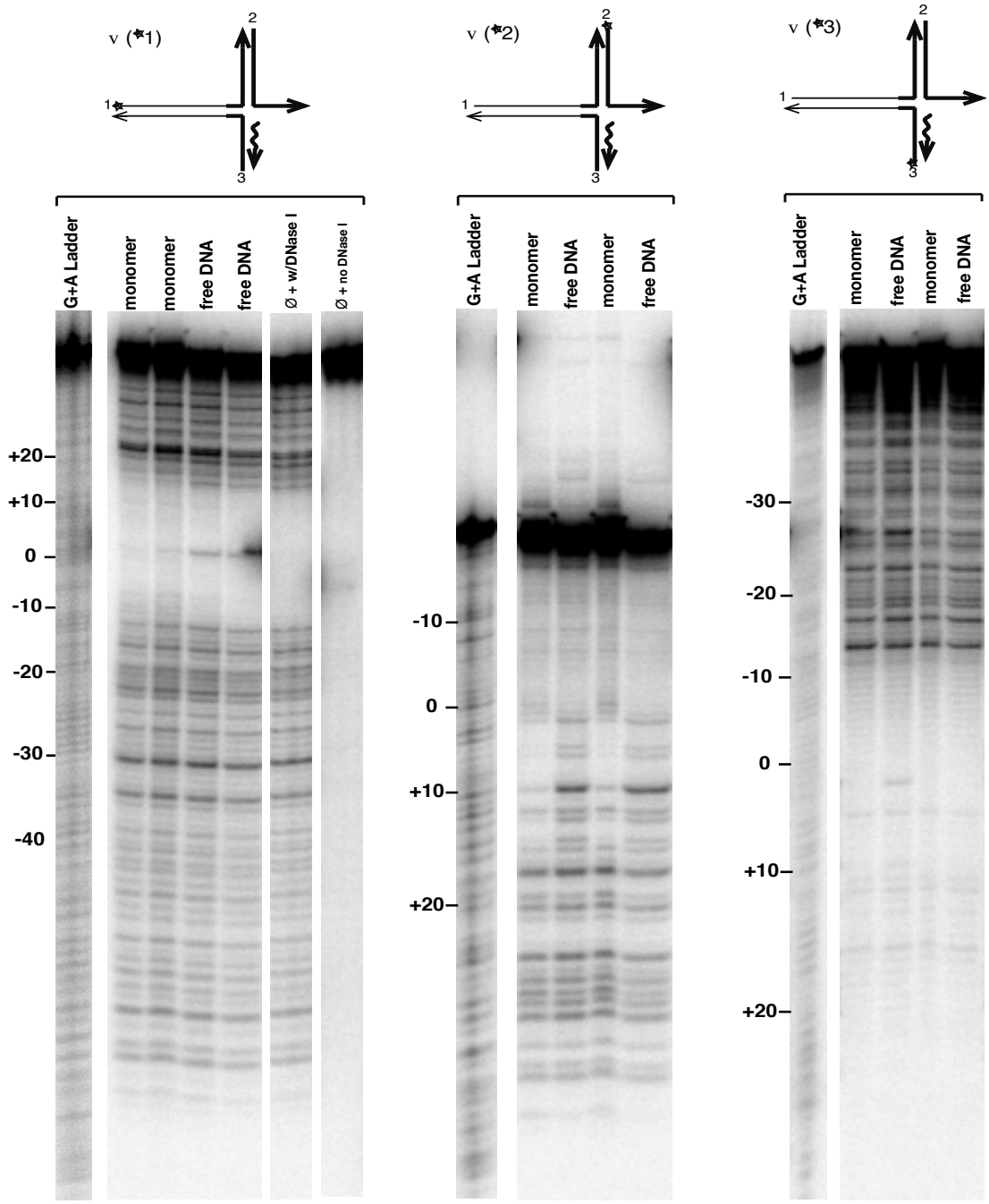


Figure 6. DNase I footprint analysis of WT R2 protein on DNA integration intermediate with non-specific upstream sequence and 5' PBM RNA. DNA integration intermediates with mostly non-specific upstream sequence (constructs v (*1), v (*2) and v (*3)) were used in DNase I footprint analyses. The star, "*", indicates the DNA oligo that was 5 ′ end-labeled to track DNA binding and cleavage. Each arm/DNA oligo of a construct is labeled with a number. Thick lines on each construct represent 28S DNA. Thin lines represent non-specific sequences. The squiggly line on each construct is a 25 bp 5' PBM RNA derived sequence. Also, G + A ladder is used to determine the exact location on sequence that show footprint. Both upstream and downstream positions are indicated next to G+A ladder. Below each of the construct cartoons are the corresponding footprint data generated on denaturing gel. The analyses were performed on WT R2 protein titration series with DNase I digestion. The Ø + w/DNase I lane indicates reaction with no R2 protein but with DNase I treatment. Ø + No DNase I lane indicates reaction with no R2 protein and no DNase I treatment.

DISCUSSION

Our previous paper (Chapter 2), determined that linear DNA, first-strand-cleaved linear-DNA, and the TPRT product were not good substrates for second strand cleavage. The paper established that a hypothetical open "4-way" branched structure was a good substrate for second strand DNA cleavage [21]. This paper further explored the DNA sequence and structure requirements for second-strand DNA cleavage, particularly, the sequence requirement of the north arm. The north arm equates to the 28S rDNA sequences located downstream of the insertion site. The R2 myb domain had previously been reported to bind to linear 28S rDNA 10-20 bp downstream of the R2 insertion site. The myb bound to this region as an isolated polypeptide as a full-length R2 protein in the presence of 5' PBM RNA [18] [13]. Therefore, myb binding region might be important for recognizing the north arm of the open "4-way" junction.

The inherent structure of the open "4-way" junction was probed by DNase I footprinting (Figure 2 and control lanes in subsequent figures). Junctions without the gap and flap showed a structure footprint in the central junction region that spanned about 12 by into the west and north arms, similar the structural footprint of a Holliday junction [23]. The presence of a gap and flap on the west arm generated a structure footprint in the vicinity of the gap/flap (-34 to -18) in addition to the footprint of the central junction area. The south arm was impervious to DNase I as it was DNA-RNA hybrid. It was difficult to determine if the single stranded east arm was protected or not. The fact that no signal was detected for the single stranded could be interpreted as the single stranded region was rapidly (and completely cleaved) or that it was structurally protected from DNase I cleavage.

The cleavage data for each of the junctions in Figure 2 indicate that structure, beyond the immediate junction area, is important in the west arm. The west arm gap with flap was generally

stimulatory as was non-specific sequence. In the north arm, sequence was important as a non-specific north arm was inhibitory.

We tried to determine how the protein was binding to a junction of high cleavability (v) *versus* a junction of low cleavability (ii) by DNase I footprinting. These efforts were hampered by the fact that the structural footprint of the junctions (-34 to -18, -12 to +12) coincides with most of the known DNA footprint regions with R2 protein (-40 to -20, -40 to +7, and -3 to +20) and the fact that the south and west arms were not amenable to the DNase footprint approach [13–15, 18]. That said, we were able to detect a DNA footprint in the region where we would expect the myb to bind (+7 to +20) on junction v (high cleavability) and not on junction ii (low cleavability). We were unable to tell to what extent R2 protein is binding to sequences on the west arm on junction ii, beyond the existence of R2 protein induced DNase I hypersensitive sites, because of the interfering structural footprint. No DNase footprint is observable on the nonspecific west arm of junction v. Our previous paper, however, showed that the presence of the template jump (i.e., the west arm) is, nonetheless important.

Why is junction ii not able to be cleaved as well as junction v, when junction ii has all specific sequence and the correct structural components? R2 protein binds to target sequence and inserts specifically to 28S rDNA. This integration reaction takes place in a very systematic manner *in vivo*. When the R2 protein-RNA complex binds the target DNA the protein-RNA-DNA complex adopts the pre-cleavage conformation. This conformation is driven primarily by the 3' PBM RNA and upstream target DNA sequences. After first-strand DNA cleavage the protein-RNA-DNA complex adopts the pre-TPRT conformation. After TPRT begins, the protein-RNA-DNA complex adopts the conformation driven by the 5' PBM and the three arms of the DNA-RNA intermediate. At the end of TPRT, the protein-DNA complex is thought to be in the template-jump conformation

(i.e., no PBM RNA bound). After template jump, the open "4-way" junction is formed and the protein-DNA complex adopts the second-strand DNA cleavage conformation.

The R2 protein likely has an extensive nucleic acid binding surfaces with which to interact with the 3' PBM RNA, 5' PBM RNA, upstream linear target DNA, downstream linear target DNA, nicked linear target, early TPRT product, late TPRT product, and template jump product). Some of these surfaces might interact with specific nucleic acid sequences, others might read nucleic acid structure, and other contacts are likely non-specific.

In vitro, there are a number of ways the R2 protein, in the absence of element RNA, can bind to the full-sequence full-structure substrate that is junction ii. Binding to the junction *in vitro* is not arrived at in a step wise manner, as it would be *in vivo*. By replacing specific DNA with nonspecific DNA and by adding or subtracting structural components on the DNA substrate, the ratio of the different protein-DNA conformations being sampled and then fixed by tight binding is reduced. For example, replacing the west arm with non-specific sequence reduces that chance that the R2 protein will try to bind to this region as if it were binding to linear DNA or TPRT product. Indeed, on linear DNA the protein prefers to the upstream sequences on linear DNA in the absence of RNA [18].

For second strand cleavage, we think the protein is binding to the overall DNA structure of the open "4-way"as well as to specific sequences. A major structural component that required for second-strand DNA cleavage is the presence of the template jump (Chapter 2) [21]. A major specific sequence that is stimulatory is the +5 to +20 region on the north arm. This area footprints on junction v (Figure 4). That area also coincides with the area that, when turned into nonspecific sequence, leads to inhibition of second strand cleavage (junctions iii and iv, Figure 2).

This chapter is pending additional experiments to turn it into a publication worthy paper. Additional junction constructs with smaller regions turned into nonspecific sequence will be generate and tested in cleavage reactions. Regions to turn into nonspecific sequences +5 to +10 and +10 to +20 on the north arm as well the central junction region. and non-specific region in the myb domain among others.

DNA footprints on a set of after-the-fact logical constructs might be helpful and necessary. DNA footprint experiments using higher resolution DNA footprinting techniques, like hydroxyl radical footprint, methylation interference, and/or ethylation interference might be more successful at separating, or peering beyond, the structural DNA footprint so as to observe the DNA-protein footprint.

An unknown reaction is occurring between the 3' end of the east arm and the R2 protein as labeled DNA larger than the oligo (oligo #2) length is observed in denaturing poly acrylamide gels in the presence of R2 protein. This reaction needs to be further explored.

ACKNOWLEDGEMENTS
We thank Santosh Dhamala for helpful discussions along the way and for critical reading of the manuscript.

FUNDING
Phi Sigma Graduate Student Biology Society to B.B.K. (in part); University of Texas Arlington Research Enhancement (REP) Grant Program (to S.C.). Conflict of interest statement. None declared.

REFERENCES

1. Craig,N.L. (2002) Tn7 In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 423-456.
2. Richardson,S.R., Doucet,A.J., Kopera,H.C., Moldovan,J.B., Garcia-Perez,J.L. and Moran,J.V. (2015) The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061.
3. Zingler,N., Weichenrieder,O. and Schumann,G.G. (2005) APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res* **110**, 250-268.
4. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605.
5. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910.
6. Moran,J.V. and Gilbert,N. (2002) Mammalian LINE-1 Retrotransposons and Related Elements In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 836-869.
7. Yang,J., Malik,H.S. and Eickbush,T.H. (1999) Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A* **96**, 7847-7852.
8. Eickbush,T.H. (2002) R2 and Related Site-Specific Non-Long Terminal Repeat Retrotransposons In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 813-835.
9. Eickbush,T.H. and Eickbush,D.G. (2015) Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011.
10. Eickbush,T.H. and Malik,H.S. (2002) Origins and Evolution of Retrotransposons In Craig, NL, Craigie, R, Gellert, M and Lambowitz,A.M. (eds.), *Mobile DNA II*. ASM Press, Washington, DC, pp. 1111-1146.
11. Malik,H.S., Burke,W.D. and Eickbush,T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**, 793-805.
12. Mahbub,M.M., Chowdhury,S.M. and Christensen,S.M. (2017) Globular domain structure and function of restriction-like-endonuclease LINEs: similarities to eukaryotic splicing factor Prp8. *Mob DNA* **8**, 16.
13. Christensen,S.M. and Eickbush,T.H. (2005) R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**, 6617-6628.
14. Christensen,S.M., Bibillo,A. and Eickbush,T.H. (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468.
15. Christensen,S.M., Ye,J. and Eickbush,T.H. (2006) RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* **103**, 17602-17607.
16. Han,J.S. (2010) Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA* **1**, 15.

17. Fujiwara,H. (2014) Site-specific non-LTR retrotransposons. *Microbiol Spectrum* **3**, MDNA3-0001.
18. Christensen,S. and Eickbush,T.H. (2004) Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* **336**, 1035-1045.
19. Gladyshev,E.A. and Arkhipova,I.R. (2009) Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* **448**, 145-150.
20. Govindaraju,A., Cortez,J.D., Reveal,B. and Christensen,S.M. (2016) Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287.
21. Khadgi,B.B., Govindaraju,A. and Christensen,S.M. (2019) Completion of LINE integration involves an open '4-way' branched DNA intermediate. *Nucleic Acids Res* **47**, 8708-8719.
22. Mukha,D.V., Pasyukova,E.G., Kapelinskaya,T.V. and Kagramanova,A.S. (2013) Endonuclease domain of the Drosophila melanogaster R2 non-LTR retrotransposon and related retroelements: a new model for transposition. *Front Genet* **4**, 63.
23. Churchill,M.E., Tullius,T.D., Kallenbach,N.R. and Seeman,N.C. (1988) A Holliday recombination intermediate is twofold symmetric. *Proc Natl Acad Sci U S A* **85**, 4653-4656.
24. Middleton,C.L., Parker,J.L., Richard,D.J., White,M.F. and Bond,C.S. (2004) Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res* **32**, 5442-5451.
25. Eickbush,D.G., Burke,W.D. and Eickbush,T.H. (2013) Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441.
26. Eickbush,D.G. and Eickbush,T.H. (2010) R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol*
27. Luan,D.D. and Eickbush,T.H. (1995) RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* **15**, 3882-3891.
28. Mathews,D.H., Banerjee,A.R., Luan,D.D., Eickbush,T.H. and Turner,D.H. (1997) Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3**, 1-16.
29. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B., Tinevez,J.Y., White,D.J., Hartenstein,V., Eliceiri,K., Tomancak,P. and Cardona,A. (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682.

CHAPTER 4

Conclusions

Brijesh B. Khadgi and Shawn M. Christensen

Department of Biology, University of Texas at Arlington

Arlington, TX 76019, USA

Conclusions

Restriction like endonuclease (RLE) encoding LINEs such as R2 specifically integrate into the host genome via a process called target primed reverse transcription (TPRT). R2 encodes a multifunctional protein which uses element encoded endonuclease to cleave the target DNA. This cleavage releases a 3'-hydroxyl at the site of a nick which is then used as a primer to prime reverse transcription of the element mRNA [1 2 3]. Much of the first half of the integration reaction (i.e. first-strand cleavage and first-strand synthesis/TPRT) has been extensively studied; however, it wasn't until recently where branched integration intermediate structures (i.e. "4-way junctions") were discovered to resolve long standing question regarding the completion of integration reaction of R2 [4]. Theoretically, after the first-strand cleavage and TPRT, the second-strand cleavage ensues and the insertion mechanism is completed with second-strand DNA synthesis [5 1]. While branched integration intermediates resolved immediate questions as to which structure(s) and/or protein-nucleic acid requirements are necessary for second half of the integration reaction, there still remains much uncertainties including specific sequences required for protein binding and determining whether the myb region play a role in second-strand cleavage. Here, we have attempted to elucidate most of the concerns through protein cleavage assays and DNA footprints.

Both APE-bearing LINEs and RLE-bearing LINEs encode for a single multifunctional protein for integration mechanism. The recent experiments involving branched integration intermediates suggest that the single protein subunit might be responsible for full integration of the element. The idea of a single protein subunit performing full integration reaction can be explained through "rock and roll" model. In this model, the protein binds upstream of the insertion site in the presence of 3' PBM RNA, upon which the target DNA unwinds locally at the R2 insertion site. The protein subunit bound upstream of the insertion site is then rocked into a specific

conformation allowing for first-strand cleavage at the R2 site via the element encoded endonuclease. Next, the reverse transcriptase is rocked in place which would then initiate TPRT. The 3' PBM RNA is removed from the protein as the heteroduplex is formed. For second half of the integration, the protein binds downstream of the insertion site in the presence of 5' PBM RNA, and results in a specific confirmation which allows the protein to "roll" over towards downstream region (i.e. North arm in the case of branched integration intermediate). At this stage, the protein is also making enough contacts with upstream region (i.e. West arm) and East arm of the branched structure. Once the TPRT is completed, the 5' PBM is dissociated from the protein due to the formation of a heteroduplex followed by template jump. Once the endonuclease from downstream protein subunit rolls in place, second-strand cleavage occurs. The reverse transcriptase is then finally rocked into place followed by the second-strand synthesis [4].

Integration of RLE LINEs are dependent upon element encoded endonuclease. In recent studies, the endonucleases are determined to be a member of larger endonuclease family, the PD-(D/E)XK endonucleases [6]. While most early identified of PD-(D/E)XK endonucleases were limited to type II restriction enzymes such as EcoRI, BamHI, and FokI, various other newly found endonucleases are involved in many functions including DNA repair, Holliday junction resolutions, and RNA processing [7] [8] [9] [10] [11] [12]. While various endonucleases do not show any sequence homology with one another, most of these share a common conserved core. This specific structure of the conserved core has been found to consist a four-stranded mixed β-sheet flanked by α-helix on each side [13].

The element encoded endonuclease and/or R2 protein can associate with double-stranded region and cleave a nearby single-stranded region in branched integration intermediate strucutres [4]. It has been hypothesized that the protein associating with the downstream sequence (i.e. double-

stranded region) is possibly providing the endonuclease to perform second-strand cleavage and that the cleavage site could actually be a single-stranded region. This case is further substantiated by inclusion of gap and a flap in the structure of branched integration intermediates. The flexibility rendered by the presence of this gap and a flap allows the R2 protein to sort of gain more access to the cleavage site and therefore perform second-strand cleavage via the endonuclease. Indeed, constructs with gap and a flap show better cleavage than their companion constructs with no gap and a flap.

Branched integration intermediates are key to second-strand cleavage and second-strand synthesis

Chapter 2 introduces an idea of branched integration intermediates (i.e. "4-way" like junctions) which has been the major step forward in understanding the mechanistic pathway for R2 integration reaction. R2 protein can recognize various target sequences including linear DNA, three-way junctions/constructs and/or "4-way" branched integration intermediates. While linear DNA and three-way junctions are recognized and bound by the protein, second-strand cleavage from these substrates are abysmal (i.e. less that 5% of the total bound unit of protein) (See Figure 2, Chapter 2). Instead, open "4-way" integration intermediate show significant increase in the amount of cleavage (between 10% to 22%) and therefore are considered as the good substrates for second-strand cleavage (See Figure 3, Chapter 2).

Various combination of 28S/R2 derived sequences were used to engineer specific branched integration intermediates to experiment with in cleavage assays. Second-strand cleavage occurs when the protein is bound to the double-stranded region and cleaves on a single-stranded region

near the R2 site (See construct ii and iii; Figure 3, Chapter 2). Specific requirements such as in construct xii, where there is non-specific sequence upstream (West arm), 28S sequence downstream (North arm) and single non-heteroduplex East arm, is shown to be the best candidate for second-strand cleavage (See graphs from Figure 3, Chapter 2). Similarly, introduction of gap and flap allowed for a sort of flexibility in the overall structure of the branched integration intermediate. This theory is substantiated by the result seen for construct xvi (See Figure 3, Chapter 2) where the substrate/junction show respectable 15% of second-strand cleavage. Part of this result would suggest that protein is associating both upstream and downstream sequences in the branched structure, possibly in the central region. Hence, the idea of R2 protein being a single, multimeric and multifunctional protein is quite conceivable; however, additional experiments including cleavage assays and footprint were warranted. Chapter 3 was a step forward into this assumption. Second-strand cleavage via branched integration intermediates allowed us to study second-strand synthesis. In fact, Chapter 2 puts forward a full integration reaction for RLE LINEs via R2 integration mechanism. Cleavage of the branched integration intermediate generated a specific primer-template important for the synthesis. Instances of second-strand synthesis is seen on "best" substrates for second-strand cleavage and include constructs v, xii, xvi (partially), and xvii (partially). Second-strand cleavage were seen, however, in only higher end of the protein titration series on the denaturing gels (See Figure 4, Chapter 2).

Sometimes, on denaturing gels, there is a signal seen above the full-length oligo (See Figure 4 in Chapter 2, and Figure 4, 5 and 6 in Chapter 3). We postulate that these signals are the result of the original full-length oligo being extended by the protein. Indeed, R2 protein is known to extend any 3' end of the template, either on a *cis* or *trans* [14 15].

R2 protein associates with downstream sequences

From previous studies on linear DNA, we know that R2 protein binds to downstream myb region between positions +10 to +15 [1]. DNase I footprint studies (detailed in Chapter 3) suggest a role of the myb region in binding of the protein to the junction for second-strand DNA cleavage as well.

The junctions tested in Chapter 3 were engineered in order to help identify DNA sequences necessary for protein binding and second-strand cleavage. Undeniably, R2 protein can recognize and therefore bind to branched structures. This idea is further corroborated by the inclusion of gap and a flap where the second-strand cleavage is shown to be better than companion constructs (i.e. same backbone of the structure but without gap and a flap). For instance, DNA integration intermediate structure with non-specific upstream shows R2 protein binding to the downstream sequence (as well as the overall junctions) and therefore resulting in significant amount of second-strand cleavage (see construct v (*1) in Figure 2, Chapter 3). Moreover, when the gap and a flap is added to the same backbone structure of construct v (*1) (see construct vi (*1) in Figure 2, Chapter 3), even higher amount of second-strand cleavage is seen.

For DNase I footprints, we used the "worst" and the "best" cleavers/substrates (i.e. construct ii (*1) and construct v (*1)) for second-strand cleavage (See Figure 2, Chapter 3) for footprint studies. DNase I footprint on construct v (*1) with non-specific sequence upstream of the insertion site showed protein is seen associating with the downstream sequence (presumably along with the junction as a whole). Unfortunately, footprint studies in Chapter 3 were significantly limited by the fact that branched integration intermediate structures exhibit large structure-derived footprints which occluded probable protein-DNA footprint signals.

Nonetheless, there are essentially few things that we could still learn: (1) Branched intermediate structures are good substrates for second-strand cleavage; (2) R2 protein sees the inherent structure of the branched intermediates, especially the template jump (West arm) structure and the single stranded East arm; (3) R2 protein can associate with the downstream sequence (i.e. North arm) of branched integration intermediates and that something beyond the 5 bp (i.e. specific sequence) in each arm is important for protein binding; (4) Sequence at the upstream of the insertion site is less important for protein binding, but the structure is still important; and (5) Presence of 5' PBM RNA can force the R2 protein to bind to downstream sequence as in the case of linear DNA, however, this reaction is not the correct pathway of the R2 insertion mechanism.

Limitations

The endonuclease of R2 is involved in both the first-strand/bottom strand cleavage and second-strand/top strand cleavage. Strangely, these cleavage sites for first-strand and second-strands are completely different. It is unclear how the same endonuclease manages to recognize, and perform cleavage in very specific manner. Nonetheless, instances of *in vitro* second-strand cleavage had always been a challenge since it required a narrow range of protein, DNA and RNA ratio. Most *in vitro* reactions with branched integration intermediates in Chapter 2 were limited due to inefficiency of second-strand cleavages. Part of the reason could be that most engineered branched structures were either not the right substrate or reaction conditions were incoherent. Second-strand cleavage is a prerequisite for second-strand synthesis. To better assist in our chances of getting second-strand synthesis, the substrates that showed highest amount of second-strand cleavage were selected for synthesis experiments as seen in Figure 4 of Chapter 2. Most of the

instances of second-strand synthesis seen were under conditions of protein excess and possibly because the synthesis occurred primarily from the released "primer template" (i.e. released from protein/DNA complex) generated by second-strand cleavage. Indeed, it was shown that the cleavage of branched integration intermediate generated a natural primer-template used for second-strand DNA synthesis.

While the branched integration intermediate structures are important for the second-strand cleavage, we still are uncertain as to which specific sequences are required for proper protein binding. In fact, we could not address how the protein binds to the "arms" of the branched structure. We assume that the R2 protein is binding at the central region of the branched structure but results are not conclusive. In addition, the bound vs. cleaved data (i.e. second-strand cleavage assay) suggests something beyond the 5 bp in the downstream sequence is important in branched structures; the DNase I footprint studies show similar findings. The footprint results showed some evidence of the R2 protein binding to the myb interaction region at the downstream (i.e. North arm) when the West arm was made of non-specific sequence, but not when the West arm was of specific sequence. Interpretation of the data was somewhat and limited because of the inherent structure-derived footprints seen on all branched integration intermediates.

Finally, we know that R2 protein binds both upstream and downstream sequences on linear DNA in the presence of PBM RNAs. In branched integration intermediates, specific sequences upstream of the insertion site lead to non-productive protein binding in the context of the *in vitro* reactions because the protein is not arriving at binding the branched DNA through a stepwise reaction. Nevertheless, protein binding does, apparently, see the structure of the upstream (West) arm. The protein also sees the North arm and appears to make productive specific contacts with

the North arm. Additional studies will be required to conclusively determine if the myb is being used to bind the North arm, as our results seem to indicate.

Future directions

Many uncertainties remain regarding protein binding with respect to branched integration intermediate structures. While branched structures were the key in mapping out full element insertion mechanism, various ideas regarding how the protein interacts with different regions including binding to specific sequences and protein-nucleic acid conformation required for second-strand cleavage remains to be addressed. The immediate follow-up experiments would focus on different versions of branched integration intermediates. The substrate with non-specific upstream plus gap and a flap still remains a clear winner for demonstrating second-strand cleavage. We plan on using this construct as a backbone structure to engineer few more constructs. For instance, the middle portion/sequences of the branched integration intermediate could be changed to non-specific sequences and address whether the protein requires the central region to bind to the structure. Also, myb region can be replaced with non-specific sequences to determine if the protein is still bound to the downstream sequence. Also, we can use other DNA footprinting techniques including methylation interference, hydroxyl radical footprints or ethylation interference to try to differentiate and resolve the structure footprint from the protein-DNA footprint. These methods could possibly address protein binding to branched structures from different perspective and resulting data would be in higher resolution. All of these possible routes could undoubtedly lead to better understanding of LINE integration. We plan on addressing most of the shortcomings on Chapter 3 to make it worthy for future publication.

References

1. Christensen,S.M. and Eickbush,T.H. (2005) R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**, 6617-6628.
2. Christensen,S.M., Bibillo,A. and Eickbush,T.H. (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468.
3. Christensen,S.M., Ye,J. and Eickbush,T.H. (2006) RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* **103**, 17602-17607.
4. Khadgi,B.B., Govindaraju,A. and Christensen,S.M. (2019) Completion of LINE integration involves an open '4-way' branched DNA intermediate. *Nucleic Acids Res* **47**, 8708-8719.
5. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605.
6. Govindaraju,A., Cortez,J.D., Reveal,B. and Christensen,S.M. (2016) Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287.
7. Ban,C. and Yang,W. (1998) Structural basis for MutH activation in E.coli mismatch repair and relationship of MutH to restriction endonucleases. *EMBO J* **17**, 1526-1534.
8. Tsutakawa,S.E., Jingami,H. and Morikawa,K. (1999) Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex *Cell* **99**, 615-623.
9. Hadden,J.M., Convery,M.A., Déclais,A.-C., Lilley,D.M.J. and Phillips,S.E.V. (2001) Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I *Nature structural biology* **8**, 62-67.
10. Nishino,T., Komori,K., Tsuchiya,D., Ishino,Y. and Morikawa,K. (2001) Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition. *Structure* **9**, 197-204.
11. Middleton,C.L., Parker,J.L., Richard,D.J., White,M.F. and Bond,C.S. (2004) Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res* **32**, 5442-5451.
12. Dias,A., Bouvier,D., Crépin,T., McCarthy,A.A., Hart,D.J., Baudin,F., Cusack,S. and Ruigrok,R.W.H. (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit *Nature* **458**, 914-918.
13. Feder,M. and Bujnicki,J.M. (2005) Identification of a new family of putative PD-(D/E) XK nucleases with unusual phylogenomic distribution and a new type of the active site *BMC genomics* **6**, 21.

14. Bibillo,A. and Eickbush,T.H. (2002) The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol* **316**, 459-473.
15. Bibillo,A. and Eickbush,T.H. (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* **279**, 14945-14953.