PROTEIN/NUCLEIC-ACID STRUCTURE AND SEQUENCE REQUIREMENTS FOR SPECIFYING SECOND-STRAND CLEAVAGE AND SECOND-STRAND SYNTHESIS DURING THE INTEGRATION OF THE SITE-SPECIFIC LINE R2Bm

By

ARUNA GOVINDARAJU

DISSERTATION

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy at
The University of Texas at Arlington
August, 2017

Arlington, Texas

Supervising Committee:

Dr. Shawn M. Christensen, Supervising Professor
Dr. Esther Betran
Dr. Michael R. Roner
Dr. Clay Clark
Dr. Kathryn O'Donnell-Mendell

Abstract

Long interspersed nucleotide elements (LINEs) are major drivers of genomic landscaping events in many eukaryotic genomes. LINEs encode a multifunctional protein that binds to their own mRNA, recognize and cleave target DNA, and performs target-primed reverse transcription (TPRT). R2 element from *Bombyx mori* (R2Bm) has served as an important experimental system for the biochemical understanding of the structure, function, and integration of LINEs. First-strand target site cleavage and TPRT in the current model of integration have been well understood both *in vitro* (using R2Bm) and *in vivo* (using LINE-1 elements); however, interpretation of factors involved in second-strand cleavage and second-strand cDNA synthesis has been experimentally difficult. A 3D modeling of R2Bm endonuclease was performed in conjunction with *in vitro* studies of point mutants. The 3D model revealed that the endonuclease of R2, and R2-like elements consisted of a variant of PD-(D/E)XK motif that serves as the core catalytic motif for many restriction-like endonuclease-fold (RLE) family of nucleases. Point mutations revealed the position and function of the catalytic K in that motif and other DNA binding residues. The model is largely built based on the structure of archaeal Holliday junction resolvases and few restriction endonucleases. These findings invoked the investigation of R2 activity on branched DNA structures and the possibility of DNA structure playing a role in second-strand cleavage and synthesis. A preferential binding of R2 protein to a non-target four-way Holliday junction over a non-target linear DNA was shown. In addition, an efficient second-strand cleavage in a previously unrecognized four-way structure that mimics template switching event was achieved. Upon second-strand cleavage, the same intermediate could provide a template-primer for second-strand synthesis. Based on the findings of this dissertation work, a new updated model of LINE integration was proposed that emphasizes on both sequence and structural aspects of integration intermediates.

ACKNOWLEDGEMENTS

I thank Dr. Shawn Christensen, my supervising professor for giving me this opportunity and believing in me. I could not have accomplished this dissertation work without his motivation, his dedicated mentorship and guidance. I am grateful for this wholesome experience and gratified that I could be part of this work. I thank Dr. Esther Betran for her constant support since the beginning of my graduate study. Her subject knowledge and her ability to catch nuances of my experiments have been of great help for my dissertation progress. I thank Dr. Mike Roner for his advice and questions that made me think out of the box. Thank you for your support, insightful comments and sharing your expertise. I thank Dr. Trey Fondon for encouraging me to think about my project with multiple dimensions. His advice has been very valuable to keep me grounded to fundamentals of Biology. His seminar course on evolution was an eye-opener and I thank him for making me read the 'Origin of species' and for other book recommendations. I thank Dr. Kate O'Donnell-Mendell for being an inspiration to me. She has shown her genuine interest and support for my project. Thank you for your insightful comments and encouragement. I thank Dr. Clay Clark for agreeing to be on my committee this year and for being an excellent chair for biology department.

I thank Dr. Kim Bowles for being a friend, a mentor and our best lab mom. She has been my source for intellectual and emotional support from day one. I am grateful to my lab family- Jeremy Cortez, Murshida Mahbub, Monika Pradhan, Eyad Shihabeddin, Brijesh Khadgi, Leine Newby-Estrella and undergraduates from the past and present. Each one them has contributed in multiple ways for the fulfillment of this dissertation. I thank you all for your friendship. I thank my good old friends- Neha Agrawal, Saikat Banerjee, Anu Aakash and Meena Balakrishnan for keeping my humor alive.

I thank Biology office staffs for their timely help and support throughout my stay in UT Arlington. I am also grateful to Rachel Wostl, my teaching lab coordinator for making my teaching hours worthwhile. I am grateful to National Science Foundation (NSF-MCB 0950983) funding to Dr. Christensen and Phi Sigma Biological Society, UTA chapter for the funding me for my dissertation.

TABLE OF CONTENTS

CHAPTER 1

Overview of Non-LTR Retrotransposons (Non-LTRs/LINEs):

Non-LTRs are a major class of retrotransposons. They are high in abundance and diversity. They have a wide range of host systems including fungi, plants, invertebrates, and mammals. Non-LTR retroelements are generally termed as LINE (**L**ong **IN**terspersed **E**lement), coined after its discovery in mammalian genome. A phylogenetic tree based on the reverse transcriptase domain showed that non-LTRs are earlier branching and rooted to group II introns from mitochondria and eubacteria (1). Retrotransposition by LINEs begin when an element encoded endonuclease cleaves one strand of host target DNA and creates a free 3'-OH end, which is then used as a primer by element encoded reverse transcriptase in a process called **t**arget **p**rimed **r**everse **t**ranscription (TPRT) to synthesize cDNA, followed by second strand synthesis (2). Thus, LINEs encode a multifunctional protein, that can bind nucleic acids, cleave target DNA, perform TPRT, and, it is hypothesized, perform second-strand synthesis to complete a new insertion. A subtype of LINEs called SINE (**S**hort **IN**terspersed **E**lement) is a non-autonomous element and depends on enzymatic activities of LINEs for replication (3,4).

Phylogenetically, LINEs are classified into at least 16 clades that include CRE, NeSL, R4, R2, Genie, L1, RTE, Tad1, R1, LOA, I, Ingi, Jockey, CR1, Rex1, and L2. While the phylogenetic tree and the resultant clades are based on the reverse transcriptase sequences, each clade also tends to represent LINEs that share additional structural features (5). The earlier-branching clades have single open reading frame (ORF) and encode a restriction-enzyme-like endonuclease (RLE) downstream of their reverse transcriptase (RT) (6) (Figure 1). For this reason, the early branching LINEs are sometimes called the RLE LINEs. The RLE LINEs include the Genie, CRE, R2, R4 and NeSL clades and tend to be site-specific in their integration. R2 is the most well studied of the five clades.

The later branching clades generally have two open reading frames. The second open reading frame encodes the RT and the DNA endonuclease (7). However, the DNA endonuclease is encoded upstream of the RT and is not of the RLE type, rather it is an apurinic apyrimidinic DNA endonuclease (APE). The APE LINEs tend to be non-site specific. APE LINEs include the L1, RTE, Jockey, I factor, Tras, etc. (Figure 2). Among the APE LINEs, L1 elements are best studied.

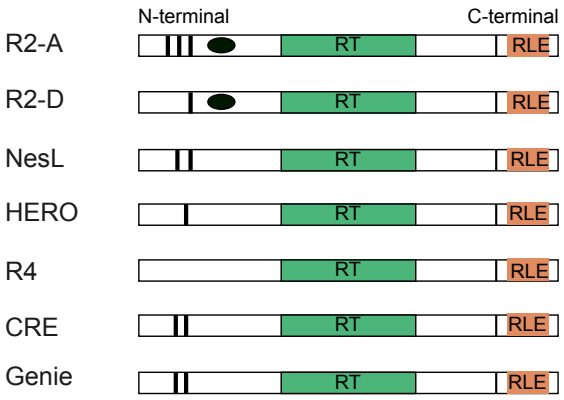


Figure 1: Domain structure of RLE LINE clades. Open boxes indicate ORFs, shaded boxes represent the enzymatic domains encoded in each element. All elements have a central reverse transcriptase (RT) domain, a C-terminal region with a CCHC domain (thin vertical lines) and a restriction like endonuclease domain (RLE). N-terminal region has variable number of CCHH zinc finger (ZF) motifs (thick vertical lines) and Myb-like (Myb) domains (black oval). This figure is adapted from Fujiwara, 2015 (8).

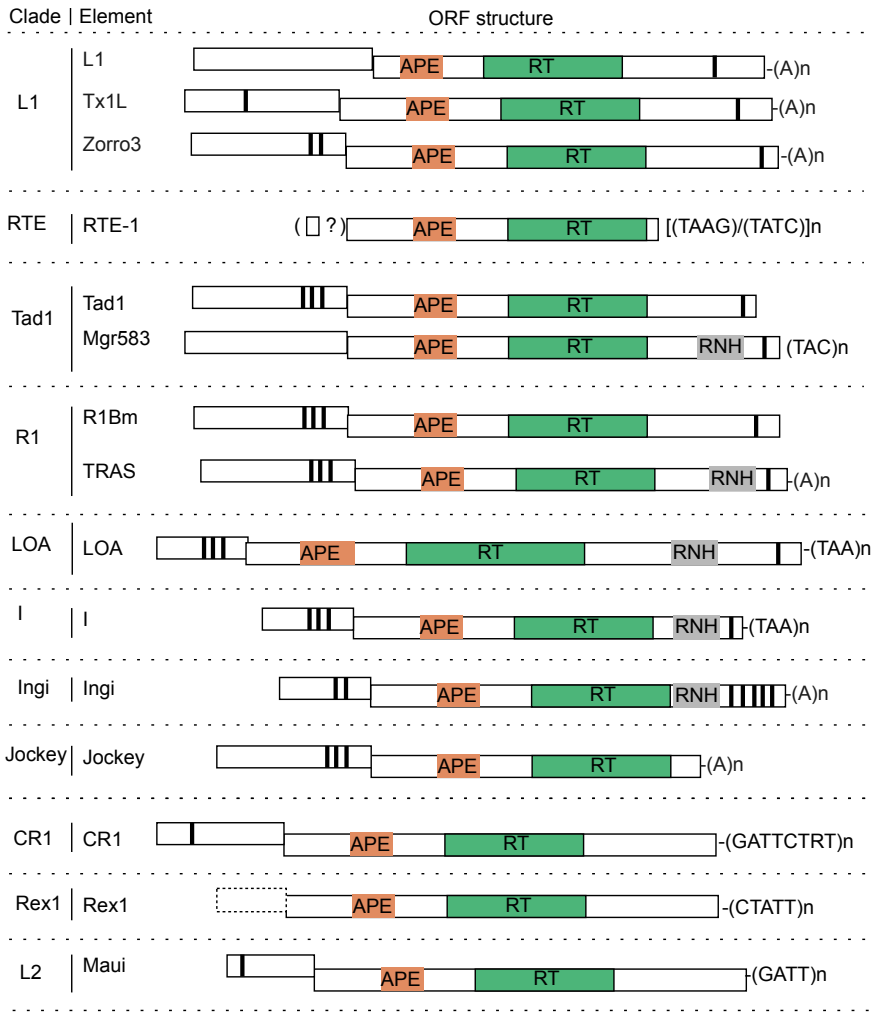


Figure 2: Domain structure of APE LINE clades. Representative elements from each clade are shown. Open boxes indicate ORFs, shaded boxes represent the enzymatic domains encoded in each element. The dotted box in Rex1 indicates that their 5' end has not been identified yet. The ORF1 of RTE is putative and shows no similarity to other elements (9). All elements have a reverse transcriptase (RT) domain and apurinic/apyrimidinic endonuclease (APE) in ORF2. Some elements have RNase H (RNH) in their ORF2 C-terminal region. Elements having 3' terminal repetitive sequences are shown at the C-terminal end of ORF2. Vertical lines represent cysteine-rich motifs. This figure is adapted from Zingler et al (10).

RLE LINEs:

RLE encoding LINEs usually target repetitive sequences such as rDNA and telomeric repeats. This is considered as a symbiotic strategy to reduce damage to host genome. Out of five clades of RLE LINEs, most members of NeSL, R2 and R4 and some members of HERO and CRE are site-specific. For example, CRE1/CRE2/SLACS/CZAR in trypanosomes from CRE clade and NeSL-1 in nematodes from NeSL clade target the spliced leader exons (11-13). R4 in Ascaris and Dong from R4 clade target a specific location in 28S rDNA and microsatellite TAA repeats, respectively (14,15).Genie 1 from Genie clade targets sub-telomeric repeats (16). R2 clade elements generally target a specific location(s) in the 28S rDNA, usually the 'R2-site' (17-19).

The R2 clade can be subdivided into four subclades—R2-A, R2-B, R2-C and R2-D—based on reverse transcriptase phylogeny (18). R2 constitutes a large group of retrotransposons with the target sequence specificity for 28S rDNA. It is present in six animal phyla: Arthropoda, Nematoda, Chordata, Echinodermata, Platyhelminthes, and Cnidaria (17,18,20-22) and have been recently reported in fishes, birds and reptiles (18,19,23). R2 belongs to R2-D clade and inserts at specific location at 28s rRNA gene (24,25). Most of our understanding of RLE LINEs comes from the R2 element from *Bombyx mori* (R2Bm) expressed and purified from *E.coli* (2,26). R8 element belonging to R2-A subclade targets 18S rDNA and is identified in *Hydra magnipapillata* in the Cnidaria phylum and it is to be noted that there is a lack of target similarity between R2 and R8 (20). R9, a member of R2-A subclade from rotifer targets 28S rDNA, but in a different location than that of R2 (27). Recent data from Christensen lab has suggested that the members of R2-A and R2-D subclades might differ in the specific roles of the two protein subunits thought to be involved in the insertion reaction (28,29).

Most of the biochemical information on the insertion mechanism of RLE LINEs comes from in vitro studies with the protein encoded by the R2 element from Bombyx mori (R2Bm). The single ORF of R2Bm, the ability to purify the R2Bm protein, the ability of that purified protein to bind to element RNA and target DNA specifically, make *in vitro* biochemical studies feasible. The protein nucleic acid interactions involved in forming the integration complex as well as a detailed study of the integration mechanism itself have been under investigation. These biochemical studies are also complemented by phylogenetic analysis and evolutionary studies (20).

R2 ORF structure and function:

The R2 element from R2Bm, a representative R2-D clade element, has single ORF that codes for a multifunctional protein that interacts with its own RNA, binds target DNA, cleaves target DNA, and performs TPRT. The ORF is composed of several domains including an a zinc finger (ZF), a MYB motif, an RT domain, and a carboxyl terminal region with a restriction-like endonuclease (RLE) domain and a cysteine/histidine rich motif (CCHC) (Figure 3).

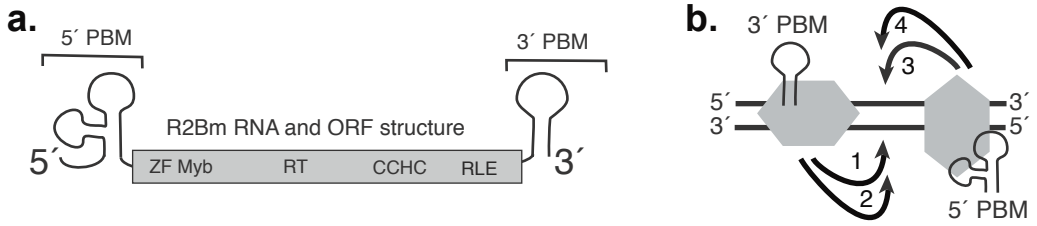


Figure 3: R2Bm structure. a) R2Bm RNA and open reading frame (ORF) structure. The ORF of R2Bm encodes several conserved motifs of known and unknown functions. Abbreviations: zinc finger (ZF), Myb (Myb), reverse transcriptase domain (RT), a cysteine-histidine rich motif (CCHC), and a PD-(D/E)XK type restriction-like endonuclease (RLE). RNA motifs present in the 5′ and 3′ untranslated regions that bind R2 protein are marked as 5′ and 3′ protein binding motifs (PBMs), respectively. Brackets indicate the individual segments of the R2Bm RNA (i.e., the 5′ and 3′ PBM RNAs) that were used for our experiments. b) The R2 insertion mechanism is depicted: (1) DNA cleavage of the bottom strand, (2) TPRT, (3) DNA cleavage of the top strand, and (4) second strand DNA synthesis. (This figure is adapted with permission from Govindaraju et al. 2016 (30)).

The ZF and Myb. The ZF and Myb of the R2Bm protein have been implicated as the protein domains responsible for recognizing the target DNA sequences downstream of the insertion site (31). In a missing nucleoside and N7-methylated guanosine interference footprints analyses, R2Bm protein was shown to bind downstream of the insertion site in the presence of 5’ PBM RNA. The DNA in the major

groove near insertion site around position +13 (downstream of insertion site) and from 1 to 3 bases upstream of the cleavage site (32). A polypeptide containing only the ZF and Myb was shown to footprint to the same target sequences. Mutating the ZF of the polypeptide revealed that the region from -3 to -1 was contacted by the ZF motif, while the region between +10 and +16 is contacted the Myb domain (33).

The R2Bm ZF is comparable to Cys-His ZF motif of eukaryotic transcription factor, TFIIIa that exhibits DNA binding ability (34). In TFIIIa ZF motif forms an α-helix starting from the second cysteine to the first histidine residues of the motif. The amino acids at position -1, 3 and 6 of the helix contact three consecutive bases in the DNA, thereby providing sequence-specificity of the binding (34). These three residue positions in the R2 protein are highly conserved across the R2D clade members (31). R2-A, R2-C, and R2-D clades have three, two, and one N-terminal zinc-finger motif(s), respectively. R2-B clade retrotransposons have three or two zinc-finger motifs at the N-terminus (18).

The R2 Myb motif is comparable to the c-Myb (Myb) domain. c-Myb proteins are known to bind DNA through two 50-residue domains, each composed of three helixes with the third helix involved in primary DNA contacts (35). R2 protein has conserved regions that correspond to the first and third helixes of a c-Myb domain. There are some clades of RLE LINEs that have no Myb motif, but in R2 clade Myb is important for target specificity. R2-A and R2-D clade Myb domains target the R2 site differently. R2-A clade Myb is likely used to target a protein subunit upstream of the insertion site while the R2-D clade Myb is used to target a subunit downstream of the insertion site. Myb-motif of R2Bm that belongs to R2-D clade binds DNA sequences from 10 to 15 base pair downstream of the insertion site (32). The difference in binding among R2-clade elements may suggest a plasticity of integration mechanisms (29).

The reverse transcriptase. R2 specifically recognizes the 3' PBM RNA. The 3' PBM RNA is necessary and sufficient to support TPRT (36). Priming of TPRT, however, occurs without complementarity between the primer and template (37). R2 RT has higher processivity than viral RTs (38). Another unusual property of R2 RT is that it can jump from the 5′ end of one template to the 3′ end of another template. This type of template jumping, or template switching does not require sequence identity between the templates. R2 RT can add up to five non-templated nucleotides to the cDNA when it reaches the end of a template

and micro homologies between these added nucleotides and sequences near the 3′ end of the acceptor template enable the polymerase to jump between templates (37,39). Similar end-to-end template jumps are also observed for viral RNA directed RNA polymerases (40), Mauriceville retroplasmid (41) and group II intron RTs (42). The intrinsic strand displacement property of the R2Bm RT is speculated to displace the RNA from the RNA-cDNA hybrid (formed during TPRT) as second-strand synthesis is being templated from the cDNA (38). Though, it is possible that LINEs also depend on host RNAse H for integration. For example, the RT of group II introns also have strong strand displacement property but they do depend on host RNase H for mobility (43). There are two conserved sequence motifs that have been termed 0 and −1, immediately N-terminal to motifs 1 through 7 of the RT domain. Mutations in these -1 and 0 domains result in decreased 5' PBM RNA binding, 3' PBM RNA binding, and polymerase functions (44). Domain 0 shares similar sequence and position in all lineages of LINEs and as well in group II introns and telomerase (45-47).

The CCHC motif. The C-terminal region has a conserved three-helix-bundle motif (ending in the sequence HINALP) and a conserved zinc-knuckle-like cysteine histidine motif (CCHC) that could play a role in nucleic acid binding (33,48). It is notable that the spacing of cysteine and histidine residues did not match with that of characterized canonical zinc knuckles (34). Canonical zinc knuckles are most often involved in nucleic acid binding (32). This indicates that C-terminal of R2 might involve in binding target DNA or might involve in RNA binding. A similar CCHC motif is found in ORF2 of L1 elements as well and will be discussed further in the APE LINEs section.

The RLE. R2 RLE is shown to have conserved residues- Lys/Arg-Pro-Asp-$X_{12\text{-}13aa}$-Asp (abbreviated as PD-(D/E)). Mutating the two D residues abolished the cleavage activity, but not binding or the subsequent TPRT process (6). The PD-(D/E) motif and thus the endonuclease itself appear to be a member of larger endonuclease family, the PD-(D/E)XK endonucleases. The first of PD-(D/E)XK endonucleases to be identified were the type II restriction enzymes (e.g. EcoRI, BamHI, and FokI). They have active sites composed of similar conserved residues, and mutations in these residues abolished the catalytic cleavage activity but not the binding. Now, the PD-(D/E)XK superfamily comprises a large group of diverse proteins that are involved in various functions including, but not limited to DNA repair (49,50), Holliday junction resolution (51-53), and RNA processing (54,55). The PD-(D/E)XK superfamily proteins

share little to no sequence homology, however they all share a structurally conserved core. The consensus core consists of a four-stranded mixed β-sheet flanked by α-helix on each side (αββββαβ) (56).

R2 Transcription and RNA processing:

Transcription of non-LTR elements are controlled either by an internal promoter, as in the case of human L1 element (57), or by the upstream cellular promoter that transcribes their RNA as a co-transcript, as in the case of site specific elements like R2 (58). The 28S-R2 co-transcript was shown to perform self-cleavage at the 5' end. The conserved region at 5' end of R2 RNA could be folded into double pseudoknot structure like that of HDV (hepatitis delta virus) ribozyme (59) and less conserved regions can form variable peripheral domain as commonly seen in HDV-like ribozymes (60,61). HDV ribozymes are autocatalytic RNAs that are capable of folding into intricate double-pseudoknot structure composed of five helical regions joined by single stranded linkers (62). They catalyze a *trans*-esterification that leads to the cleavage of the RNA sugar–phosphate backbone leaving 5′-OH and 2′,3′-cyclic phosphate ends (63). HDV-like ribozymes are also found in some of the non-site-specific APE encoding LINEs as well (64,65). The activity of the R2 ribozyme from Drosophila differed from that of other diverse animals in the location of RNA self-cleavage. For instance, *D. simulans* R2 ribozyme cleaves at the precise 28S/R2 5′ junction whereas other R2 ribozymes cleaved in GC-rich regions of the 28S rRNA either 13 or 28 nucleotides upstream of the R2 insertion site. The 5' end sequences depending on whether 28S sequence is present after self-cleavage, can result in different integration products *in vivo* (36,66). On the other hand, presence or absence of downstream 28S sequence at 3' end of RNA did not affect the TPRT integrated products *in vitro* (67,68). The co-transcribed 28S sequence of 5' end in the first strand cDNA is postulated to play a role in priming second strand synthesis by annealing to target site (69). Thus, a ribozyme self-cleavage that includes the upstream 28S sequences in some species gives rise to insertions with uniform 5′ junctions while self-cleavage at the exact 5′ end in other species gives rise to variable junctions (31,69,70) as RT might add some non-templated nucleotides enabling micro homology based second strand initiation (39). When *Drosophila* is injected with R2 RNA with and without upstream 28S sequences, integration showed precise and highly variable 5' ends respectively (36). Variations include deletions or truncations ranging from 100 bp to 3 kb and addition of non-templated sequence.

The 3' ends of R2 transcripts are conserved except for the variable length of A-rich regions and repeats. On the other hand, 5' UTR region is truncated and it is attributed to RNA degradation during TPRT, reverse transcriptase falling off from template before reaching the 5' end and template switching (39) as observed in R2 retrotransposon. Template switching will be discussed in detail in another section of this chapter.

R2 RNA Translation:

Although R2 protein is well characterized, translation of R2 transcript to yield R2 protein remains to be illustrated. Multiple start codon AUG for R2 transcript can be found (71) and variable N-terminal domain of R2 ORF may result in variation in start codon occurrence. Because of the R2 RNA being co-transcribed with ribosomal RNA and the ribozyme activity, it is unlikely that R2 transcripts undergo canonical translation initiation like that of capped mRNAs. It is possible that RNA structure in the R2 5' UTR forms an internal ribosome entry sites (IRESes) seen in viruses and some cellular mRNAs that can interact with translational machinery to initiate protein synthesis (59,71,72). R2 ribozyme is shown to act like an IRES both *in vitro* and *in vivo* presumably initiating translation by binding to translation machinery. This mode of translation can compensate the absence of 5 -methylguanosine cap and a conserved AUG codon on the R2 RNA *(64)*.

R2 Ribonucleoprotein complex (RNP) Formation:

R2 protein binds to 3′ and 5' untranslated region (UTR) of the R2 mRNA. The 5' UTR that constitutes the protein binding motif (PBM) is about 300 nt starting from the beginning of 5′ UTR and ends just before the sequences encoding the N-terminal zinc-finger. This 5' PBM forms a distinctive pseudoknot structure that is conserved across silk moths (72,73). The 250 bp of 3' UTR forms the 3' PBM.

The two protein subunits are thought to be tethered together by binding to specific secondary structures (PBMs) of its mRNA (48). R2 protein depending on which segment of R2 RNA it is bound to

define the subunit's function in the integration reaction (33,74). In the presence of the 3′ PBM, the R2 protein binds the 28S gene upstream of the insertion site. This subunit is responsible for first strand DNA cleavage and TPRT. The subunit has high affinity for the 3' PBM as even 1000-fold excess competitor RNA did not inhibit R2 protein from binding to its own RNA and proceeding with TPRT. Interestingly, *B. mori* R2 protein can recognize the 3′ UTR of the *D. melanogaster* R2 element, even though the *D. melanogaster* R2 RNA contained only little primary nucleotide sequence identity to the 3′ UTR of the R2 element from *B. mori* (67). This finding suggests that, like many RNA-protein interactions, protein recognition is mediated by the secondary and tertiary structures of the R2 RNA. The secondary structures for the two RNAs share three helical regions and the sequence AAC/UAUC in the loop generated by one of these helixes. These conserved regions are considered to make specific interactions with R2 protein (75). In the presence of the 5′ PBM the R2 protein binds the 28S gene down- stream of the insertion site (48). The downstream subunit is responsible for second-strand DNA cleavage.

First strand cleavage and TPRT:

In the presence of 3' PBM RNA the R2 protein binds upstream of the insertion site in the region from -40 to -20 bp from the site of DNA cleavage (33). It is unclear what part of the R2 protein is responsible for binding to the target DNA. Candidate regions include a conserved region just upstream of the CCHC motif, the CCHC motif, and the RLE. It is hard to conceptualize, however, how the RLE could make extensive contacts so far from the site of DNA cleavage. First strand cleavage occurs at the bottom strand i.e. anti-sense strand of target DNA with respect to 28S rRNA gene promoter. The R2Bm protein uses the 3' PBM RNA as template and nicked target DNA as primer for TPRT. While any non-specific RNA will position a R2 protein subunit upstream of the insertion site, only protein bound to the 3' PBM RNA will engage in TPRT (2). The priming of TPRT often includes unsuccessful attempts before successful reverse transcription, resulting in addition of untemplated nucleotides or short repeats at the 3' end of integrated elements (67,76). For accurate initiation of TPRT, annealing of RNA template to the DNA target is not a prerequisite. It is not known how or if 3' end of R2 is processed or must be processed away from the rRNA co-transcript. The presence of downstream 28s rRNA sequences after the 3'UTR in the *in vitro* reactions produced the accurate R2-target 3' junction as found in endogenous R2 of many insects. This explains the

utilization of co-transcript for the proper integration owing to the absence of any known mechanism to process the 3' end of R2 RNA (68).

Second strand cleavage and second strand synthesis:

The downstream subunit cleaves the second DNA strand after the reverse transcription removes 5' PBM bound from this subunit. The downstream subunit's RT is then hypothesized to perform second-strand synthesis using the newly cleaved DNA strand as a primer, completing the TPRT reaction. Thus, formation of correct ribonucleoprotein particle (RNP) is essential for successful TPRT and second strand synthesis.

Second strand cleavage and synthesis have been difficult to study and to understand. There are many unresolved questions and issues that need to be studied in greater detail. Second strand cleavage currently requires a narrow window of protein:DNA:RNA ratios. The sequence space of the first and second strand cleavage sites differs. It is unclear how a 'site-specific' DNA endonuclease cleaves two different sites specifically. It is unclear if second strand cleavage requires a protein dimer. It is also unclear if second strand cleavage requires the first DNA strand to be cleaved.

The R2 endonuclease has also been shown to cleave ssDNA at its junction with duplex DNA. This activity was proposed to be related to second strand cleavage as second strand cleavage is believed to be non-site specific and some local denaturation after first strand cleavage may result in ssDNA-duplex DNA junction (38). It is possible that this side reaction is, in part, responsible for generating 5' end truncations during insertion. The endogenous 5' end truncations may also be the result of template jumping. Template jumping may also be an important port of initiating second strand synthesis. If upstream 28S sequences are present on the element RNA and a template jump may be initiated after the RT has reverse transcribed a short distance of it. If template jump occurs before the RT reaches the end of the R2 template, a 5' truncation may occur (76).

Transcription and regulation:

R2 activity can cause large deletions that could be detrimental to host and insertions causing reduced 28S rRNA synthesis can affect host development (77-79). Concerted evolution at rDNA locus acts as a homogenizing force in the locus, driving out TEs. TEs in the rRNA locus must actively transpose or be eliminated from the locus. Regulation measures at the rRNA locus level appear to be primarily for control of TEs not for control of the rRNA (80).

The control of R2 activity is executed at the level of transcription of rDNA locus. All rDNA loci are not active all the time. An estimated 30-40 units out of several hundred units are transcriptionally active in *D. melanogaster* at a given time (77). R2 transcripts are controlled at the level of transcription rather than at a post-transcription level. The control of transcription is dependent on the composition of rDNA itself. The chromatin where R2 gets inserted can be repressed by epigenetic marks linked to heterochromatin formation (81). But, the transcription domain model of the rDNA locus explains the mode of altering the transcriptional state. The model explains the host preference to transcribe big blocks of un-inserted rDNA units or the one with fewer R2 insertions. If the selected unit has no R2 inserts then there are no R2 transcripts and no R2 retrotransposition. However, crossovers, the very characteristic of rDNA locus to bring about concerted evolution is also responsible for changing the rDNA block that are active. Crossovers could change the boundary of active blocks of rDNA, thus eliminating or including R2 insertions (82). Additionally, differences in various organisms' ability to identify R2 insertion free locus can explain the occurrence of different percentages of R2 insertions in different species (17,83,84).

The stability of R2 in rDNA locus is remarkable. Among many factors, the choice of target site and the nature of target site regulation take precedence. In a computer simulation where R2 insertions each can inactivate a rDNA locus along with the input of variable crossover rate, R2 retrotransposition rate and number of un-inserted units needed for host's viability, one can still generate stable populations (85). Thus, the key to long-term survival of R2 in rDNA loci is its ability to reside in target DNA outside of transcriptionally active domains for many generations. The stochastic nature of crossovers in rDNA locus holds the key to repopulate and maintain few copies without any harm to the host (80).

APE LINEs:

Most of the APE-encoding LINEs do not have target-specificity in their insertion. But some of them might exhibit weak target site preference. For example, human LINE-1 elements have specificity for TAAA repeats (86,87). Exception to this trend are the members of Tx1 and R1 clades of APE LINEs exhibiting target-specificity. Tx1 includes many elements such as Mutsu, targeting rRNA genes, Dewa targeting tRNA genes, Tx1-1_ACar targeting telomeric repeats, etc (22,88). R1 clade includes elements such as R1 (targeting rRNA genes), TRAS (targeting telomeric repeats) and Waldo (targeting microsatellites) (89).

APE LINEs can be divided into four groups and 11 clades based on their RT phylogeny and domain structure (1,5,31,90). Much of the following discussion about APE-type LINEs will be based on an autonomously active human LINE called LINE-1 or L1. It is notable that LINE-1 and their corresponding SINEs constitute at least one-third and LINE-1 derived sequences represent 17%of human genomic DNA (91).

ORF structure, function and variations across APE LINEs:

APE-encoding LINEs typically have two ORFs. Human L1 element measures ~6 kb in length. Human LINE-1 5′ UTR is ~910 bp in length and contains an internal RNA polymerase II promoter and cis-acting binding sites for few transcription factors necessary for LINE-1 transcription (57). Most LINE-1 mRNAs contain a 7-methyl guanosine cap structure that facilitates their translation (92). Analysis of a subset of LINE-1 (L1Ta) revealed that only ~30-35% of insertions are full length and ~40-45% are truncated at their 5’ ends and ~25% show inversion/deletion events (91,93,94). Following 5’ UTR are two open reading frames, ORF1 and ORF2 and a 3′ UTR that terminates in a poly-A tract (95). This poly-A tract is not required for L1 retrotransposition in cultured cells (96), but it can inhibit L1 RT activity in *in vitro* reactions (97). L1 3′ UTRs also contain a functional RNA polymerase II polyadenylation signal (98).

ORF1. The first ORF protein (ORF1p) from human LINE-1 is an ~40 kDa protein, translated from LINE-1 mRNA by a traditional cap-dependent mechanism (92). Structural studies have revealed that the

N-terminus of ORF1p contains a coiled-coil domain facilitating their trimerization (99). ORF1 of telomere-specific APE LINEs such as HeT-A, TART, and TAHRE encodes a Gag-like protein and transports it back into the nucleus (100). C-terminal of ORF1 proteins in many clades of APE-encoding LINEs contains one to three cysteine-rich motifs (CCHC motifs). Similar CCHC motifs called zinc knuckle motifs in Gag proteins of retroviruses are involved in interaction between retroviral RNA and Gag proteins (101,102). In APE encoding SART1 LINE element, mutational analyses of three CCHC motifs in ORF1p showed that they are involved in the interaction with its mRNA in a sequence-specific manner and are essential for SART1 retrotransposition *in vivo* (103). This motif is conserved in both site-specific Tx1 and R1 clades as well (104). C-terminal domain of ORF1p along with the central region also contains a non-canonical recognition motif and is shown to bind single-strand RNA in a sequence-independent manner and resides in cytoplasmic RNPs (105-108). ORF1p is thus believed to be involved in RNP assembly (109) and as a chaperone to protect cis-acting RNA (99,110). Human, mouse and non-mammalian ORF1p contain nucleic acid chaperone activity that promotes re-annealing of single strand DNA *in vitro* (110,111). While the functions of ORF1p has been established, deleting ORF1 from ZfL2-1 element did not abolish retrotransposition activity in cultured human cells (112). Also, only LINE-1 ORF2p is required for Alu retrotransposition (113). These findings question the extent to which ORF1p is involved in mobility.

ORF2. The second ORF protein (ORF2) is an ~150 kDa protein encoding an APE type endonuclease (EN) (7) in the N-terminal domain, a reverse transcriptase (RT) and a C-terminal domain (114). Except for L1Tc from *Trypanosoma cruzi*, APE domain has lost its ability to cleave apurinic/apyrimidic (AP) sites (115). While L1 EN is not site specific, it may have preference for target DNA structural features (116). L1 EN makes a single-strand nick at 5′-TTTT/A-3′ sequence (slash indicates scissile phosphate), exposing a 5′ phosphate and 3′ hydroxyl group (7). RT domain shares sequence similarity to that of telomerase, Penelope-like retrotransposons, group II introns, other non-LTR retrotransposons, LTR retrotransposons, and retroviruses (5,117). L1 RT has both RNA-dependent and DNA-dependent polymerase activities (118). L1 RT has high processivity, lacks RNase H activity and exhibits cis-preference for its own RNA similar to R2Bm RT (97). Additionally, it can extend terminally mismatched primer–

templates (119). Between L1 EN and RT domains there is a PCNA interaction protein domain (PIP box). PCNA is an essential protein for DNA replication. Mutating PIP box abolished L1 retrotransposition (120).

The C-terminal of ORF2 of most APE-type retrotransposons also encode at least one CCHC motif (96) with similar spacing as that of C-terminal R2, but different from that of the CCHC motifs in ORF1 (103,121). This motif in ORF2 is also conserved in both Tx1 and R1 clades. It is suggested to function like a zinc-knuckle domain (122). Point mutations in this motif in human L1 and Bombyx TRAS1 result in loss of the retrotransposition activity (96,123). Further, mutations in L1 ORF2 CCHC motif reduce the reverse transcriptase activity of the ribonucleoprotein (RNP), implying a possible role in binding to L1 RNA (124). Recently, the last 180 amino acids of ORF2p that contains this CCHC motif is shown to bind RNA in a non-sequence specific manner, however cysteine to serine mutations do not affect this RNA binding (125). Though, it is speculated that this CCHC motif is involved in binding nucleic acids and/or protein motifs, the exact function has not been elucidated yet.

In TRAS1, a telomere-specific element, a Myb-like domain is found between endonuclease and RT domains in ORF2 (126). A Myb-like three-helix motif is shared by many telomere-binding proteins (127), implying a possible role of TRAS1 Myb-like motif in binding the telomere.

The proteins encoded by APE LINE-1s can act in trans to mobilize non-autonomous retrotransposons (e.g., human Alu and SVA elements and mouse B1 and B2 elements) (113,128,129) and cellular mRNAs to new genomic locations, which give rise to pseudogenes (130). These non-autonomous elements contain necessary ORF structures that are needed for being mobilized by the LINE-1 proteins. (131). Human Alu element is one of the best studied SINEs. Alus are derived from the 7SL RNA component of the signal recognition particle, a ribonucleoprotein that targets a subset of nascent polypeptides to the endoplasmic reticulum (132). A full-length Alu element forms a dimeric structure that consists of highly similar left and right monomers separated by an adenosine-rich linker sequence. The left Alu monomer contains the sequences that are required for RNA polymerase III dependent transcription (133). The right monomer ends in a poly (A) tract and contains an RNA polymerase III terminator sequence (134).

Transcription, RNA structure and function:

While R2 HDV-like ribozymes are used to process R2 transcripts, L1 ribozyme that maps downstream of the retrotransposon 5' terminus is not involved in processing L1 transcripts (135). Exception to this is L1Tc, a LINE element from *Trypanosoma cruzi*. The first 77 nt of this element forms an internal promoter that generates translatable transcripts (136). In the same region, an active ribozyme is shown to exist making this element to be the first described LINE with internal promoter–ribozyme dual function (137). However, it is shown that L1 ribozyme was lost from a linage of L1 called L1PA suggesting that L1 ribozyme is not beneficial for retrotransposition (64).

Translation:

An internal ribosome entry site located near the 3' end of mouse ORF1 is shown to facilitate ORF2 translation (138). In human counterparts, there are two in-frame stop codons between ORF1 and ORF2. Genetic studies show that these stop codons are involved in terminating ORF1 translation and re-initiating ORF2 translation. ORF2 is shown to be translated in an AUG-independent manner. Since it is unlikely of human ORF1 to have an internal ribosome entry site, it is possible that host translation mechanism is used (139). SART elements undergo translational coupling as seen in prokaryotes and viruses (18). SART ribozyme is also capable of initiating translation in a non-canonical fashion (140).

Ribonucleoprotein complex (RNP) Formation:

After transcription, a full-length bicistronic LINE-1 mRNA is exported to the cytoplasm. Upon translation, ORF1p and ORF2p with a strong cis-preference, bind to their respective mRNA, forming an RNP (141). SART1 ORF1p includes domains for the ORF1p–ORF1p and ORF1p–ORF2p interactions (103). L1 RNP constitutes L1 mRNA, multiple ORF1p trimers, at least one ORF2p and likely contains other cellular proteins and RNAs (124,141,142). While ORF1p and RT activity are readily detectable in RNP preparations, ORF2p has been difficult to identify. Immunofluorescence studies revealed that ORF2p co-localizes with both ORF1p and LINE-1 mRNA in cytoplasmic foci, which are closely associated with stress

granule proteins (124). In yeast, cytoplasmic foci localization, termed P-body localization of proteins is important for RNP assembly (143,144). However, the stoichiometry of ORF1p and ORF2p bound to LINE-1 mRNA require further elucidation.

DNA targeting:

In APE encoding LINEs, DNA target specificity can be dictated by APE itself, in addition to other factors such as target DNA accessibility in nuclei, interaction between mRNA and target DNA. Like the protein-DNA interactions of DNase I, DNA recognition by L1 endonuclease is also likely mediated by minor groove interactions at least on a local scale (116). Among the APE LINEs that exhibit non-specific targeting, human L1 inserts randomly in host genome with a preference for 5′-TTTT/AA-3' sequence. A similar sequence preference is shown for their *in vitro* retrotransposition and additionally any mutation in the endonuclease domain is shown to abolish DNA cleavage and retrotransposition (7). Retrotransposons like I and *jockey* integrate into AT-rich regions of the genome (145,146). Substitutions at target DNA which result in pyrimidine- purine junctions are cleaved, whereas purine-pyrimidine junctions greatly reduce or eliminate cleavage activity by APE endonuclease.

Among the APE LINEs that exhibit site-specific targeting, R1Bm and Tx1L APEs have been shown *in vitro* to cleave the 28S rDNA and Tx1D, respectively (147,148). Purified APE of telomere-specific TRAS1 element cleaves the telomeric repeats between T and A on the $(TTAGG)_n$ bottom strand and between C and T on the $(CCTAA)_n$ top strand in a highly specific manner and is not affected by the flanking sequence, implying that the target-site specificity of TRAS1 is mainly determined by APE domain itself (149).

Crystal structures of some of the APE domains have four-layered α/β sandwich structure with a topology like those of AP endonucleases, however, the structures have an extra β-hairpin loop (βB_6-βB_5 in the case of L1 and β_{10}-β_{11} in TRAS1) at the edge of the DNA-binding surface that fits into the wider minor groove of TpA junction of the target DNA. Mutagenesis in this loop (Tyr-98 and N-180 for R1Bm EN and Asp-130 for TRAS1) is shown to affect the sequence specificity. In general, modulation in DNA specificity can be achieved by varying the individual surface loops of APE structure (150-152). Swapping the APE domains of SART1 and TRAS1 in a retrotransposition assay showed that their sequence for cleavage

specificity was interchanged, suggesting that the primary determinant of *in vivo* target selection is the APE domain (123).

First strand cleavage and TPRT:

Most of the mechanistic studies of APE LINEs are performed using a retrotransposition cell culture assay with LINE-1. The 3′ UTR sequences of LINE-1s are tagged with a retrotransposition indicator cassette that consists of a backward copy of a neomycin phosphotransferase reporter gene equipped with its own promoter and polyadenylation signals. There is an intron in the same transcriptional orientation as the LINE-1 disrupting the reporter gene. This ensures that the expression of the neomycin phosphotransferase gene only occurs upon a successful LINE-1 retrotransposition, allowing the host cell to grow in the presence of the neomycin analog G418 (96).

L1 TPRT process is analogous to that of R2Bm. Additionally, a modification of TPRT process called twin-priming is proposed, where second-strand cleavage occurs before the completion of TPRT, producing an internal primer. The internal primer then anneals to L1 RNA, priming reverse transcription. When the RNA is removed from the RNA/cDNA structure, single-stranded cDNAs pair at a region of microhomology, and the remaining DNA synthesis is completed. The entire process results in an inversion of small regions of L1 flanked by target site duplications (93).

Second strand cleavage and second strand synthesis:

Although the first nick occurs at L1 EN consensus cleavage site (5′-TTTT/A), the analysis of flanking canonical TSDs did not call for sequence-specific second-strand cleavage site except for a weak preference for the sequence 5′-TYTN/R observed in few inversion/deletion and inversion/duplication events in L1s. Following possibilities are outlined for second-strand cleavage: (i) L1 APE exhibits site-specific bottom-strand cleavage activity but has a weak or no sequence specificity for second-strand cleavage, (ii) L1 encodes a second nuclease activity that is required for top-strand cleavage, or (iii) host factors are involved (153).

The top-strand cleavage occurs at variable distances downstream of the bottom-strand cleavage site, preferably within 15-16 bp (131). A comparison study of top-strand cleavages occurring upstream and downstream of first-strand nick and the one right on top of first-strand nick indicated that only downstream cleavages resulted in TSDs as observed in human genome reference sequence (86,153). The subsequent steps of the integration process of elements may involve 5′-truncation and 5′-invertion. The 5' truncation is explained by L1 RT unable to copy the entire L1 RNA, falling off of the template before completion or RNA template being degraded (154). The 5'-inversion is explained by twin-priming, in which second-strand synthesis is primed by annealing of the newly synthesized cDNA to complementary sequences in the genomic target (93). It is possible that different mechanisms are employed for attachment of the 5′-end of L1 to the top-strand and initiate second-strand synthesis (155). As observed for R2, short stretches of non-templated nucleotides can be added after TPRT by the LINE-1 RT, which by micro complementarity may facilitate annealing of the LINE-1 cDNA to single- strand DNA exposed at the top-strand target site DNA (156).

Annealing of the LINE-1 cDNA to top-strand genomic DNA may specify top-strand cleavage, and priming for subsequent second-strand synthesis. Additional steps in second-strand cleavage and second-strand cDNA synthesis require detailed investigation.

Transcription and regulation:

Methylation of CpG sequences, especially 5-hydroxymethylation of cytidine regulate LINE-1 transcription and is involved in suppression of LINE-1 expression in a variety of cell types (157,158). Human APOBEC3 gene family proteins catalyze the deamination of cytidine to uridine residues in single-strand DNA substrates (159). In cell culture, inhibition of LINE-1 and Alu retrotransposition is caused by several members of the APOBEC3 family, especially APOBEC3A (A3A) and APOBEC3B (A3B) (160-162). Several RNA-based restriction mechanisms have been employed in LINE-1 control. PIWI (P-element induced wimpy testes) protein is a subclade of the Argonaute family of small RNA binding proteins. They interact with small RNAs (26-31 nt) termed piwi-interacting RNAs (piRNAs) (163). piRNAs are known to actively defend the mammalian germline from transposons (164). L1 transcripts can be processed to small interfering RNAs (siRNAs) that can suppress retrotransposition through an RNA interference (RNAi)

mechanism (165). Recently, it is shown that members of the Krüppel- associated box (KRAB) zinc-finger (KZNF) protein family can recruit KRAB-associated protein-1 (KAP1) and its associated repressive complex to inhibit L1 and SVA elements to control their expression in embryonic stem cells (166).

Full-length human L1 RNAs contain a conserved splice donor sequence, which is shown to involve in the generation of shorter LINE-1 mRNAs that can affect retrotransposition (167). LINE-1 mRNA splicing is considered to serve as a regulatory mechanism to restrict LINE-1 expression and/or retrotransposition in a tissue-specific manner (168). Recently, a RNA binding protein that facilitates alternative splicing is shown to bind mouse L1 RNA and human L1 RNP indicating a direct role in modulating retrotransposition (169,170).

Gap in the knowledge and scope of this dissertation work:

First half of the integration mechanism, first-strand DNA cleavage and TPRT, in both RLE and APE LINEs have been fairly extensively characterized, although there is much not known about DNA target site recognition. The second half of the integration reaction, second-strand cleavage and second-strand synthesis, is much less understood, especially the biochemical and mechanistic details. The existing model for RLE LINE insertion proposes a mode of integration where a pseudo dimer (two protein-subunits bound to the same RNA) interact on either side of the insertion site on the target DNA. First-strand DNA cleavage and TPRT performed by one unit of protein and second -strand cleavage and second-strand synthesis is performed by another unit. The weak top strand cleavage that we observed on target duplex DNA could be due to the lack of right intermediate structure in our *in vitro* reactions.

My dissertation work focuses on DNA recognition and cleavage requirements, at both the protein and DNA levels, for both first strand and especially second-strand DNA cleavage steps. In chapter 2, I addressed the unknown aspects of R2 endonuclease domain, as DNA cleavage is the crucial and initial step for both first and second half of the reactions. We have modeled the 3D structure of R2Bm endonuclease domain and outlined the biochemical experiments to identify the catalytic K residue in the PD-(D/E)XK catalytic motif. We identified the presence of R-box (RXXR) located at the beginning of PD-(D/E)XK catalytic motif and showed that this R-box has strong affinity to bind DNA, which corroborated its

role in a subset of type II restriction enzymes. In addition, the findings from chapter 2 opens the possibility of R2 activity on branched structures, owing to the close relation to Holliday junction resolvase PD-(D/E)XK catalytic motif as revealed by structural database during 3D modeling.

In Chapter 3 I explored R2 protein's activity on Holliday junctions and other branched DNA structures. I identified a hitherto unrecognized DNA intermediate required for efficient second strand DNA cleavage: one that, when cleaved appears to be a substrate for second strand synthesis. Finally, I proposed a new model of integration based on the combined information from chapter 2 and chapter 3 that can be relevant to both RLE and APE type LINEs.

REFERENCES:

1. Eickbush, T.H. and Malik, H.S. (2002), *Mobile DNA ii*. American Society of Microbiology, pp. 1111-1144.
2. 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.
3. Kajikawa, M. and Okada, N. (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell*, **111**, 433-444.
4. Roy-Engel, A.M. (2012) A tale of an A-tail: The lifeline of a SINE. *Mobile genetic elements*, **2**, 282-286.
5. 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.
6. 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.
7. Feng, Q., Moran, J.V., Kazazian, H.H., Jr. and Boeke, J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905-916.
8. Fujiwara, H. (2015) Site-specific non-LTR retrotransposons. *Microbiol Spectr*, **3**, MDNA3-0001-2014.
9. Malik, H.S. and Eickbush, T.H. (1998) The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol Biol Evol*, **15**, 1123-1134.
10. 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.
11. Malik, H.S. and Eickbush, T.H. (2000) NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. *Genetics*, **154**, 193-203.
12. Aksoy, S., Williams, S., Chang, S. and Richards, F.F. (1990) SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. *Nucleic acids research*, **18**, 785-792.

13. Gabriel, A., Yen, T.J., Schwartz, D.C., Smith, C.L., Boeke, J.D., Sollner-Webb, B. and Cleveland, D.W. (1990) A rapidly rearranging retrotransposon within the miniexon gene locus of Crithidia fasciculata. *Molecular and cellular biology*, **10**, 615-624.
14. Burke, W.D., Muller, F. and Eickbush, T.H. (1995) R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res*, **23**, 4628-4634.
15. Xiong, Y. and Eickbush, T.H. (1993) Dong, a non-long terminal repeat (non-LTR) retrotransposable element from Bombyx mori. *Nucleic Acids Res*, **21**, 1318.
16. 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. *Molecular biology and evolution*, **19**, 619-630.
17. 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.
18. Kojima, K.K. and Fujiwara, H. (2005) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol*, **22**, 2157-2165.
19. 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.
20. 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.
21. Kapitonov, V. and Jurka, J. (2014) A family of HERO non-LTR retrotransposons from the Californian leech genome. *Repbase Reports*, **14**, 311.
22. 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. *Molecular biology and evolution*, **21**, 207-217.
23. Kapitonov, V. and Jurka, J. (2009) R2 non-LTR retrotransposons in the bird genome. *Repbase Rep*, **9**, 1329.
24. Burke, W.D., Calalang, C.C. and Eickbush, T.H. (1987) The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol*, **7**, 2221-2230.
25. Jakubczak, J.L., Xiong, Y. and Eickbush, T.H. (1990) Type I (R1) and type II (R2) ribosomal DNA insertions of Drosophila melanogaster are retrotransposable elements closely related to those of Bombyx mori. *J Mol Biol*, **212**, 37-52.
26. Xiong, Y.E. and Eickbush, T.H. (1988) Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. *Cell*, **55**, 235-246.
27. 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.
28. Shivram, H., Cawley, D. and Christensen, S.M. (2011) Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements*, **1**, 169-178.
29. Thompson, B.K. and Christensen, S.M. (2011) Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: Plasticity of integration mechanism. *Mob Genet Elements*, **1**, 29-37.

30. 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 research*.
31. 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. *Molecular biology and evolution*, **16**, 502-511.
32. 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.
33. 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.
34. Berg, J.M. and Shi, Y. (1996) The galvanization of biology: a growing appreciation for the roles of zinc. *Science*, **271**, 1081-1085.
35. Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. and Nishimura, Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, **79**, 639-648.
36. 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. *Molecular and cellular biology*, **20**, 213-223.
37. 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.
38. 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.
39. 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.
40. Arnold, J.J. and Cameron, C.E. (1999) Poliovirus RNA-dependent RNA polymerase (3Dpol) is sufficient for template switching in vitro. *The Journal of biological chemistry*, **274**, 2706-2716.
41. Chen, B. and Lambowitz, A.M. (1997) De novo and DNA primer-mediated initiation of cDNA synthesis by the mauriceville retroplasmid reverse transcriptase involve recognition of a 3' CCA sequence. *Journal of molecular biology*, **271**, 311-332.
42. Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V.R., Hunicke-Smith, S., Swamy, S. *et al.* (2013) Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *Rna*, **19**, 958-970.
43. Yao, J., Truong, D.M. and Lambowitz, A.M. (2013) Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. *PLoS Genet*, **9**, e1003469.
44. Jamburuthugoda, V.K. and Eickbush, T.H. (2014) Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic acids research*, **42**, 8405-8415.

45. Clements, A.P. and Singer, M.F. (1998) The human LINE-1 reverse transcriptase: effect of deletions outside the common reverse transcriptase domain. *Nucleic acids research*, **26**, 3528-3535.
46. Gu, S.Q., Cui, X., Mou, S., Mohr, S., Yao, J. and Lambowitz, A.M. (2010) Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA*, **16**, 732-747.
47. Rouda, S. and Skordalakes, E. (2007) Structure of the RNA-binding domain of telomerase: implications for RNA recognition and binding. *Structure*, **15**, 1403-1412.
48. 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.
49. 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.
50. 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.
51. Hadden, J.M., Convery, M.A., Declais, A.C., Lilley, D.M. and Phillips, S.E. (2001) Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I. *Nat Struct Biol*, **8**, 62-67.
52. 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.
53. 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.
54. Dias, A., Bouvier, D., Crepin, T., McCarthy, A.A., Hart, D.J., Baudin, F., Cusack, S. and Ruigrok, R.W. (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature*, **458**, 914-918.
55. Yuan, P., Bartlam, M., Lou, Z., Chen, S., Zhou, J., He, X., Lv, Z., Ge, R., Li, X., Deng, T. *et al.* (2009) Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature*, **458**, 909-913.
56. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. and Ginalski, K. (2012) Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res*, **40**, 7016-7045.
57. Swergold, G.D. (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and cellular biology*, **10**, 6718-6729.
58. Eickbush, T.H. (1992) Transposing without ends: the non-LTR retrotransposable elements. *New Biol*, **4**, 430-440.
59. Eickbush, D.G. and Eickbush, T.H. (2010) R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA cotranscript. *Mol Cell Biol*, **30**, 3142-3150.
60. Jimenez, R.M., Polanco, J.A. and Luptak, A. (2015) Chemistry and Biology of Self-Cleaving Ribozymes. *Trends Biochem Sci*, **40**, 648-661.
61. Been, M.D., Perrotta, A.T. and Rosenstein, S.P. (1992) Secondary structure of the self-cleaving RNA of hepatitis delta virus: applications to catalytic RNA design. *Biochemistry*, **31**, 11843-11852.

62. Chen, J.-H., Yajima, R., Chadalavada, D.M., Chase, E., Bevilacqua, P.C. and Golden, B.L. (2010) A 1.9 Å crystal structure of the HDV ribozyme precleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry*, **49**, 6508-6518.
63. Cochrane, J.C. and Strobel, S.A. (2008) Catalytic strategies of self-cleaving ribozymes. *Acc Chem Res*, **41**, 1027-1035.
64. Ruminski, D.J., Webb, C.H., Riccitelli, N.J. and Luptak, 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.
65. Webb, C.H., Riccitelli, N.J., Ruminski, D.J. and Luptak, A. (2009) Widespread occurrence of self-cleaving ribozymes. *Science*, **326**, 953.
66. 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 research*, **32**, 1555-1565.
67. 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.
68. 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.
69. 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.
70. 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 biology*, **10**, R49.
71. George, J.A. and Eickbush, T.H. (1999) Conserved features at the 5 end of Drosophila R2 retrotransposable elements: implications for transcription and translation. *Insect Mol Biol*, **8**, 3-10.
72. 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.
73. Kierzek, E., Kierzek, R., Moss, W.N., Christensen, S.M., Eickbush, T.H. and Turner, D.H. (2008) Isoenergetic penta- and hexanucleotide microarray probing and chemical mapping provide a secondary structure model for an RNA element orchestrating R2 retrotransposon protein function. *Nucleic acids research*, **36**, 1770-1782.
74. Yang, J. and Eickbush, T.H. (1998) RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol Cell Biol*, **18**, 3455-3465.
75. 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.
76. Eickbush, T.H. (2002), *Mobile DNA II*. American Society of Microbiology, pp. 813-835.
77. Hawley, R.S. and Marcus, C.H. (1989) Recombinational controls of rDNA redundancy in Drosophila. *Annu Rev Genet*, **23**, 87-120.

78. Zhang, X., Zhou, J. and Eickbush, T.H. (2008) Rapid R2 retrotransposition leads to the loss of previously inserted copies via large deletions of the rDNA locus. *Mol Biol Evol*, **25**, 229-237.
79. Hollocher, H. and Templeton, A.R. (1994) The molecular through ecological genetics of abnormal abdomen in Drosophila mercatorum. VI. The non-neutrality of the Y chromosome rDNA polymorphism. *Genetics*, **136**, 1373-1384.
80. Eickbush, T.H. and Eickbush, D.G. (2015) Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr*, **3**, MDNA3-0011-2014.
81. Ye, J. and Eickbush, T.H. (2006) Chromatin structure and transcription of the R1- and R2-inserted rRNA genes of Drosophila melanogaster. *Molecular and cellular biology*, **26**, 8781-8790.
82. Zhou, J., Eickbush, M.T. and Eickbush, T.H. (2013) A population genetic model for the maintenance of R2 retrotransposons in rRNA gene loci. *PLoS genetics*, **9**, e1003179.
83. Lathe, W.C., 3rd 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.
84. Ghesini, S., Luchetti, A., Marini, M. and Mantovani, B. (2011) The non-LTR retrotransposon R2 in termites (Insecta, Isoptera): characterization and dynamics. *J Mol Evol*, **72**, 296-305.
85. Zhang, X., Eickbush, M.T. and Eickbush, T.H. (2008) Role of recombination in the long-term retention of transposable elements in rRNA gene loci. *Genetics*, **180**, 1617-1626.
86. Gilbert, N., Lutz-Prigge, S. and Moran, J.V. (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell*, **110**, 315-325.
87. Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G. and Boeke, J.D. (2002) Human l1 retrotransposition is associated with genetic instability in vivo. *Cell*, **110**, 327-338.
88. Kojima, K. and Jurka, J. (2013) Non-LTR retrotransposons from green anole. *Repbase Rep*, **4**, 1615.
89. Kojima, K.K. and Fujiwara, H. (2003) Evolution of target specificity in R1 clade non-LTR retrotransposons. *Molecular biology and evolution*, **20**, 351-361.
90. Lovsin, N., Gubensek, F. and Kordi, D. (2001) Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Mol Biol Evol*, **18**, 2213-2224.
91. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
92. 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. *Molecular and cellular biology*, **27**, 4685-4697.
93. Ostertag, E.M. and Kazazian, H.H., Jr. (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome research*, **11**, 2059-2065.
94. Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V. *et al.* (2002) A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet*, **71**, 312-326.

95. Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F. and Kazazian, H.H., Jr. (1991) Isolation of an active human transposable element. *Science*, **254**, 1805-1808.
96. Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D. and Kazazian, H.H., Jr. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917-927.
97. Piskareva, O. and Schmatchenko, V. (2006) DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. *FEBS letters*, **580**, 661-668.
98. Moran, J.V., DeBerardinis, R.J. and Kazazian, H.H., Jr. (1999) Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530-1534.
99. Khazina, E., Truffault, V., Buttner, R., Schmidt, S., Coles, M. and Weichenrieder, O. (2011) Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat Struct Mol Biol*, **18**, 1006-1014.
100. Rashkova, S., Karam, S.E. and Pardue, M.L. (2002) Element-specific localization of Drosophila retrotransposon Gag proteins occurs in both nucleus and cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3621-3626.
101. Williams, M.C., Gorelick, R.J. and Musier-Forsyth, K. (2002) Specific zinc-finger architecture required for HIV-1 nucleocapsid protein's nucleic acid chaperone function. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 8614-8619.
102. D'Souza, V. and Summers, M.F. (2004) Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. *Nature*, **431**, 586-590.
103. 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.
104. 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.
105. Hohjoh, H. and Singer, M.F. (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J*, **15**, 630-639.
106. Kolosha, V.O. and Martin, S.L. (1997) In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A*, **94**, 10155-10160.
107. Khazina, E. and Weichenrieder, O. (2009) Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc Natl Acad Sci U S A*, **106**, 731-736.
108. Januszyk, K., Li, P.W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J.A., Martin, S.L. and Clubb, R.T. (2007) Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. *J Biol Chem*, **282**, 24893-24904.
109. Kulpa, D.A. and Moran, J.V. (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet*, **14**, 3237-3248.
110. Martin, S.L. and Bushman, F.D. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Molecular and cellular biology*, **21**, 467-475.

111. Nakamura, M., Okada, N. and Kajikawa, M. (2012) Self-interaction, nucleic acid binding, and nucleic acid chaperone activities are unexpectedly retained in the unique ORF1p of zebrafish LINE. *Molecular and cellular biology*, **32**, 458-469.
112. Kajikawa, M., Sugano, T., Sakurai, R. and Okada, N. (2012) Low dependency of retrotransposition on the ORF1 protein of the zebrafish LINE, ZfL2-1. *Gene*, **499**, 41-47.
113. Dewannieux, M., Esnault, C. and Heidmann, T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics*, **35**, 41-48.
114. Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D. and Gabriel, A. (1991) Reverse transcriptase encoded by a human transposable element. *Science*, **254**, 1808-1810.
115. Olivares, M., Thomas, M.C., Alonso, C. and Lopez, M.C. (1999) The L1Tc, long interspersed nucleotide element from Trypanosoma cruzi, encodes a protein with 3'-phosphatase and 3'-phosphodiesterase enzymatic activities. *The Journal of biological chemistry*, **274**, 23883-23886.
116. Cost, G.J. and Boeke, J.D. (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37**, 18081-18093.
117. Evgen'ev, M.B. and Arkhipova, I.R. (2005) Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenetic and genome research*, **110**, 510-521.
118. Piskareva, O., Denmukhametova, S. and Schmatchenko, V. (2003) Functional reverse transcriptase encoded by the human LINE-1 from baculovirus-infected insect cells. *Protein Expr Purif*, **28**, 125-130.
119. Monot, C., Kuciak, M., Viollet, S., Mir, A.A., Gabus, C., Darlix, J.L. and Cristofari, G. (2013) The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS genetics*, **9**, e1003499.
120. Taylor, M.S., LaCava, J., Mita, P., Molloy, K.R., Huang, C.R., Li, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T. *et al.* (2013) Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell*, **155**, 1034-1048.
121. Kajikawa, M., Ohshima, K. and Okada, N. (1997) Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. *Mol Biol Evol*, **14**, 1206-1217.
122. Fanning, T. and Singer, M. (1987) The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic Acids Res*, **15**, 2251-2260.
123. Takahashi, H. and Fujiwara, H. (2002) Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J*, **21**, 408-417.
124. Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A., Moran, J.V. *et al.* (2010) Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet*, **6**.
125. Piskareva, O., Ernst, C., Higgins, N. and Schmatchenko, V. (2013) The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio*, **3**, 433-437.
126. Kubo, Y., Okazaki, S., Anzai, T. and Fujiwara, H. (2001) Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. *Mol Biol Evol*, **18**, 848-857.

127. Konig, P., Fairall, L. and Rhodes, D. (1998) Sequence-specific DNA recognition by the myb-like domain of the human telomere binding protein TRF1: a model for the protein-DNA complex. *Nucleic acids research*, **26**, 1731-1740.
128. Dewannieux, M. and Heidmann, T. (2005) L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *Journal of molecular biology*, **349**, 241-247.
129. Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Lower, J., Stratling, W.H., Lower, R. and Schumann, G.G. (2012) The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic acids research*, **40**, 1666-1683.
130. Esnault, C., Maestre, J. and Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature genetics*, **24**, 363-367.
131. Jurka, J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A*, **94**, 1872-1877.
132. Ullu, E. and Tschudi, C. (1984) Alu sequences are processed 7SL RNA genes. *Nature*, **312**, 171-172.
133. Chu, W.M., Liu, W.M. and Schmid, C.W. (1995) RNA polymerase III promoter and terminator elements affect Alu RNA expression. *Nucleic acids research*, **23**, 1750-1757.
134. Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nature reviews. Genetics*, **3**, 370-379.
135. Salehi-Ashtiani, K., Luptak, A., Litovchick, A. and Szostak, J.W. (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*, **313**, 1788-1792.
136. 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.
137. 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 research*, **39**, 8065-8077.
138. 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 research*, **34**, 853-864.
139. 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.
140. Lupták, A. and Szostak, J.W. (2007), *Ribozymes and RNA catalysis*, pp. 123-133.
141. Hohjoh, H. and Singer, M.F. (1997) Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. *Journal of molecular biology*, **271**, 7-12.
142. Martin, S.L. (1991) Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. *Mol Cell Biol*, **11**, 4804-4807.
143. Dutko, J.A., Kenny, A.E., Gamache, E.R. and Curcio, M.J. (2010) 5' to 3' mRNA decay factors colocalize with Ty1 gag and human APOBEC3G and promote Ty1 retrotransposition. *J Virol*, **84**, 5052-5066.

144. Larsen, L.S., Beliakova-Bethell, N., Bilanchone, V., Zhang, M., Lamsa, A., Dasilva, R., Hatfield, G.W., Nagashima, K. and Sandmeyer, S. (2008) Ty3 nucleocapsid controls localization of particle assembly. *J Virol*, **82**, 2501-2514.
145. Priimagi, A.F., Mizrokhi, L.J. and Ilyin, Y.V. (1988) The Drosophila mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. *Gene*, **70**, 253-262.
146. Chaboissier, M.C., Finnegan, D. and Bucheton, A. (2000) Retrotransposition of the I factor, a non-long terminal repeat retrotransposon of Drosophila, generates tandem repeats at the 3' end. *Nucleic Acids Res*, **28**, 2467-2472.
147. 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.
148. Christensen, S., Pont-Kingdon, G. and Carroll, D. (2000) Target specificity of the endonuclease from the Xenopus laevis non-long terminal repeat retrotransposon, Tx1L. *Molecular and cellular biology*, **20**, 1219-1226.
149. Anzai, T., Takahashi, H. and Fujiwara, H. (2001) Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol Cell Biol*, **21**, 100-108.
150. Weichenrieder, O., Repanas, K. and Perrakis, A. (2004) Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure*, **12**, 975-986.
151. Maita, N., Anzai, T., Aoyagi, H., Mizuno, H. and Fujiwara, H. (2004) Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J Biol Chem*, **279**, 41067-41076.
152. 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.
153. 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.
154. Ostertag, E.M. and Kazazian, H.H., Jr. (2001) Biology of mammalian L1 retrotransposons. *Annual review of genetics*, **35**, 501-538.
155. 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 research*, **15**, 780-789.
156. 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**, MDNA3-0061-2014.
157. Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, **13**, 335-340.
158. Branco, M.R., Ficz, G. and Reik, W. (2011) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*, **13**, 7-13.
159. Chiu, Y.L. and Greene, W.C. (2008) The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol*, **26**, 317-353.
160. Bogerd, H.P., Wiegand, H.L., Hulme, A.E., Garcia-Perez, J.L., O'Shea, K.S., Moran, J.V. and Cullen, B.R. (2006) Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc Natl Acad Sci U S A*, **103**, 8780-8785.

161. Chen, H., Lilley, C.E., Yu, Q., Lee, D.V., Chou, J., Narvaiza, I., Landau, N.R. and Weitzman, M.D. (2006) APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Current biology : CB*, **16**, 480-485.
162. Wissing, S., Montano, M., Garcia-Perez, J.L., Moran, J.V. and Greene, W.C. (2011) Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. *J Biol Chem*, **286**, 36427-36437.
163. Siomi, M.C., Sato, K., Pezic, D. and Aravin, A.A. (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews. Molecular cell biology*, **12**, 246-258.
164. Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T. and Hannon, G.J. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell*, **31**, 785-799.
165. Yang, N. and Kazazian, H.H. (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature structural & molecular biology*, **13**, 763-771.
166. Jacobs, F.M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R. and Haussler, D. (2014) An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, **516**, 242-245.
167. Belancio, V.P., Hedges, D.J. and Deininger, P. (2006) LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic acids research*, **34**, 1512-1521.
168. Belancio, V.P., Roy-Engel, A.M. and Deininger, P. (2008) The impact of multiple splice sites in human L1 elements. *Gene*, **411**, 38-45.
169. Peddigari, S., Li, P.W., Rabe, J.L. and Martin, S.L. (2013) hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. *Nucleic acids research*, **41**, 575-585.
170. Goodier, J.L., Cheung, L.E. and Kazazian, H.H., Jr. (2013) Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res*, **41**, 7401-7419.

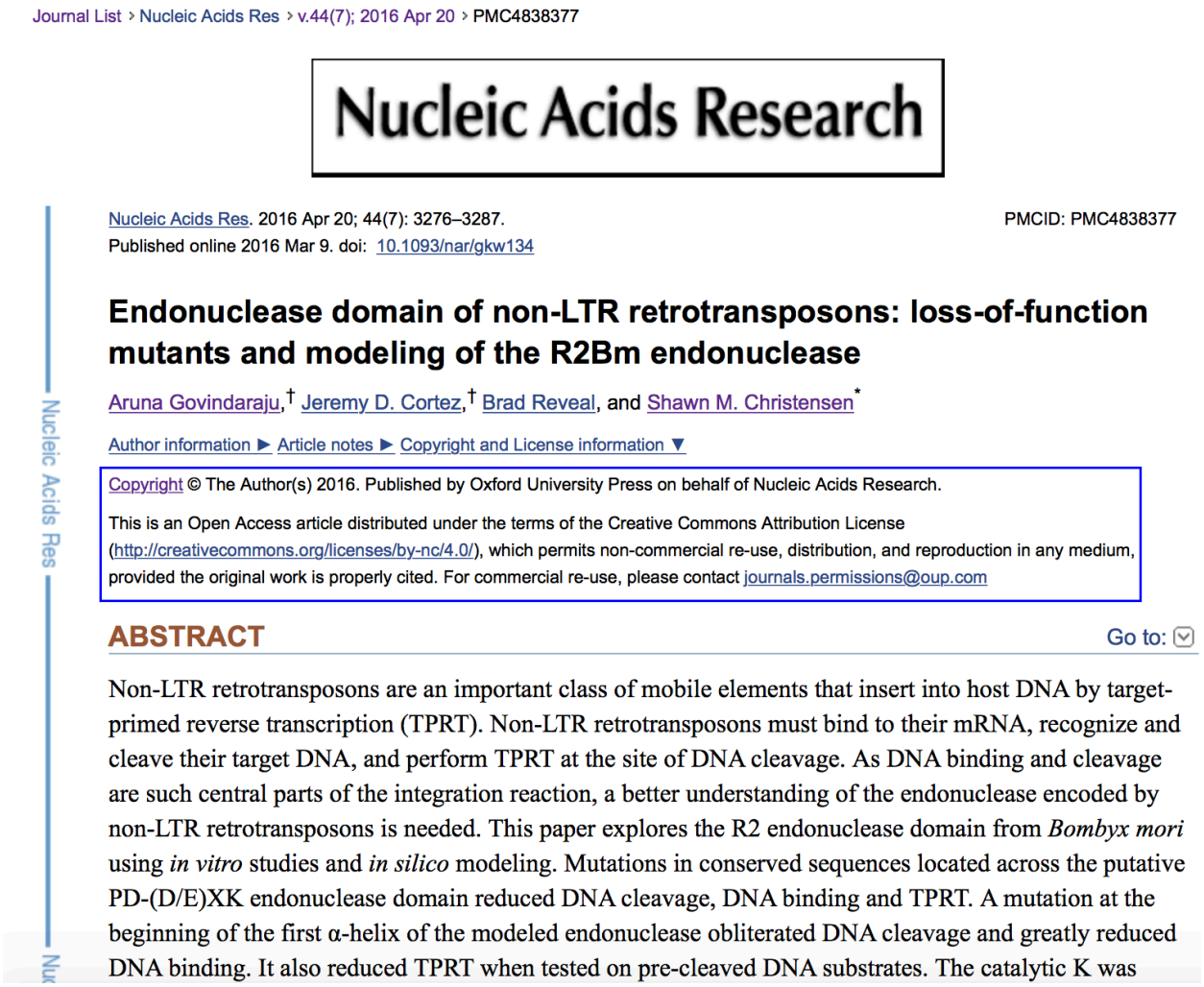
CHAPTER 2

ENDONUCLEASE DOMAIN OF NON-LTR RETROTRANSPOSONS: LOSS-OF-FUNCTION MUTANTS AND MODELING OF THE R2BM ENDONUCLEASE[1]

Govindaraju A, Cortez JD, Reveal B, Christensen SM. Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. Nucleic Acids Res 2016;44:3276-87.

[1] Manuscript presented here is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4838377/. Used with permission from authors, 2017

Copyright information for chapter 2 (outlined by a blue rectangle):



Journal List › Nucleic Acids Res › v.44(7); 2016 Apr 20 › PMC4838377

Nucleic Acids Research

Nucleic Acids Res. 2016 Apr 20; 44(7): 3276–3287. PMCID: PMC4838377
Published online 2016 Mar 9. doi: 10.1093/nar/gkw134

Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease

Aruna Govindaraju,† Jeremy D. Cortez,† Brad Reveal, and Shawn M. Christensen*

ABSTRACT

Go to:

Non-LTR retrotransposons are an important class of mobile elements that insert into host DNA by target-primed reverse transcription (TPRT). Non-LTR retrotransposons must bind to their mRNA, recognize and cleave their target DNA, and perform TPRT at the site of DNA cleavage. As DNA binding and cleavage are such central parts of the integration reaction, a better understanding of the endonuclease encoded by non-LTR retrotransposons is needed. This paper explores the R2 endonuclease domain from *Bombyx mori* using *in vitro* studies and *in silico* modeling. Mutations in conserved sequences located across the putative PD-(D/E)XK endonuclease domain reduced DNA cleavage, DNA binding and TPRT. A mutation at the beginning of the first α-helix of the modeled endonuclease obliterated DNA cleavage and greatly reduced DNA binding. It also reduced TPRT when tested on pre-cleaved DNA substrates. The catalytic K was

All authors of this paper have expressed consent for reprinting this article in my dissertation. Individual author roles are as follows:

Shawn M. Christensen: Principle investigator and responsible for conception of idea for paper

Brad Reveal: Made few endonuclease mutants and performed initial screening of mutants

Jeremy D. Cortez: Made few endonuclease mutants, performed cleavage experiments, analyzed data and involved in writing paper

Aruna Govindaraju: Performed binding experiments for wild type and mutants, 3D modeling and alignment work, determined fraction TPRT for endonuclease mutants, analyzed data and involved in writing paper

Abstract

Non-LTR retrotransposons are an important class of mobile elements that insert into host DNA by target-primed reverse transcription (TPRT). Non-LTR retrotransposons must bind to their mRNA, recognize and cleave their target DNA, and perform TPRT at the site of DNA cleavage. As DNA binding and cleavage are such central parts of the integration reaction, a better understanding of the endonuclease encoded by non-LTR retrotransposons is needed. This paper explores the R2 endonuclease domain from *Bombyx mori* using *in vitro* studies and *in silico* modeling. Mutations in conserved sequences located across the putative PD-(D/E)XK endonuclease domain reduced DNA cleavage, DNA binding, and TPRT. A mutation at the beginning of the first α-helix of the modeled endonuclease obliterated DNA cleavage and greatly reduced DNA binding. It also reduced TPRT when tested on precleaved DNA substrates. The catalytic K was located to a noncanonical position within the second α-helix. A mutation located after the fourth β-strand reduced DNA binding and cleavage. The motifs that showed impaired activity form an extensive basic region. The R2 biochemical and structural data are compared and contrasted with that of two other well characterized PD-(D/E)XK endonucleases, restriction endonucleases and archaeal Holliday junction resolvases.

Introduction

R2 is a widely distributed site-specific non-LTR retrotransposon that targets the host genome's 28S ribosomal RNA genes (1, 2). R2 elements, like all non-LTR retrotransposons, replicate by inserting into the host's chromosomes using a process called target-primed reverse transcription (TPRT) (3-5). The element encoded endonuclease cleaves the host chromosome to generate a free 3' OH. The element encoded reverse transcriptase uses the liberated 3' OH to prime reverse transcription of the element RNA into DNA at the site of insertion (3-6). The protein encoded by the R2 element from *Bombyx mori,* R2Bm, is expressible in bacteria, and purified components are readily obtainable—allowing TPRT to be extensively studied *in vitro*. Two R2Bm protein subunits have been shown to be involved in the TPRT reaction: one bound to the 5' end of the element RNA and one bound to the 3' end of the RNA (Figure 1B). The protein subunit bound to the element RNA's 3' protein binding motif (PBM) interacts with target DNA upstream of

the insertion site, cleaves the first (bottom) DNA strand, and performs TPRT (5). The protein subunit bound to the 5' PBM RNA, binds to target DNA downstream of the insertion site, cleaves the second (top) DNA strand, and is thought to perform second strand synthesis (5, 7, 8). The upstream and downstream protein subunits bind and cleave different (i.e., non palindromic) DNA sequences.

R2 elements are divided into subclades based upon sequence homology and number of amino-terminal zinc fingers coded for by the open reading frame (ORF). The two major subclades are the R2-A and R2-D clades. The R2-A clade elements have three amino-terminal zinc fingers, while the R2-D clade elements have a single zinc finger. R2 (and R2-like) elements have been found to target various genomic repeats in their host genomes (9-15). The R2 site is located near the middle of the 28S rDNA, the R9 site is near the beginning of the 28S rDNA, and the R8 site is in the 18S rDNA (16-18). The ORF of R2Bm, an R2-D clade element, encodes a number of conserved motifs of known and unknown functions (Figure 1A).

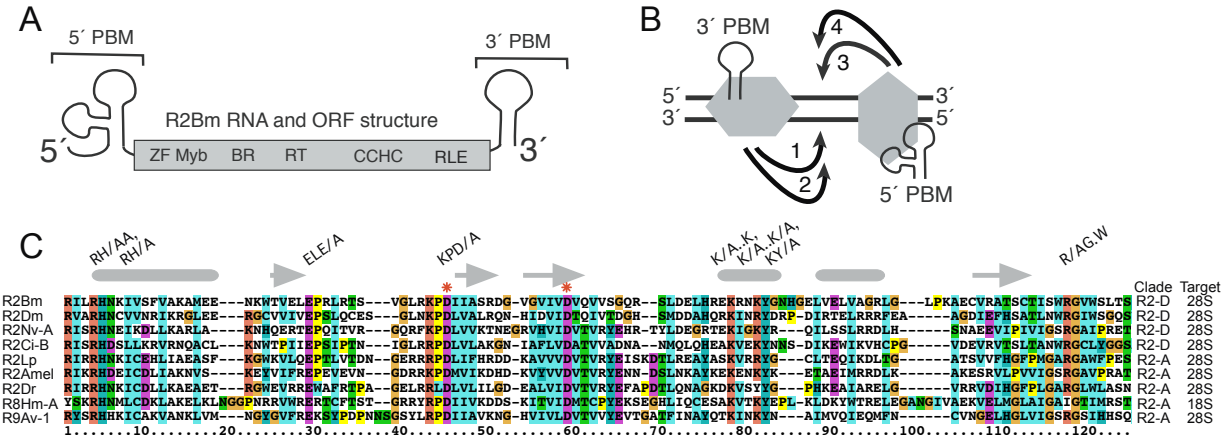


Figure 1: R2Bm structure. (A) R2Bm RNA and open reading frame (ORF) structure. The ORF of R2Bm encodes a number of conserved motifs of known and unknown function. Abbreviations: zinc finger (ZF), Myb (Myb), basic region (BR), reverse transcriptase domain (RT), a cysteine-histidine rich motif (CCHC), and a PD-(D/E)XK type restriction-like endonuclease (RLE). RNA motifs present in the 5′ and 3′ untranslated regions that bind R2 protein are marked as 5′ and 3′ protein binding motifs (PBMs), respectively. Brackets indicate the individual segments of the R2Bm RNA (i.e., the 5′ and 3′ PBM RNAs) that were used in this paper. The data for reactions containing the 3′ PBM are presented in the paper. The 5' PBM RNA data is located in the supplemental material. (B) The R2 insertion mechanism is depicted: (1) DNA cleavage of the bottom strand, (2) TPRT, (3) DNA cleavage of the top strand, and (4) second strand DNA synthesis. (C) Clustal alignment of the RLE domain of D-clade and A-clade R2 elements that target the R2, R8, and R9 sites. Abbreviations: *Bombyx mori* R2 (R2Bm) (52), *Drosophila melanogaster* R2 (R2Dm) (16), *Nasonia vitripennis* R2 (R2Nv-A) (53), *Ciona intestinalis* R2 (R2Ci-B) (54), *Limulus polyphemus* R2 (R2Lp) (55), *Apis mellifera* R2 (R2Amel) (2), *Danio rerio* R2 (R2Dr) (54), *Hydra magnipapillata* R8 (R8Hm-A) (17), and *Adineta vaga* R9 (R9Av-1) (18). The red asterisks indicate active site residues that were previously identified and characterized by mutation to alanine in R2Bm (25). The KPD/A mutation is used as a "wild type" control in Figures 2D, 3D, 4D, and 5B (25). The R2Bm mutants generated for this study were RH/AA, RH/A, ELE/A, K/A..K/A, K/A..K, KY/A, and R/AG.W. The rounded

gray bars delineate the α-helices as determined by the 3D model presented in Figure 6. The gray arrows delineate the β-strands.

The zinc finger (ZF) and Myb motifs are the major motifs used to bind a protein subunit to a DNA sequence located downstream of the insertion site (7, 19, 20). The protein motifs used to secure the upstream subunit to the target site remain uncharacterized. The basic region (BR) is involved in RNA binding (21). The reverse transcriptase domain (RT) performs the actual TPRT reaction (3, 4, 6). A cysteine-histidine rich motif (CCHC) of largely unknown function is downstream of the RT (22-24). Beyond the CCHC motif is a PD-(D/E) restriction endonuclease-like motif (RLE). R2 elements appear to encode an endonuclease variant of the PD-(D/E)XK motif (25). The D-(D/E) residues of the R2 endonuclease have been located. The lysine has not been discovered, and the overall size and structure of the R2 endonuclease is not known (25). As DNA cleavage is such a central part of the integration reaction, a better understanding of the endonuclease encoded by R2 elements is needed.

Endonucleases of PD-(D/E)XK family include most bacterial restriction enzymes, a number of DNA repair enzymes (e.g., MutH), certain group I homing intron endonucleases, and certain Holliday junction resolvases (26, 27). Members of the PD-(D/E)XK superfamily adopt a common structural core—a restriction endonuclease-like fold (terminology from SCOP2 database)—consisting of a four-stranded, mixed β-sheet flanked by two α-helices on both sides (αββββαβ topology) (26-30). Endonucleases of this family perform various biological roles, including their involvement in DNA repair, replication, and even tRNA-intron splicing (26-28, 31). The family members share little to no sequence or structural homology beyond the conserved PD-(D/E)XK residues and the restriction endonuclease-like fold. The restriction endonuclease-like fold often contains multiple insertions and extensions, making the core fold difficult to detect (26-29). The similarity of the R2 encoded endonuclease to restriction enzymes has been long noted, and until recently the best hit on structure prediction and threading programs like Phyre2 and HHPRED was the restriction enzyme Fok I, although it modeled only part of the fold (unpublished) (25). Like the restriction DNA endonuclease Fok I, the R2 protein is thought to use distinct DNA binding domain(s) coupled to a non-specific DNA endonuclease to cleave the nonpalindromic DNA target. The recent top hits in the structural databases now include the Holliday junction resolvases E and C (Hje and Hjc) and allow more extensive modeling of the

R2 endonuclease. An article by Mukha et al. was published reporting this similarity using the *Drosophila melanogaster* R2 element as a query sequence, corroborating our work-in-progress (32).

In the canonical PD-(D/E)XK arrangement of the catalytic residues within the restriction endonuclease-like fold (αββββαβ), the aspartate residue of the PD is at or near the beginning of the second β-strand (26-29). The second aspartate (or glutamate, D/E) is in the third β-strand, and the lysine is one residue away from the aspartate/glutamate in the third β-strand (26-29). These active site residues can sometimes vary with respect to their relative positions within the αββββαβ topology, but regardless of position, they form a conserved active site geometry. The active site residues play various catalytic roles, including coordination of up to three divalent metal ions (26, 31, 33). This paper maps, models, and biochemically demonstrates that sequence motifs beyond the CCHC motif of R2 are part of the DNA endonuclease. The mutations in the RH (α-helix 1), K..KY (α-helix 2), and RG.W (carboxyl terminal to β-strand 4) motifs either knockout or reduce DNA cleavage activity. Mutations in the RH motif decrease DNA binding activity as well as DNA cleavage. Results for both the 3′ PBM RNA (presented in the paper) and the 5' PBM RNA (provided in the supplemental material) *in vitro* reactions were collected. We compare and contrast our biochemical and structural data for R2 with that of restriction endonucleases and archaeal Holliday junction resolvases.

Materials and Methods

Nucleic acid preparation

Target DNA was generated by PCR with ^{32}P end labeled primers as previously reported with the target primers listed in Supplemental Table 1 (20, 34) (See supplemental files section). The target DNA was 120 bp in length, with 70 bp of upstream DNA and 50 bp of downstream DNA, relative to the insertion dyad. DNA concentration was determined by ethidium bromide dot analysis where DNA is mixed with ethidium bromide (0.1µg/µl) for detection. Dilutions of Lambda DNA (Promega #21674203) were used to generate a standard curve. Pre-nicked target DNA was labeled on the bottom strand and nicked by reacting the target

DNA with R2 protein under conditions that nick the bottom strand. The nicked DNA was purified by phenol-chloroform extraction followed by ethanol precipitation.

5′ PBM RNA and 3′ PBM RNA were made essentially as previously reported except that RNA was gel purified on a denaturing 5% polyacrylamide gel and refolded (8, 34, 35). RNA was quantified by OD260 and additionally checked by staining with SYBR Green (LONZA #50523) and comparing the RNA to a known sample. All quantitations were done using ImageJ software analysis of digital photographs (36).

The protein expression construct contained an *Escherichia coli* codon optimized R2Bm ORF. The ORF contained amino acid residues 70-1114 (i.e. KKSN...GGVG) of the translated genbank sequence (M16558) and contained a carboxyl-terminal 6X His-tag. A QuikChange site-directed mutagenesis kit (Stratagene #200523-5) was used to generate the RH/AA, RH/A, ELE/E, K/A..K/A, K/A..K, KY/A, and R/AG.W mutants. The primers used are listed in Supplemental Table 1 (See supplemental files section). The specific mutations were chosen based on sequence alignments and protein threading models (see Figures 1 and 6). The mutated construct was then transformed into a BL21 strain of *E. coli* (Agilent #200133) by making the cells chemically competent and heat shocking them.

Protein purification

To express the proteins, 250 mL expression cultures containing the appropriate R2Bm expression construct were grown in LB broth supplemented with 50 µg/ml kanamycin in an incubator-shaker (37°C, 240 rpm). At an OD600 of between 0.8-1.0, cells were induced with 0.1mM IPTG and grown for an additional hour at 37°C. Cells were harvested by centrifugation at 3,724 X g for 20 minutes at 4°C. The cells were resuspended in 25 ml 10mM Tris pH 7.5 and centrifuged again at 400 X g for 10 minutes. The rinsed cells were stored at -80°C.

To purify R2Bm protein, cell pellets were thawed and resuspended in 2.5 ml of Buffer A (100 mM Hepes pH 7.5, 50% glycerol, 5mM β-mercapto ethanol, 2 mg/mL lysozyme). The resuspended cells were incubated at 23º C for 10 minutes. After incubation, 13.2 ml of Buffer B (100 mM HEPES pH 7.5, 1 M NaCl, 0.2% triton X-100, 5 mM β-mercapto ethanol) was added. The mixture was gently inverted 5-6 times and held on ice for 30 minutes. The cell lysate was cleared of DNA and insoluble material by centrifugation

under vacuum for 20 hours at 110,000 X g in 2° C. The clarified lysate was poured into a 15 ml tube containing 250 µl bed volume of Talon metal affinity resin (Clontech #635501) pre-equilibrated with 3 ml of wash buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 0.02% triton X-100, 10 mM Imidazole pH 7.5). The lysate plus Talon resin was agitated for 20 minutes on ice. The Talon resin was spun down at 1000 X g for 3 minutes at 4°C and the supernatant was discarded. The Talon resin was washed 2X quickly with 10 ml aliquots of wash buffer, centrifuging between washes to discard supernatant. Two more identical washes were done with an additional 10 minutes incubation with agitation prior to centrifugation. The Talon resin was then transferred to a 1.2 ml chromatography column (Biorad 732-6008) and washed with two successive 1 ml aliquots of wash buffer (gravity flow, total wash volume of 2 ml). Proteins were eluted from the resin by addition of 600 µl elution buffer (50 mM HEPES pH 7.5, 100 mM NaCl, 50% glycerol, 0.1% triton X-100, 150 mM imidazole). Proteins were stored in elution buffer supplemented with 0.1 mg/ml BSA and 2 mM DTT (final concentrations) at -20° C. Proteins were quantified by SYPRO Orange (Sigma #S5692) staining of samples run on SDS-PAGE prior to addition of BSA for storage. Dilutions of a BSA standard (Biorad #500-0202) were used to generate a standard curve. All quantitations were done using ImageJ software analysis of digital photographs (36).

R2Bm reactions and analysis

Binding, cleavage, and TPRT reactions were performed essentially as described previously (5, 8). Reactions were 13 µl and contained 80 fmol labeled substrate DNA, 150 ng of unlabeled poly-dIdC, 360-12 fmol R2Bm protein, 10 mM Tris-HCl (pH 8.0), 200 mM NaCl, 5 mM $MgCl_2$, 1mM dithiothreitol, 0.1 mg/ml bovine serum albumin, 0.01% Triton X-100, and 12% glycerol. In addition, either 1.2 pmol of R2Bm 5′ PBM RNA or 3′ PBM RNA was present. TPRT reactions contained 2.5 µM of each deoxynucleotide triphosphate (dNTP). The reactions in Figure 4A-C lacked poly-dIdC as these reactions were done prior to poly-dIdC becoming a standard component of the R2 reactions.

Reactions used to determine the relative DNA binding potential of the mutants were assembled as a master mix of all of the components minus protein and aliquoted into 10 µl reactions. Reactions were started with the addition of 3 µl of protein containing equal moles of either WT or mutant protein. The amount

of protein added was such that WT protein bound around 40-60% of the labeled DNA. The KPD/A mutant was used as the WT benchmark for mutants that had impaired DNA cleavage ability.

Reactions for determining DNA cleavage and TPRT were assembled as a master mix lacking protein and target DNA. Protein was added to aliquots and allowed to preincubate at 25º C for 15 minutes to allow the RNA to bind to the R2Bm protein. Reactions were started by the addition of substrate DNA (1µl) and continued at 37°C for 30 minutes. The reactions were chilled on ice prior to loading 9 µl of the reactions onto 5% native (1X Tris-borate-EDTA, 9.5 cm by 12.5 cm) polyacrylamide gels and run for ~50 minutes at 220 V at 4°C. The remainder of the R2Bm reaction was mixed with 95% formamide 0.5X TBE to a final concentration of formamide ≥ 70%. A 5 µl aliquot of each formamide treated sample was loaded onto denaturing (8M urea) 7% polyacrylamide gels. Typically, four protein concentrations (e.g., spanning from 120 fmol protein to 12 fmole) were used per comparative experiment.

For both approaches, two replicate experiments were performed per protein purification round and usually two independent protein purification rounds were examined. WT and mutant proteins were age matched as the DNA binding and cleavage activities decrease with time post purification.

All gels were exposed to a phosphorimager screen, which was then scanned on a molecular dynamics STORM 840 phosphorimager. The resulting 16 bit TIFF images were linearly adjusted in Photoshop so that the most intense bands were dark gray. Adjusted TIFF files were quantified using ImageJ (36). The fraction of DNA bound by the R2Bm protein in a given reaction was determined using EMSA data. The fraction of DNA cleaved by the R2Bm protein in a given reaction was determined using data from the denaturing gel.

3D modeling

R2 sequences from various R2 elements (especially R2Bm, R2Lp, R8Hm, R9Av) were submitted to Phyre2 structure prediction program as queries (37). Varying lengths of upstream and downstream of the KPD-D motif (especially for R2Bm and R2Dm) were submitted for modeling. Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (38).

Results

R2 Alignment

The size and structure of the R2 RLE is not known. To better understand the R2 RLE structure-function, point mutations were generated in conserved regions based upon 3D modeling and amino acid sequence conservation. A Clustal alignment of the RLE domain from representative D-clade and A-clade R2 elements that target the R2, R8, and R9 sites is presented in Figure 1C. The red asterisks indicate active site residues that were previously identified (25). The α-helices and β-strands are also indicated in Figure 1C. The α-helices and β-strands were determined by the 3D model (see the modeling section of the results). The R2Bm mutants generated by site directed mutagenesis for this study were located in α-helices 1 and 2 and just after β-strand 4. The mutations were RH/AA, RH/A, ELE/A, K/A..K/A, K/A..K, KY/A, and R/AG.W, as noted in Figure 1C.

Mutations in α-helix 1 affect DNA cleavage and DNA binding

To determine if the conserved RH residues located at the beginning of α-helix 1 (see Figure 1C) had an effect on DNA cleavage activity, the cleavage activity of purified RH/AA and RH/A mutant R2Bm protein were compared to the cleavage activity of purified wild type (WT) R2Bm protein *in vitro*. DNA cleavage reactions were carried out on ^{32}P end-labeled (bottom strand) target dsDNA in the presence of 3' PBM RNA. Protein was pre-bound to RNA prior to the addition of DNA. After incubation, the aliquots of the DNA cleavage reactions were assayed by electrophoretic mobility shift assays (EMSAs), run on native polyacrylamide gels (Figure 2A), as well as denaturing polyacrylamide electrophoresis (Figure 2B, full length denaturing gels are in Supplemental Figure 8 in Supplemental files section).

The EMSA allowed a qualitative assessment of the mutant's ability to form native-like protein-DNA-RNA complexes—assayed by migration distance. To assay for DNA cleavage activity, the amount of DNA cleaved per protein-bound unit of DNA was determined by combining information generated from the

EMSAs and information generated from denaturing (8M urea) polyacrylamide gels. The EMSA gel allowed the fraction of DNA that had been bound to protein to be calculated. The denaturing polyacrylamide gels allowed the fraction of DNA that had undergone a DNA cleavage event to be calculated. The results are graphed in Figure 2C as a scatter plot of DNA cleavage activity (fraction cleaved) as a function of DNA that had been bound by protein (fraction bound). The RH mutations did not appear to greatly affect the migration pattern of the protein-DNA-RNA complexes (Figure 2A). The RH/AA mutation, however, completely abolished DNA endonuclease activity (Figure 2B and 2C). The endonuclease activity of RH/A single mutation was much less affected. No cleavages beyond the R2 cleavage site were detected in either WT or mutants.

A separate experiment was performed that was tailored to focus on DNA binding activity instead of DNA cleavage activity. As the RH/AA abolished DNA cleavage, a previously characterized endonuclease mutant, KPD/A, was used as the "wild type" control (referred to as $WT^{KPD/A}$) in the DNA binding assays (Figure 2D). The KPD/A mutant lacks endonuclease activity but is otherwise WT in form and function (25). In order to reduce errors inherent in pipetting small volumes of DNA, the master mix contained all of the components except R2 protein. The binding reaction was started by the addition of R2 protein. The amount of protein added was such that the $WT^{KPD/A}$ control-reactions resulted in nearly 50% of DNA being bound by protein. The actual fraction bound was determined for each matched set of $WT^{KPD/A}$ and RH/AA mutant lanes, and the results were normalized to $WT^{KPD/A}$ DNA binding activity. $WT^{KPD/A}$ activity defined to be 100%. Replicate experiments were performed both within and across independent rounds of protein purification, and the results averaged and reported in Figure 2D. The RH/AA mutation was observed to decrease the ability of R2BM protein to bind to target DNA by about 40% in the presence of the 3' PBM RNA. The RH/A mutation was not directly tested, but appeared to have improved DNA binding over the double mutant in the DNA cleavage reaction EMSAs (Figure 2A).

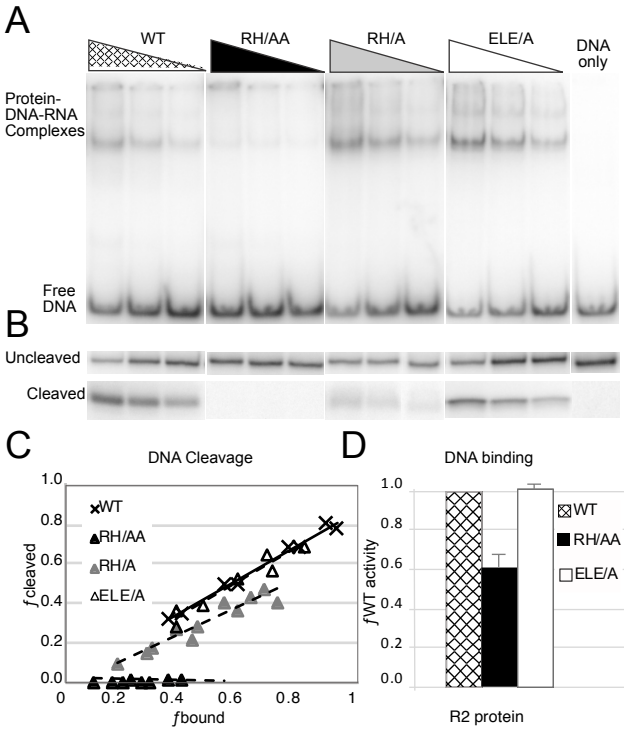


Figure 2: The RH residues in α-helix 1 affect DNA cleavage and DNA binding. DNA cleavage reaction (A-C) differ from DNA binding reactions (D). In the DNA cleavage reactions, purified mutant R2Bm protein and wild type (WT) R2 protein were pre-bound to 1.2 pmol of 3' PBM RNA and the reactions were initiated with the addition of a 120 bp segment of 28S rDNA that contained the R2 insertion site. The 120 bp target DNA had been generated by PCR and contained within it the upstream and downstream R2 protein binding sites (8). The antisense 28S PCR primer (i.e., the "bottom" strand primer) had been 5' end labeled with ^{32}P to facilitate tracking of bottom strand cleavage events of target DNA bound by R2Bm protein. After an incubation period, each reaction was split into two aliquots and loaded onto native and denaturing polyacrylamide gels for analysis. In the binding reactions, DNA was contained in the master mix and the reactions were initiated by the addition of protein in order to reduce pipetting error. (A) Representative electrophoretic mobility shift assay (EMSA) gel used to calculate the fraction of DNA bound by R2Bm protein across a range of protein concentrations (triangles). (B) Representative denaturing (8M urea) polyacrylamide gel electrophoretic analysis—of the reactions in panel A—used to determine the fraction of cleaved DNA. For graphical and space considerations, the gel has been trimmed to show just the uncleaved and cleaved DNA regions of the gel as R2 is highly site-specific and the mutations did not affect specificity. (C) A scatter plot of the fraction of cleaved DNA ($f$cleaved) as a function of the fraction of protein bound to target DNA ($f$bound). The $f$bound data were derived from EMSA gels, similar to panel A and the $f$cleaved DNA were derived from the corresponding denaturing gels.The black and gray triangles on the graph represent RH/AA and RH/A data respectively, while the "X" points on the graph represent WT data. The white triangles are the ELE/A data. (D) Bar graph reporting the DNA binding efficiency of the RH/AA and ELE/A mutant proteins calculated from binding reactions assayed on a EMSA gels. The black bar reports the DNA binding activity of RH/AA mutant protein as a fraction of $WT^{KPD/A}$ activity. The white bar reports the DNA binding activity of ELE/A mutant protein as a fraction of WT activity. The standard deviation is indicated above the bar. The DNA binding activity of WT and $WT^{KPD/A}$ was set at 1 and is represented as the hatched bar.

The RH/AA mutant was also tested for the ability to bind to target DNA in the presence of the 5' PBM RNA. 5' PBM RNA drives protein to bind to sequences downstream of the insertion site and to cleave

the top strand of the target DNA.The RH/AA mutation reduced DNA binding activity by approximately 25% in the presence of 5' PBM RNA and abolished top strand cleavage activity (Supplemental Figures 1 and 2, see Supplemental files section).

Conserved glutamate at the end of β-strand 1 does not appear to be directly involved in either DNA cleavage or DNA binding

Mutating the conserved glutamate residue positioned at the end of the first β-strand to an alanine (ELE/A) does not appear to affect DNA binding (Figure 2D), DNA cleavage (Figure 2A-C), or the migration of the protein nucleic acid complexes in mobility shift gels (Figure 2A).

Mutations in α-helix 2 affect DNA cleavage

Three mutations were generated in the K..KY motif located at the amino-terminal end of α-helix 2: K/A..K/A, K/A..K, and KY/A. DNA cleavage reactions were set up to test the mutants for loss of endonuclease function (Figure 3). The DNA cleavage reactions and subsequent analysis were identical to those already described in Figure 2. At least two replicate experiments were performed per protein purification round, and two independent protein purifications were examined. Figure 3A shows a representative EMSA gel. Each of the mutants appeared to be able to form the correct migrating protein-RNA-DNA complexes. The companion denaturing gel is shown in Figure 3B. The scatter plot of DNA cleavage as a function of DNA binding for the cumulative data sets is presented in Figure 3C. The K/A..K/A and the K/A..K mutations virtually abolished DNA cleavage, while the KY/A mutation partially impaired DNA cleavage activity per protein bound unit of DNA. None of the helix 2 mutants affected DNA cleavage specificity.

In order to determine the role, if any, the K..KY motif might have on DNA binding, the DNA binding activity of the K/A..K/A and KY/A mutants were measured relative to the DNA binding activity of the $WT^{KPD/A}$ protein (Figure 3D). Both mutants had WT, or near WT, DNA binding activity. The experimental setup in Figure 3D was identical to the analysis in Figure 2D for the RH/AA mutant.

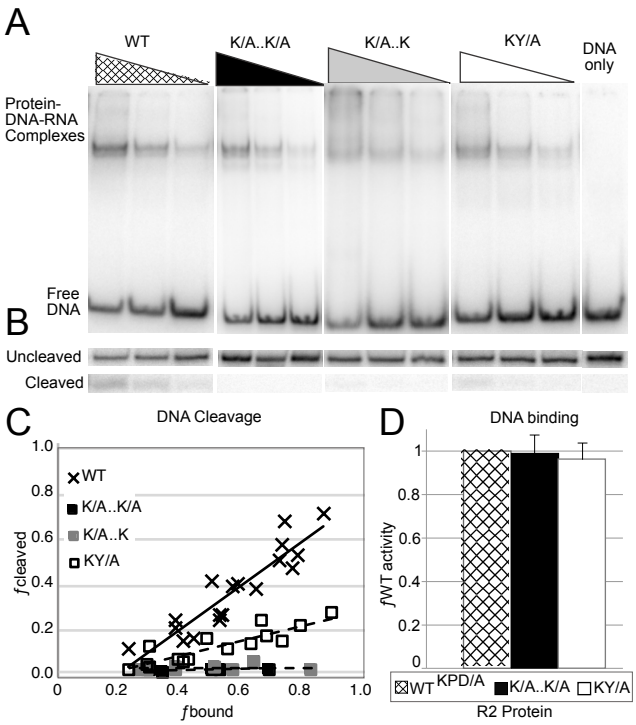


Figure 3: α-Helix 2 K..KY motif residues affect DNA cleavage but not DNA binding. DNA cleavage reactions (A-C) and DNA binding reactions (D) were performed like Figure 2. (A) Representative EMSA gels used to calculate the fraction of DNA bound by R2Bm protein across a protein concentration series. Triangles represent a protein titration series. (B) Representative denaturing gel of the reactions in panel A. (C) Scatter plot of fraction of cleaved target DNA (*f*cleaved) as a function of fraction of bound target DNA (*f*bound). All reactions, abbreviations, and graphs are as in Figure 2 except where noted. The black boxes represent the K/A..K/A data, the gray boxes represent the K/A..K data, and the white boxes represent the KY/A data. (D) Bar graph reporting the relative DNA binding efficiency of the K/A..K/A, and KY/A mutants relative to $WT^{KPD/A}$ protein. Activity of $WT^{KPD/A}$ was set at 1. All other conditions, abbreviations, and symbols are as in previous figures.

In addition to the experiments shown in Figure 3, we mutated the RE residues just in front of the K..KY motif at the beginning of α-helix 2 in R2Bm, as an internal control. The RE residues are not conserved across R2 elements and mutating these residues has no observable effect on either DNA binding or DNA cleavage (Cortez, unpublished data). Finally, the α-helix 2 mutants were also tested for the ability to bind to target DNA in the presence of 5' PBM RNA and to cleave the top strand of the target DNA. The α-helix 2 mutations had little to no effect on overall DNA binding activity in the presence of 5' PBM RNA. Both the K/A..K/A, and KY/A reduced top strand cleavage activity to below detection levels (Supplemental Figures 3 and 4, see Supplemental files section).

Mutations in the loop region following β-strand 4 affects DNA binding and DNA cleavage

As per the previous mutations, DNA cleavage and DNA binding analysis of the R/AG.W R2 mutant were carried out (Figure 4). While the R/AG.W mutant appeared to largely form the correct migrating protein-RNA-DNA complex(es) (Figure 4A), the mutant was partially impaired in DNA cleavage activity per protein bound unit at the higher protein to DNA ratios (Figure 4C). DNA cleavage specificity was maintained. The R/AG.W mutation appeared to have a moderate reduction (25%) on DNA binding activity (Figure 4D).

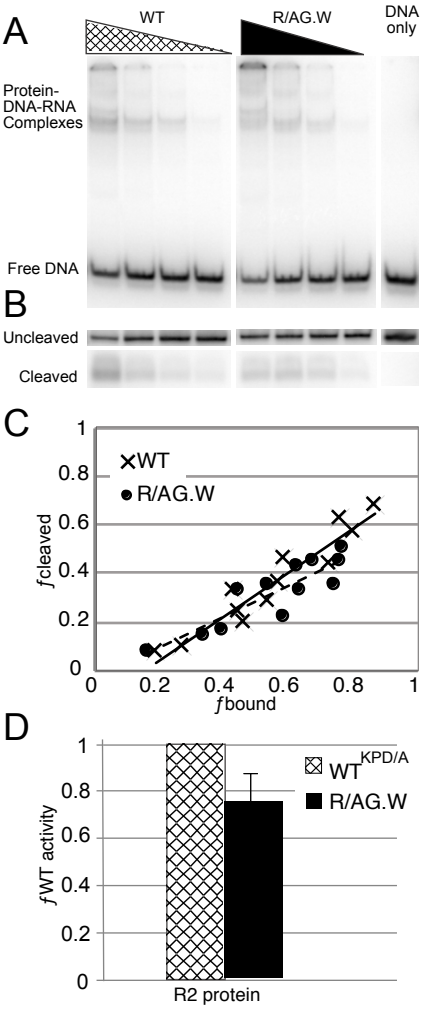


Figure 4: DNA cleavage and DNA binding activity of the R/AG.W mutant. (A) Representative EMSA gel used to calculate the fraction of target DNA bound by R2Bm protein across a protein concentration series. (B) Representative denaturing gel of the reactions in panel A. (C) Scatter plot of fraction cleaved target DNA ($f$cleaved) as a function of fraction target DNA bound by protein ($f$bound). Black circles represent data from the R/AG.W mutant. (D) Bar graph reporting the relative DNA binding efficiency of the R/AG.W mutant (black bar) relative to $WT^{KPD/A}$ protein. Activity of $WT^{KPD/A}$ was set at 1. All other abbreviations and symbols are as in previous figures.

Again, as with the previous mutations examined in this paper, the R/AG.W mutation was examined for loss-of-function in the presence of 5' PBM RNA. The results were similar to those observed in the presence of 3' PBM RNA (see Supplemental Figures 5 and 6 in supplemental files section).

Target primed reverse transcription

In order to determine if the DNA endonuclease impaired mutants RH/AA, K/A..K/A, KY/A, and R/AG.W were capable of performing TPRT, TPRT activity had to be assayed on a pre-nicked target substrate. See Figure 5A for a diagram of the experimental setup. See materials and methods for information on how the nicked DNA (60% nicked) was generated. TPRT activity of the endonuclease impaired mutants was compared to the TPRT activity of $WT^{KPD/A}$ R2Bm protein instead of WT R2Bm. $WT^{KPD/A}$ R2Bm protein is known to have WT TPRT activity given pre-nicked DNA substrates (25). The α-helix 1 mutant, RH/AA, was the only mutant that was found to be deficient in TPRT activity (Figure 5B).

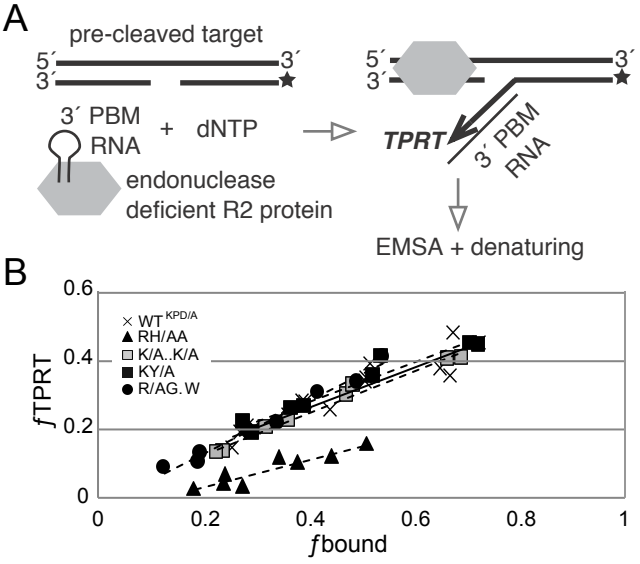


Figure 5: TPRT activity of the endonuclease impaired mutants. The RH/AA, K/A..K/A, KY/A, R/AG.W were tested for the ability to perform TPRT on pre-nicked DNA substrates. (A) Experimental design. The master mix contained 3' PBM RNA, poly-dIdC, ^{32}P end-labeled (asterisk) pre-nicked target DNA (60% nicked). Reactions were started by the addition of R2Bm protein. (B) The results are given as a scatter plot of TPRT activity, expressed as the fraction of DNA that had undergone TPRT (ƒTPRT), as a function of the fraction of target DNA bound by protein (ƒbound) at each protein concentration. All other conditions, abbreviations, and symbols are as in previous figures.

3D modeling of the endonuclease domain

In order to model the endonuclease domain of R2Bm, we submitted C-Terminal sequences of varying lengths upstream and downstream of the known catalytic residues in R2Bm to the Phyre2 structure prediction program (37). The CCHC domain and sequence downstream of the RG.W motif were not modeled well. The presence of the CCHC domain and sequence downstream of the RG.W motif often lead to perturbations of the predicted restriction endonuclease-like fold (αββββαβ) geometry and structure, especially in the extended loop between β-strands 1 and 2, in the placement of the RG.W motif, and the geometry and extent of the two α-helices (data not shown). The sequence block that generated the most reliable model for the R2Bm endonuclease is the block presented in Figure 6: the sequence starting just downstream of the CCHC and ending just past the RG.W motif. The resulting model (Figure 6A) was of generally high confidence (Figure 6G). The mutations that had a measurable effect on DNA binding or cleavage were largely on one face of the modeled protein, forming a positively charged patch (Figure 6B).

Phyre2 generated the models using modern protein threading algorithms. Three of the crystal structures that were highest weighted (Figure 6G) in the modeling of R2Bm were a domain of unknown function (duf3 fov), Hje, and Hjc (Figure 6D, 6E, and 6F, respectively) (29, 39). The individual helices and strands of the core αββββαβ architecture for duf 3fov, Hje, and Hjc are roughly equivalently positioned in 3D space except that the two α-helices in Hjc are comparatively less perpendicular. In addition, the loop between the first two β-strands of the αββββαβ core in duf 3fov is more structured and Hje contains an additional β-strand leading into the second helix. Beyond αββββαβ core, both Hjc and Hje contain additional structural motifs.

We have similarly modeled the endonuclease domains for several of the other R2 elements listed in Figure 1 (R2Dm, R2Lp, R8Hm-A, R9Av1) spanning the R2-A and R2-D clades and the R2, R8, and R9 target sites. The models for R2BM and R2Dm superimpose nicely (Figure 6A and 6C). The other R2 elements modeled also give a similar overall model. The most variation occurred in the extended loop between β-strands 1 and 2 and the loop between β-strand 3 and α-helix 2. R8Hm was the most poorly modeled, with α-helix 2 not being modeled as well (see Supplemental File 10). (Supplemental File 10 is 125 pages long containing the concatenated PDB files of R2Bm, R2Dm, R2Lp, R8Hm-A, and R9Av-1. It

can be accessed at the following website:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4838377/bin/supp_44_7_3276__index.html)

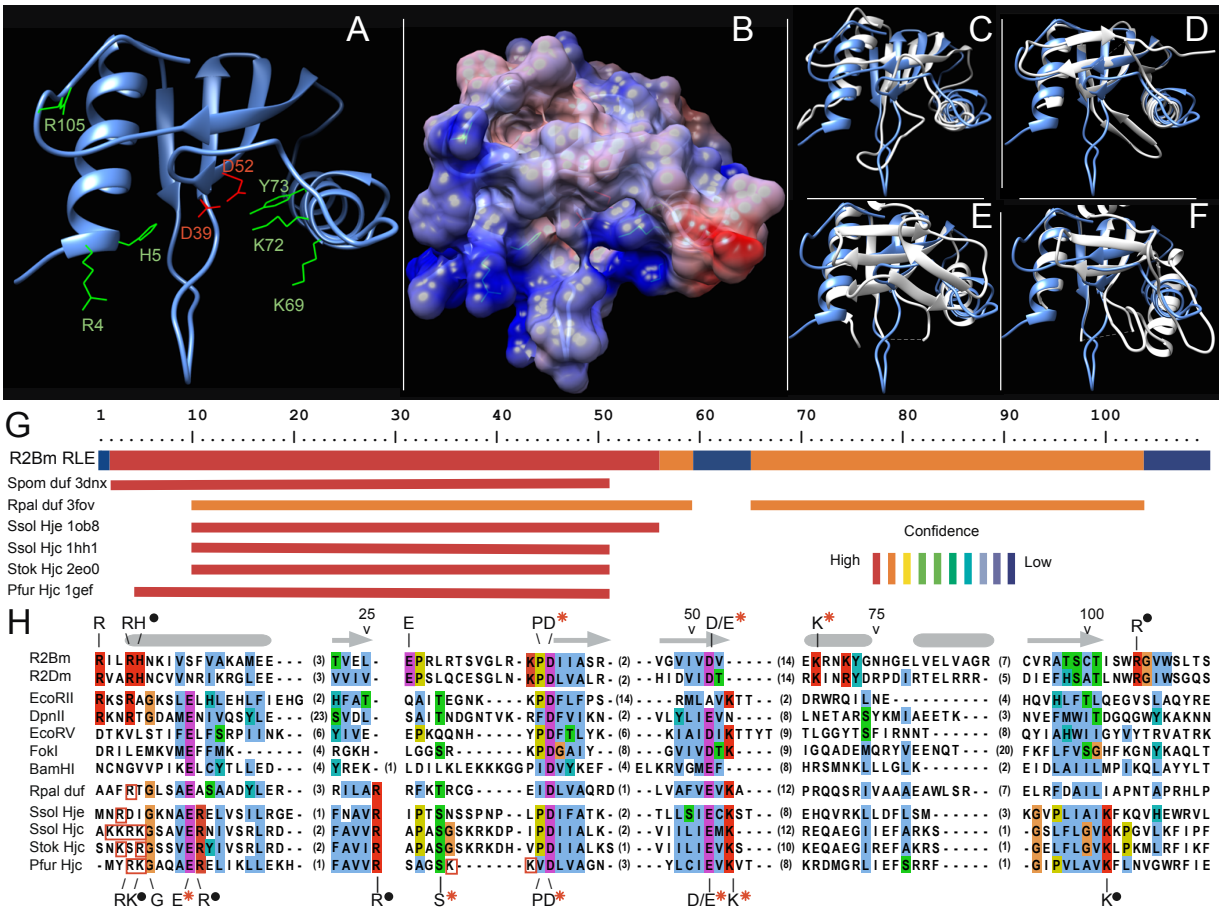


Figure 6: Structural modeling of endonuclease domain. (A) R2Bm model. The R2Bm model was generated using the Phyre2 server (37). The side chains in red are the previously characterized active site residues (25). The side chains in green are residues that were changed to alanine in this study. (B) Coulombic surface map of R2Bm. (C) R2Bm in blue and R2Dm in white overlay. (D) Overlay of R2Bm and a domain of unknown function from *Rhodopseudomonas palustris*—Rpal duf, PDB code 3fov. (E) Overlay of R2Bm and the Holliday junction resolving enzyme from *Sulfolobus solfataricus*—Ssol Hje, PDB code 1ob8. (F) Overlay of R2Bm and Holliday junction cleavage protein—Ssol Hjc, PDB code 1hh1. (G) R2Bm model construction and confidence report from Phyre2. PDB code 3dnx is a domain of unknown function (duf) derived from the uncharacterized protein SPO1766 in *Silicibacter pomeroyi*. PDB code 3fov is a duf from the uncharacterized protein rpa0323 in *Rhodopseudomonas palustris*. Holliday junction resolvases Ssol Hje 1ob8 and Ssol Hjc 1hh1 are from *Sulfolobus solfataricus.* Holliday junction resolvase Stok Hjc 2eo0 is from *Sulfolobus tokodaii* and Pfur Hjc 1gef is from *Pyrococcus furiosus*. (H) Structural plus Clustal alignment of the core αββββαβ fold from type II restriction enzymes and from the archaeal Holliday junction resolvases used to build the R2Bm model. The α-helices (rounded bars) and β-strands (arrows) are marked for R2. The amino acid numbers are given for the R2 endonuclease sequence. Known conserved catalytic (red asterisk) and DNA binding (black dot) residues for R2 and for Holliday junction resolvases are listed above and below the alignment, respectively. Arginine and Lysine residues near the start of α-helix 1 of the

Holliday junction resolvases are highlighted with a red rectangle. Other conserved residues are marked with the default Clustal coloring scheme.

Discussion

PD-(D/E)XK superfamily of endonucleases comprise a diverse set of endonucleases that share an αββββαβ restriction endonuclease-like fold and a PD-(D/E)XK motif (reviewed in (26, 28, 40)). In this superfamily, the D-(D/E)XK residues of the PD-(D/E)XK motif are known to coordinate the metal ions that are required for DNA cleavage (reviewed in (40)). Many members of the PD-(D/E)XK superfamily have an additional glutamic acid residue located in the first α-helix of the αββββαβ restriction endonuclease-like fold, yielding a motif of E-PD-(D/E)XK (26, 31). The α-helix 1 glutamate, at least in the case of *Eco*RV and *Bgl*I, is catalytic and helps to coordinate metal ions in the active site (31, 41-44). A small subset of the E-PD-(D/E)XK family members appear to have a conserved histidine (or other residue) of untested function instead of glutamate (26, 28). Holliday junction resolvases, like many of the bacterial restriction endonucleases, are of the E-PD-(D/E)XK construction: having a conserved glutamate residue (E10 as numbered in Figure 6H) in α-helix 1 that is important for DNA cleavage (29, 45-47). Point mutations in any of the E-PD-(D/E)XK catalytic residues (red asterisks in Figure 6H) of restriction endonucleases, Hjc, and Hje virtually abolish catalytic activity but do not greatly affect DNA binding (29, 48, 49). In addition to the E-PD-(D/E)XK residues, Hjc and Hje have an invariant catalytic serine located between β-strands 1 and 2 that is critical for cleavage of junctions (29).

The R2 endonuclease was modeled using archaeal Hjc and Hje structural data (Figure 6G), as such it shares the αββββαβ core fold and the placement of the PD-(D/E) residues of these enzymes. However, R2 lacks the catalytic serine and the α-helix 1 glutamate of the Holliday junction resolvases. Instead, R2 has a conserved glutamate located between β-strands 1 and 2, near the Hjc/Hje catalytic serine (Figure 6H). Unlike the Hjc/Hje serine, however, the R2 glutamate residue does not appear to be catalytic (Figure 2). The glutamate residue is not a replacement for either the Hjc/Hje serine or the missing α-helix 1 catalytic glutamate, nor is the R2 α-helix 1 histidine serving that catalytic function. The role of the R2 glutamate is as yet undetermined. The role of the RH motif is discussed below. Finally, R2 differs in the

placement of the catalytic lysine of the PD-(D/E)XK motif. The location and existence of the catalytic lysine has been an open question for R2 since PD-(D/E) catalytic residues were located (23, 25). The lysine residue in R2 is located in α-helix 2 instead of the more canonical position in β-strand 3. The first lysine (K69 of the K69..K72Y73 motif) in α-helix 2 is the catalytic lysine as it is more highly conserved (Figure 1C) and mutating it knocks out DNA cleavage (Figure 3). The second lysine (K72) is deeper into the cleft than K69. It remains a possibility that both K69 and K72 are catalytic. The side chain of the tyrosine is also oriented toward the cleft. Mutating the tyrosine reduces DNA cleavage (Figure 3C). It is unknown at this point if the tyrosine residue has any role in catalysis or if mutating the residue simply changed the geometry of the catalytic lysine residue(s). Mutations in the K..KY motif have negligible effect on DNA binding (Figure 3D). The R2 D39-(D/E)52 motif is located at the start of β-strand 2 and in β-strand 3, respectively, and are colored red in Figure 6A. These residues are in canonical positions and have been previously tested for catalytic activity in R2Bm (25).

The R2 endonuclease appears to have a large potential DNA-interaction surface—a contiguous electropositive surface surrounding the active site cleft (Figure 6B). The RH of α-helix 1, the K..K of α-helix 2, and the arginine of the RG.W motif form parts of that surface. The RH motif, when mutated, had the greatest effect on DNA binding. In the model, the histidine residue of the RH motif lies inside the active site cleft and points towards the catalytic residues. The arginine residue sits on the surface of the protein at the edge of the active site cleft. Preceding the arginine of the RH is an additional conserved arginine.

There is an arginine rich region consisting of RXXR or NXRXXR found within or in front of the first α-helix of the αββββαβ endonuclease fold of a subset of type II restriction enzymes (e.g., EcoRII, DpnII, MboI, PspGI, and Sso II, Figure 6H) (31, 40, 50, 51). Mutation of this "R-box" region can reduce DNA binding in these enzymes by several orders of magnitude and abolish DNA cleavage (40, 50, 51). In some instances, the R-Box motif is thought to make contacts with consecutive G residues in the DNA target 3' of the cleavage site (50, 51). Restriction enzymes that do not have an R-box still often use this region as a DNA contacting region. The R-box residues function as DNA recognition and correct positioning of the DNA in the active site for cleavage.

Similar to a restriction endonuclease R-box, arginine and lysine residues are found near the start of α-helix 1 of PD-(D/E)XK family Holliday junction resolvases. In the crystal structure for *Pyrococcus*

furiosus Hjc (Figure 6H, Pfur Hjc), the RK residues boxed in red coordinate a sulfate residue (45, 46). The sulfate is thought to mimic the placement of a DNA phosphate group in the crystal structure (45, 46).The RK residues are generally conserved across other Hjc members (45). Mutating the RK residues to alanines reduces binding to Holliday junctions by increasing the disassociation rate, but retain the ability to disrupt the junction basepairs and to cleave the DNA (45). The RK residues help form a stable complex with the junction DNA. Deleting the amino terminal MYRKG residues in *Pyrococcus furiosus* abolishes DNA binding altogether (46). The equivalent residues in Sulfolobus solfataricus (Ssol) Hjc crystal structure are disordered, and thought to become ordered upon DNA binding (39). The glycine residue in this region is highly conserved in the Hjc and Hje Holliday junction resolvases. The glycine is also present in some restriction endonucleases, especially those with an R-box. In Holliday junction resolvases, the glycine is thought to aid conformational changes of the N-terminal region upon binding a Holliday junction (45). The protein conformational changes and the local unstacking of DNA bases at the junction point that occur upon binding provide the induced fit required for DNA cleavage (49). R2 lacks this important glycine residue.

In addition to the R and K residues near the beginning of α-helix 1, Hjc and Hje have highly conserved basic residues located within α-helix 1 and at the ends of both β-strands 1 and 4 (highlighted with a black dot in Figure 6H). These basic residues have been biochemically and structurally implicated in interacting with DNA phosphates and positioning the junction for cleavage (29, 39, 45, 46). R2 lacks the basic residues found in α-helix 1 and at the end of β-strand 1 of the Hjc/Hje resolvases. R2, however, does have a conserved basic residue near the end β-strand 4, similar to Hjc/Hje. In the resolvases, the residue following β-strand 4 is a lysine while in R2 the conserved residue is an arginine (the R of the RG.W) motif. As is the case for the Hjc/Hje, the R2's RG.W motif appears to be involved in DNA binding.

Less conserved Hjc/Hje residues located between β-strands 1 and 2, and between β-strand 3 and α-helix 2, have also been implicated in DNA phosphate binding interactions. Point mutants for Ssol and *Pfur* Hjc basic residues in these regions reduce DNA binding and DNA cleavage (39, 45). R2 has several basic resides in the loops between β-strands 1 and 2, especially the lysine (or arginine) next to the PD motif, but largely appears to lack a similar patch of basic residues between β-strand 3 and α-helix 2.

The loop between β-strands 1 and 2 tends to be flanked by proline residues in both R2 and Hjc/Hje. The first proline is at the dimer interface in Hjc and Hje and impacts the geometry of the two subunits relative

to each other (29, 39). The second proline is presumably important in positioning the catalytic aspartic acid residue.

With the exception of the aforementioned R-box region and basic residues near the PD motif, restriction enzymes, being very diverse with many additions to the αββββαβ core structure, are difficult to make generalizations about when comparing DNA binding motifs to R2 and to Holliday junction resolvases. All three, however, appear to use flexible regions that become more structured upon binding to DNA. The induced fit of these unstructured regions is important for positioning the DNA near catalytic D-(D/E)XK residues.

The RH motif, and likely the preceding arginine residue of R2 (Figure 1C), appears to form an R-box equivalent involved in DNA binding. As in both restriction enzymes and the Holliday junction resolvases Hjc and Hje, major perturbations of the R-box region reduce both cleavage and DNA binding. It is possible that the DNA binding role of the R2 R-box residue(s) might have a site-specific component to it, as the DNA binding defects are greater in the presence of 3' PBM RNA than they are in the presence of 5' PBM RNA. In the presence of 3' PBM RNA, the R2 protein would be expected to bind upstream of the insertion site (5). In the presence of 5' PBM RNA, the protein would be expected to bind downstream of the insertion site (8). The RH/AA mutation reduced DNA binding activity by 40% in the presence of 3' PBM RNA (Figure 2D), while in the presence of 5' PBM RNA DNA binding activity was only reduced by 25% (Supplemental Figure 1, Supplemental files section). However, the fact that the RH motif is present in both A and D clade R2 elements—regardless of the DNA sequence being targeted by the element (Figure 1C)—argues against the reduction in DNA binding being due to site-specific protein-DNA contacts. In addition, the RH/AA mutant shows a similar 40% reduction in DNA binding compared to WT R2 protein when WT and RH/AA R2 proteins are forced to bind to non-target DNA in the presence of non-specific RNA (Supplemental Figure 7 in Supplemental files section), possibly arguing against sequence specific role for the RH domain.

The other R2 mutation that decreased DNA binding activity was the R/AG.W mutant. The mutation decreased DNA binding activity relative to WT R2 by 25% in the presence of 3' PBM RNA and 21% in the presence of 5' PBM RNA, although the reduction in DNA binding might be greater than that indicated in Figure 4D because the R/AG.W mutant retains most of its DNA cleavage activity. The $WT^{KPD/A}$ R2 protein, to which the R/AG.W mutant is being compared for DNA binding, does not cleave target DNA and the R2

protein yields a greater DNA footprint post DNA cleavage (34). The difference in affinity of the R2 protein for DNA pre and post DNA cleavage appears to be small (Govindaraju, unpublished data) at the levels of protein and DNA used in our reactions. The RG.W motif is located at a sharp turn in a coiled region in the R2 models, with the R residue side chain running roughly parallel to the side chain of the R in the RH motif (Figure 6A). In Hjc and Hje the equivalent sharp turn to the RG.W turn is located between two β-strands and is near to or including the K residue following β-strand 4 of Hjc/Hje (Figures 6E, 6F, and 6H). Like the arginine in the α-helix 1 RH motif of R2, the arginine of the RG.W motif likely makes contact with the target DNA. The arginine residues from both motifs are part of the extended basic patch that likely constitutes the DNA contacting surface of the R2 endonuclease (Figure 6B).

Overall the model generated for the R2 endonuclease appears to be a good predictor of form and function. A summary of our biochemical results is presented in Table 1. A large contiguous electropositive surface is present that includes not only the catalytic cleft but also the surface surrounding the cleft (Figure 6B). The arginine of the α-helix 1 RH motif and the arginine of the RG.W motif form part of that surface. The D-(D/E) catalytic residues are properly modeled in the cleft with the correct orientations compared to known crystal structures. The first lysine residue of α-helix 2 appears to be the "missing" catalytic lysine residue identified for other members of the PD-(D/E)XK endonuclease superfamily. The spacing and placement of the R2 lysine in the restriction endonuclease-like fold, however, is noncanonical, with the lysine being far downstream of the D/E and being located in the second α-helix. We were unable to locate a catalytic residue upstream of the PD motif that might function as the α-helix 1 glutamate of an E-PD-(D/E)XK motif or serine of a Hjc/Hje E-S-PD-(D/E)XK motif.

Table 1: Summary of DNA binding and cleavage results.

R2 Protein	Target DNA Binding Activity	Target DNA Cleavage Activity	TPRT activity	Target DNA Binding Activity	Target DNA Cleavage Activity	Non-target dsDNA Binding Activity	non R2 site Cleavage Activity
	(3' PBM RNA)	(3' PBM RNA)	(3' PBM RNA)	(5' PBM RNA)	(5' PBM RNA)	(non-specific RNA)	(3' PBM RNA)
WT	N.A.	N.A.	N.A.	N.A.	N.A.	non-specific associations$^{\delta}$	none detected
RH/AA	40% reduction	none detected	reduced	30% reduction	none detected	40% reduction	none detected
RH/A	N.T.	reduced	N.T.	N.T.	N.T.	N.T.	none detected

ELE/A	normal	normal	N.T.	N.T.	N.T.	N.T.	none detected
K/A..K/A	normal	none detected	normal	normal	none detected	N.T.	none detected
K/A..K	N.T.	none detected	normal	N.T.	N.T.	N.T.	none detected
KY/A	normal	reduced	normal	normal	reduced	N.T.	none detected
R/AG.W	25% reduction	marginal reduction	normal	21% reduction	marginal reduction	N.T.	none detected

Legend: Not applicable (N.A.). Not tested (N.T.). $^{\delta}$ See Supplemental Figure 9 in supplemental files section.

It is interesting that, of the PD-(D/E)XK endonuclease superfamily with crystal structure data, R2 would be most similar to Holliday junction resolvases; the closest PD-(D/E)XK member previously had been bacterial restriction enzymes (esp., Fok I). If the structural connection to Holliday junction resolvases is more than structural resemblance but also is borne out functionally, we might predict that R2 could look and behave similar to Holliday junction resolvases because transposition intermediates in the integration reaction resemble Holliday junction-like structures. TPRT generates a three-way junction (three-quarters of a Holliday junction). The start of second strand synthesis generates a transient four-way Holliday junction-like nucleic acid structure. We doubt that R2 is a bona fide Holliday junction resolvase catalytically as R2 lacks two major catalytic residues (E, S) and several conserved basic residues characteristic of Hjc and Hje. Rather R2 must function like a restriction endonuclease on linear dsDNA yet keep track of and cleave multi-branched DNA structures that form during the integration reaction. If this conjecture is true, the reduced TPRT activity observed when the RH motif is mutated might be the result of a diminished ability to bind and coordinate a three-way junction. Alternatively, the RH motif might form part of an RNA binding surface as well as a DNA binding surface.

Funding

This work was supported by the National Science Foundation [0950983].

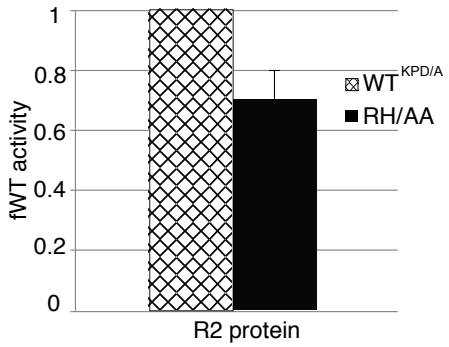
Acknowledgments

The authors would like to thank Kim Bowles, Murshida Mahbub, Monika Pradhan, and Brijesh Khadgi for critical reading of the manuscript and for helpful discussions. The authors also wish to thank Micki Christensen for copyediting, Dr. Subhrangsu Mandal for the kind use of his phosphorimager, Monika Pradhan for helping with finalizing some data, and Murshida Mahbub for initially generating some of the helix two mutants.

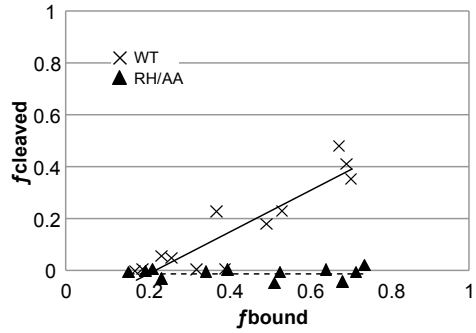
Supplemental Files section:

Supplemental Table 1: List of primers used in this study

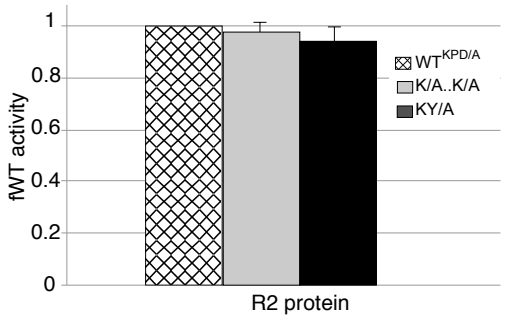
Primer	Sequence (5'-3')
RH/AA, forward	GCACCATCAGCTGGGCGGGTGTGTGGAGCTTG
RH/AA, reverse	CAAGCTCCACACACCCGCCCGCTGATGGTGC
RH/A, forward	GTGGTCGCATCCTGCGTGCGAACAAGATCGTTAGCTTTGTC
RH/A, reverse	GACAAAGCTAACGATCTTGTTCGCACGCAGGATGCGACCAC
VELE/A, forward	GAACAAATGGACCGTCGAGCTGGCGCCGCGCCTGCGTACCTCCG
VELE/A, reverse	CGGAGGTACGCAGGCGCGGCGCCAGCTCGACGGTCCATTTGTTC
K/A..K/A, forward	GACGAACTGCACCGCGAGGCGCGCAATGCGTATGGTAATCACGGC
K/A..K/A, reverse	GCCGTGATTACCATACGCATTGCGCGCCTCGCGGTGCAGTTCGTC
K/A..K, forward	GACGAACTGCACCGCGAGGCGCGCAATAAGTATGGTAATCACGGC
K/A..K, reverse	GCCGTGATTACCATACTTATTGCGCGCCTCGCGGTGCAGTTCGTC
KY/A, forward	GCACCGCGAGAAACGCAATAAGGCGGGTAATCACGGCGAGCTGG
KY/A, reverse	CCAGCTCGCCGTGATTACCATACGCATTGCGTTTCTCGCGGTGC
R/AG.W, forward	GCACCATCAGCTGGGCGGGTGTGTGGAGCTTG
R/AG.W, reverse	CGTGGTAGTCGACCCGCCCACACACCACGAAC
Target 28S DNA Primers, forward	GCTCTGAATGTCAACGTGAAGAAATTCAAGC
Target 28S DNA Primers, reverse	TAATCCATTCATGCGCGTCACTAATTAGATGACG



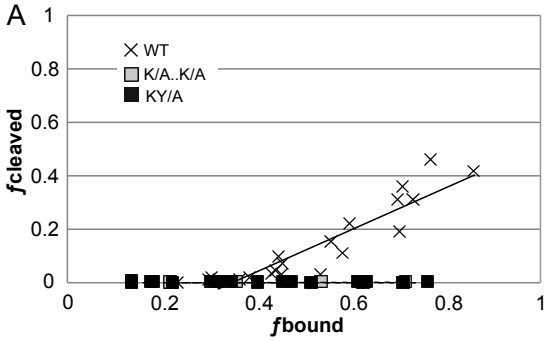
Supplemental Figure 1: DNA binding activity of the RH/AA mutant relative to WT$^{KPD/A}$ R2Bm protein in the presence of 5' PBM RNA. The master mix contained 5' PBM RNA, poly-dIdC, and top strand labeled target DNA. The reactions were started by the addition of R2Bm protein. The bar graph reports the DNA binding efficiency of the RH/AA mutant protein relative to WTKPD/A protein. The DNA binding activity of WTKPD/A was set at 1 and is represented as the hatched bar. The black bar reports the DNA binding activity of RH/AA mutant protein as a fraction of WT$^{KPD/A}$ activity—averaged over experimental replicates. The standard deviation is indicated above the bar.



Supplemental Figure 2: DNA cleavage activity of the RH/AA mutant relative to WT R2Bm protein in the presence of 5' PBM RNA. DNA cleavage reactions were carried out on ^{32}P end-labeled (top strand) target DNA in the presence of 1.2 pmole of 5' PBM RNA. Protein was pre-bound to RNA prior to the addition of DNA. The scatter plot plots the fraction of cleaved DNA (ƒcleaved) as a function of the fraction of protein bound to target DNA (ƒbound). The ƒbound data were derived from EMSA gels and the ƒcleaved DNA were derived from the corresponding denaturing gels. The black triangles on the graph represent RH/AA data while the "X" points on graph represent WT data.

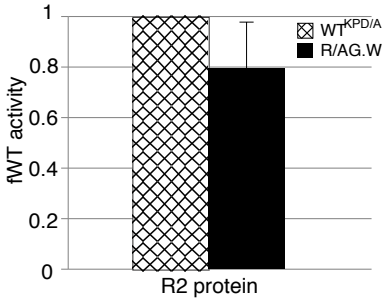


Supplemental Figure 3: DNA binding activity of the K/A..K/A and KY/A mutants relative to WT$^{KPD/A}$ R2Bm protein in the presence of 5' PBM RNA. The master mix contained 5' PBM RNA, poly-dIdC, and labeled target DNA. The reactions were started by the addition of R2Bm protein. The bar graph reports the relative DNA binding efficiency of the K/A..K/A and KY/A mutants relative to WT$^{KPD/A}$ protein. Activity of WT$^{KPD/A}$ was set at 1. The hatched bar is the WT$^{KPD/A}$ data, the grey bar is K/A..K/A data, and the black bar is KY/A data.

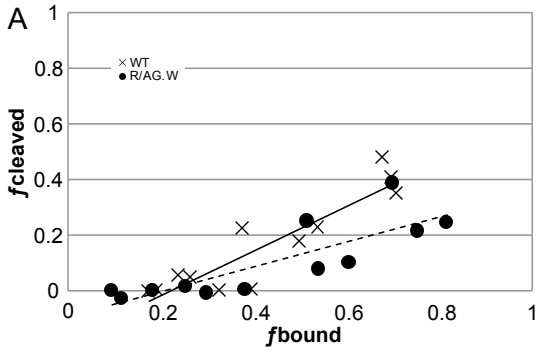


Supplemental Figure 4: DNA cleavage of the K/A..K/A and KY/A mutants. The DNA cleavage reactions were performed in the presence of poly-dIdC, R2Bm protein, and 5' PBM RNA. The reactions were started by the addition of ^{32}P end-labeled (top strand) target DNA to the reaction. The scatter

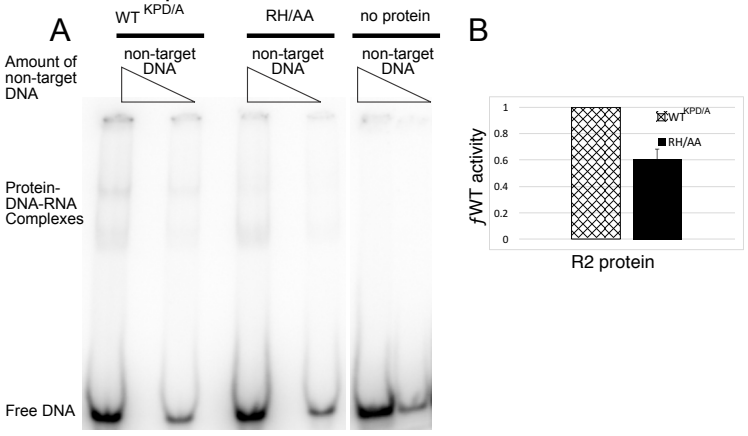
plot plots the fraction of cleaved DNA (ƒcleaved) as a function of the fraction of protein bound to target DNA (ƒbound). The "X" data points represent WT data, the grey boxes represent the K/A..K/A data, and the black boxes represent the KY/A data.



Supplemental Figure 5: DNA binding activity of the R/AG.W mutant relative to WT$^{KPD/A}$ R2Bm protein. Master mix contained 5' PBM RNA, poly-dIdC, and labeled target DNA. The reactions were started by the addition of R2Bm protein. The bar graph reports the relative DNA binding efficiency of the R/AG.W mutant (black bar) relative to WT$^{KPD/A}$ protein (hatched bar). Activity of WTKPD/A was set at 1.

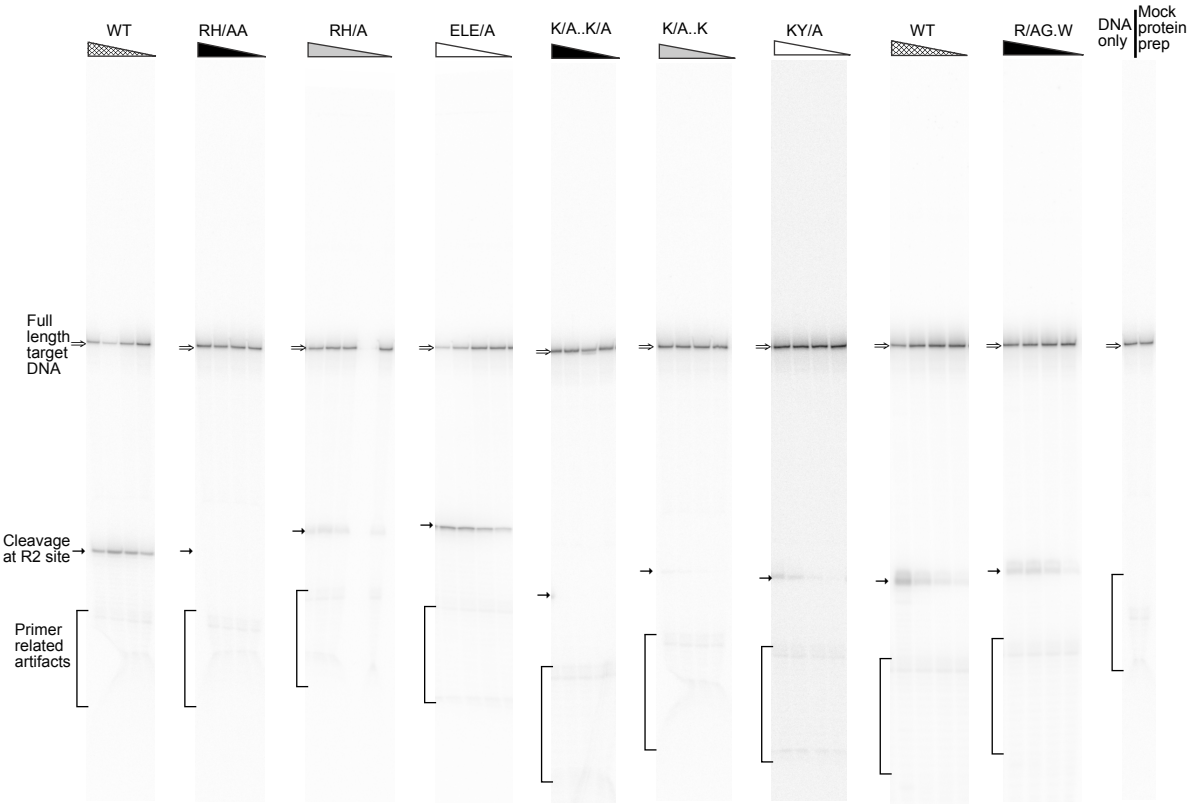


Supplemental Figure 6: DNA cleavage activity of the R/AG.W mutant. The DNA cleavage reactions were performed in the presence of R2Bm protein and 5' PBM RNA. The reactions were started by the addition of ^{32}P end-labeled (top strand) target DNA to the reaction. Scatter plot of fraction cleaved target DNA (ƒcleaved) as a function of fraction of bound target DNA (ƒbound). The "X" points represent WT data and the black circles represent data from the R/AG.W mutant.

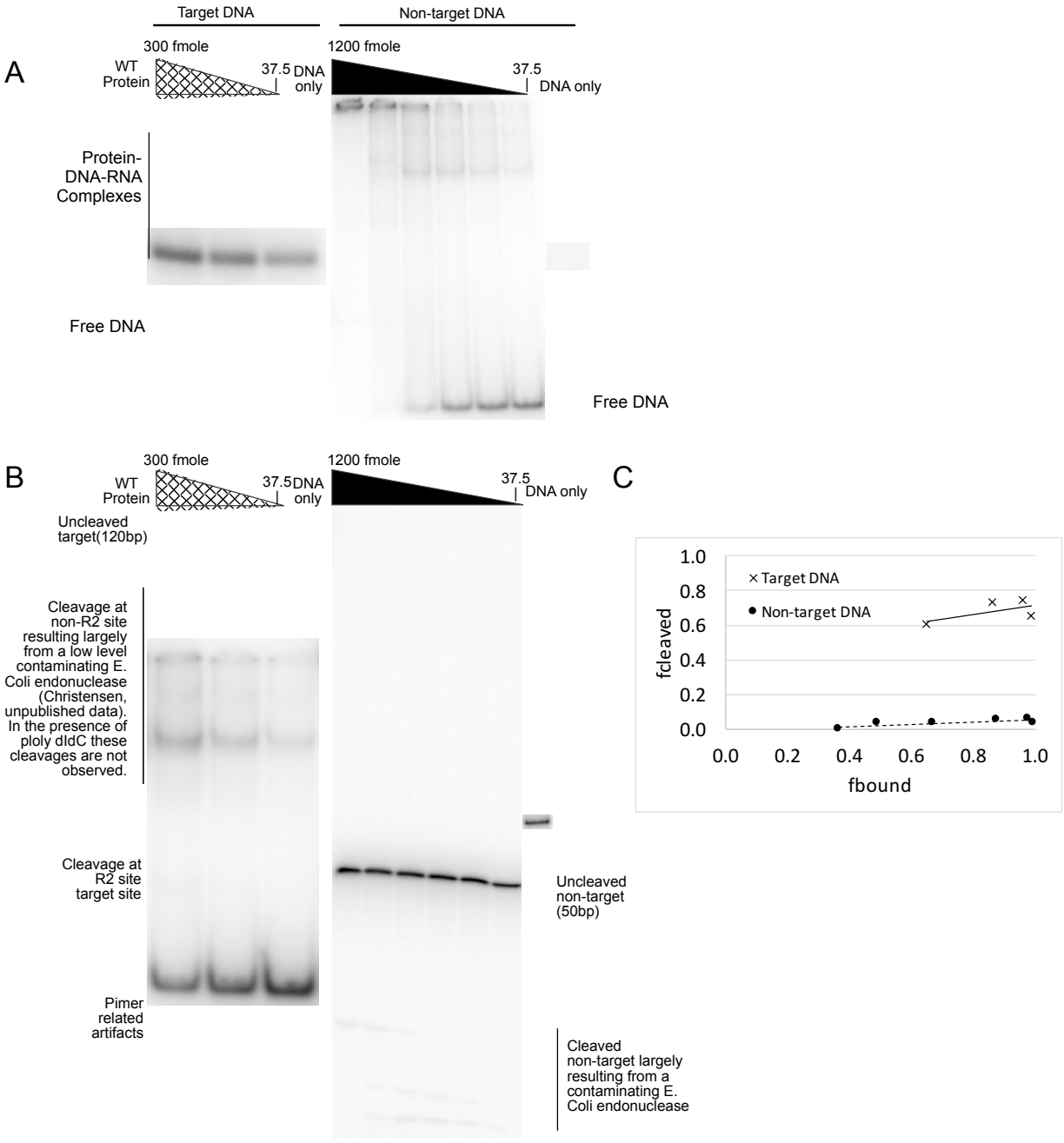


Supplemental Figure 7: Non-target DNA binding activity of the RH/AA mutant relative to WT$^{KPD/A}$ R2Bm protein in the presence of non-specific RNA. The master mix contained non-specific RNA, poly-dIdC, and labeled DNA. (A) Representative EMSA gel with fixed amount of protein (260 fmole/reaction) and

varying amounts of non-target DNA (80 fmole and 20 fmole, triangles). (B) The bar graph reports the DNA binding efficiency of the RH/AA mutant protein relative to $WT^{KPD/A}$ protein. The DNA binding activity of $WT^{KPD/A}$ was set at 1 and is represented as the hatched bar. The black bar reports the DNA binding activity of RH/AA mutant protein as a fraction of $WT^{KPD/A}$ activity—averaged over experimental replicates. The standard deviation is indicated above the bar.



Supplemental Figure 8: Representative denaturing gels from DNA cleavage reactions with the bottom strand labeled in the presence of 3' PBM RNA and various WT and mutant R2 proteins. Uncleaved DNA is marked with a double arrow (⇒). DNA cleaved at the R2 site is marked with a single arrow (→). Primer related artifacts are marked with a bracket. The position of the full length (uncleaved DNA) is graphically positioned to be in a matching position. The gels were not run a uniform distance so the distance between the uncleaved, the cleaved DNA, and the primer related artifacts varies. R2 can cleave several bases within the cleavage site area (see reference 36) sometimes leading to a doublet or blurry looking cleaved band depending on gel resolution.



Supplemental Figure 9: DNA binding and cleavage activity of R2 protein preparation on target and non-target DNA. The DNA cleavage reactions were performed in the presence of R2Bm protein and 120bp target DNA or 50bp non-target DNA. Non-target duplex DNA is synthesized by annealing complementary sequences of random bases and the annealed duplex was gel purified. The reactions were started by the addition of either 32P end-labeled (top strand) target DNA or 32P end-labeled (top strand) non-target DNA to the reaction. Poly dIdC was not used in these reactions in order to force binding to non-target DNA. (A) Representative electrophoretic mobility shift assay (EMSA) gel used to calculate the fraction of DNA bound by R2Bm protein across a range of protein concentrations (triangles). (B) Representative denaturing (8M urea) polyacrylamide gel electrophoretic analysis—of the reactions in panel A—used to determine the fraction of cleaved DNA. (C) A scatter plot of the fraction of cleaved DNA ($f$cleaved) as a function of the fraction of protein bound to target DNA ($f$bound). The $f$bound data were derived from EMSA gels, similar

to panel A and the *f*cleaved DNA were derived from the corresponding denaturing gels. The "X" points on the graph represent WT with target DNA data and solid circles represent WT with non-target DNA data.

References

1. 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.
2. Kojima,K.K. and Fujiwara,H. (2005) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165.
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. 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.
6. 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.
7. 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.
8. 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.
9. Aksoy,S., Williams,S., Chang,S. and Richards,F.F. (1990) SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. *Nucleic Acids Res* **18**, 785-792.
10. Teng,S.C., Wang,S.X. and Gabriel,A. (1995) A new non-LTR retrotransposon provides evidence for multiple distinct site-specific elements in Crithidia fasciculata miniexon arrays. *Nucleic Acids Res* **23**, 2929-2936.
11. 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.
12. Malik,H.S. and Eickbush,T.H. (2000) NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. *Genetics* **154**, 193-203.
13. Volff,J.N., Korting,C., Froschauer,A., Sweeney,K. and Schartl,M. (2001) Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. *J Mol Evol* **52**, 351-360.
14. Burke,W.D., Muller,F. and Eickbush,T.H. (1995) R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res* **23**, 4628-4634.
15. Villanueva,M.S., Williams,S.P., Beard,C.B., Richards,F.F. and Aksoy,S. (1991) A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of Trypanosoma cruzi. *Mol Cell Biol* **11**, 6139-6148.

16. Jakubczak,J.L., Xiong,Y. and Eickbush,T.H. (1990) Type I (R1) and type II (R2) ribosomal DNA insertions of Drosophila melanogaster are retrotransposable elements closely related to those of Bombyx mori. *J Mol Biol* **212**, 37-52.

17. 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.

18. 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.

19. Shivram,H., Cawley,D. and Christensen,S.M. (2011) Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements* **1**, 169-178.

20. Thompson,B.K. and Christensen,S.M. (2011) Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: plasticity of integration mechanism *Mobile Genetic Elements* **1**, 29-37.

21. Jamburuthugoda,V.K. and Eickbush,T.H. (2014) Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res* **42**, 8405-8415.

22. Wagstaff,B.J., Barnerssoi,M. and Roy-Engel,A.M. (2011) Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. *PLoS One* **6**, e19672.

23. 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.

24. Doucet,A.J., Hulme,A.E., Sahinovic,E., Kulpa,D.A., Moldovan,J.B., Kopera,H.C., Athanikar,J.N., Hasnaoui,M., Bucheton,A., Moran,J.V. and Gilbert,N. (2010) Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet* **6**,

25. 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.

26. Steczkiewicz,K., Muszewska,A., Knizewski,L., Rychlewski,L. and Ginalski,K. (2012) Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res* **40**, 7016-7045.

27. Kosinski,J., Feder,M. and Bujnicki,J.M. (2005) The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics* **6**, 172.

28. Kinch,L.N., Ginalski,K., Rychlewski,L. and Grishin,N.V. (2005) Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res* **33**, 3598-3605.

29. 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.

30. Andreeva,A., Howorth,D., Chothia,C., Kulesha,E. and Murzin,A.G. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* **42**, D310-4.

31. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell Mol Life Sci* **62**, 685-707.

32. 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.

33. Mones,L., Kulhánek,P., Florián,J., Simon,I. and Fuxreiter,M. (2007) Probing the two-metal ion mechanism in the restriction endonuclease BamHI. *Biochemistry* **46**, 14514-14523.

34. 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.

35. Kierzek,E., Kierzek,R., Moss,W.N., Christensen,S.M., Eickbush,T.H. and Turner,D.H. (2008) Isoenergetic penta- and hexanucleotide microarray probing and chemical mapping provide a secondary structure model for an RNA element orchestrating R2 retrotransposon protein function. *Nucleic Acids Res* **36**, 1770-1782.

36. Abramoff,M.D., Magelhaes,P.J. and Ram,S.J. (2004) Image Processing with ImageJ *Biophotonics International* **11**, 36-42.

37. Kelley,L.A., Mezulis,S., Yates,C.M., Wass,M.N. and Sternberg,M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858.

38. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612.

39. Bond,C.S., Kvaratskhelia,M., Richard,D., White,M.F. and Hunter,W.N. (2001) Structure of Hjc, a Holliday junction resolvase, from Sulfolobus solfataricus. *Proc Natl Acad Sci U S A* **98**, 5509-5514.

40. Pingoud,V., Sudina,A., Geyer,H., Bujnicki,J.M., Lurz,R., Lüder,G., Morgan,R., Kubareva,E. and Pingoud,A. (2005) Specificity changes in the evolution of type II restriction endonucleases: a biochemical and bioinformatic analysis of restriction enzymes that recognize unrelated sequences. *J Biol Chem* **280**, 4289-4298.

41. Kostrewa,D. and Winkler,F.K. (1995) Mg2+ binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 A resolution. *Biochemistry* **34**, 683-696.

42. Newman,M., Lunnen,K., Wilson,G., Greci,J., Schildkraut,I. and Phillips,S.E. (1998) Crystal structure of restriction endonuclease BglI bound to its interrupted DNA recognition sequence. *EMBO J* **17**, 5466-5476.

43. Groll,D.H., Jeltsch,A., Selent,U. and Pingoud,A. (1997) Does the restriction endonuclease EcoRV employ a two-metal-Ion mechanism for DNA cleavage? *Biochemistry* **36**, 11389-11401.

44. Horton,N.C. and Perona,J.J. (2004) DNA cleavage by EcoRV endonuclease: two metal ions in three metal ion binding sites. *Biochemistry* **43**, 6841-6857.

45. Nishino,T., Komori,K., Ishino,Y. and Morikawa,K. (2001) Dissection of the regional roles of the archaeal Holliday junction resolvase Hjc by structural and mutational analyses. *J Biol Chem* **276**, 35735-35740.

46. 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.

47. Pingoud,A., Wilson,G.G. and Wende,W. (2014) Type II restriction endonucleases--a historical perspective and more. *Nucleic Acids Res* **42**, 7489-7527.

48. Daiyasu,H., Komori,K., Sakae,S., Ishino,Y. and Toh,H. (2000) Hjc resolvase is a distantly related member of the type II restriction endonuclease family. *Nucleic Acids Res* **28**, 4540-4543.

49. Kvaratskhelia,M., Wardleworth,B.N., Norman,D.G. and White,M.F. (2000) A conserved nuclease domain in the archaeal Holliday junction resolving enzyme Hjc. *J Biol Chem* **275**, 25540-25546.

50. Pingoud,V., Conzelmann,C., Kinzebach,S., Sudina,A., Metelev,V., Kubareva,E., Bujnicki,J.M., Lurz,R., Lüder,G., Xu,S.Y. and Pingoud,A. (2003) PspGI, a type II restriction endonuclease from the extreme thermophile Pyrococcus sp.: structural and functional studies to investigate an evolutionary relationship with several mesophilic restriction enzymes. *J Mol Biol*

CHAPTER 3

DNA STRUCTURE AND SEQUENCE REQUIREMENTS FOR SPECIFYING SECOND-STRAND CLEAVAGE AND SECOND-STRAND SYNTHESIS DURING R2 INTEGRATION[1]

Aruna Govindaraju and Shawn M. Christensen

[1] Manuscript presented here is prepared for submission to journal article

Abstract

Long interspersed nucleotide elements (LINEs) are an ancient and important group of transposable elements (TE) that replicate by target primed reverse transcription (TPRT). TPRT technically refers to the first half of the integration reaction where a TE encoded endonuclease creates a targeted nick at an unoccupied chromosomal site. The element encoded reverse transcriptase uses the exposed target DNA 3'-OH to prime reverse transcription of the element RNA. There is a significant amount of biochemical knowledge about TPRT using in vitro reactions and bioinformatics. Very little, however, is known about the second half of the integration reaction, particularly biochemically. The second half of the reaction is thought to be conceptually similar to the first half of the reaction; the element endonuclease cleaves the remaining chromosomal DNA strand and exposed 3'-OH is used to prime second-strand DNA synthesis. Second-strand cleavage remains enigmatic and second-strand synthesis has not been observed or biochemically dissected. Using the protein and RNA components from a site-specific LINE from *Bombyx mori*, R2Bm, we explored the second half of the integration reaction in vitro. Our results indicate that DNA structure is a critical component of second-strand cleavage. A post-TPRT template-jump from the cDNA to the target DNA, prior to second-strand cleavage, is a hitherto unknown and unexplored reaction intermediate. The template-jump creates a 4-way junction that is recognized by the element endonuclease. The endonuclease recognizes key structural and sequence aspects of the 4-way junction. Second-strand cleavage occurs diagonally across from the first DNA cleavage event, similar to a Holliday junction resolvase. Cleavage of the 4-way junction results in a primer-template setup between the cleaved target DNA and the cDNA. This paper reports, for the first time, productive second-strand cleavage and demonstrable second-strand synthesis activity of a LINE in vitro.

Introduction

Long interspersed nucleotide elements (LINEs), also known as non-long terminal repeat (non-LTR) retrotransposons, are an abundant group of transposable elements (TE). LINEs replicate by priming reverse transcription of their RNA from a chromosomal nick, a process called target primed reverse transcription (TPRT) (*1-3*). Almost all LINEs encode a DNA endonuclease that generates the chromosomal nick and a

reverse transcriptase (RT) with which to reverse transcribe the element RNA (*4*). Early branching clades of elements encode a single ORF with reverse transcriptase and restriction-like endonuclease domains (*4, 5*). Later branching clades have two open reading frames, the second of which is similar to the single ORF of early branching elements except it codes for an apurinic-apyrimidinic endonuclease (APE) rather than a RLE (*6*). There are at least 16 early branching clades of LINES (*7*). The R2 element from *Bombyx mori,* R2Bm, has been a proven model system for studying the insertion reaction of LINEs, particularly the early branching LINEs. R2 elements are site specific, usually targeting a specific site in the 28S rRNA gene (*8-10*). The ORF of R2 elements encode a single open reading frame with N-terminal zinc finger(s) (ZF) and myb domains (Myb), a central reverse transcriptase (RT), a restriction-like endonuclease (RLE), and a C-terminal gag-knuckle-like CCHC motif (Figure 1A). The R2Bm protein has been expressed in *E.coli* and purified for use in vitro reactions. The R2Bm protein has been shown to be a multifunctional protein that binds to its own mRNA and to target DNA (*11*).

In vitro studies of R2Bm protein and RNA have led to the current model of integration (Figure 1B). Two subunits of R2 protein, one bound to the 3' protein binding motif (PBM) of the R2 RNA and other to the 5' PBM, are involved in the integration reaction. The 3' and 5' PBM RNAs dictate the roles of the two subunits and coordinate a series of DNA cleavage and polymerization steps that lead to integration. The protein subunit bound to the element RNA'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 bottom-strand 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 the way of ZF and Myb domains. The downstream subunit's RLE cleaves the second (top) DNA strand and its RT is hypothesized to perform second-strand DNA synthesis. Top-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 (*12-14*).

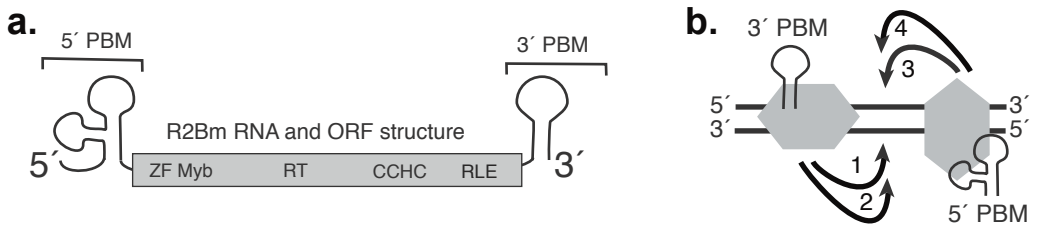


Figure 1: R2Bm structure. a) R2Bm RNA and open reading frame (ORF) structure. The ORF of R2Bm encodes several conserved motifs of known and unknown functions. Abbreviations: zinc finger (ZF), Myb (Myb), reverse transcriptase domain (RT), a cysteine-histidine rich motif (CCHC), and a PD-(D/E)XK type restriction-like endonuclease (RLE). RNA motifs present in the 5′ and 3′ untranslated regions that bind R2 protein are marked as 5′ and 3′ protein binding motifs (PBMs), respectively. Brackets indicate the individual segments of the R2Bm RNA (i.e., the 5′ and 3′ PBM RNAs) that were used in this paper. b) The R2 insertion mechanism is depicted: (1) DNA cleavage of the bottom strand, (2) TPRT, (3) DNA cleavage of the top strand, and (4) second strand DNA synthesis. (This figure is adapted with permission from Govindaraju et al, 2016 (*15*)).

While the first half of the integration reaction is fairly well understood, the second half of the reaction, second-strand cleavage and second-strand synthesis, is not well understood. Second-strand synthesis has always been enigmatic and requires a narrow range of protein, RNA, and DNA ratios. It is also not clear how two specific cleavages, top and bottom-strand, can occur as the insertion site is not palindromic. Finally, second-strand cleavage divorces the two half reactions. After second (top) strand cleavage, no covalent bonds or base pairing interactions keep the upstream and downstream half sites in contact. The two half sites separate in in vitro reactions making priming of second-strand cDNA synthesis from the top-strand cleavage event very unlikely.

The RLE found in early branching LINEs is a variant of the PD-(D/E)XK superfamily of endonucleases (*5*). In a previous paper, we had reported the similarity of the LINE RLE as having sequence and structural homology to archaeal Holliday junction resolvases (*16*). Our previous paper left open the question as to whether or not R2 protein could function as a Holliday junction resolvase and to what, if any, relevance this putative function might play in the insertion mechanism. In this paper, the ability to of R2 protein to bind to and to perform integration functions on branched DNAs is explored. Our results indicate that DNA structure is a critical component of the second half of the insertion reaction. We have, for the first time, provided biochemical evidence for two key missing steps in the integration reaction: productive second-strand cleavage and second-strand synthesis.

Materials and Methods

Protein purification

R2Bm protein expression and purification were carried out as previously published (*16*). Briefly, BL21 cells containing the R2 expression plasmid were grown in LB broth and induced with IPTG. 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 SYPRO Orange (Sigma #S5692) staining of samples run on sodium dodecyl sulphate-polyacrylamide gel electrophoresis prior to addition of BSA for storage. All quantitations were done using FIJI software analysis of digital photographs (*17*).

Nucleic acid preparation

Oligos containing 28S R2 target DNA, non-target (non-specific) DNA, and R2 sequences were ordered from Sigma-Aldrich. The upstream (28Su) and downstream (28Sd) target DNA designations are relative to the R2 insertion dyad within the 28S rRNA gene. The oligo sequences are reported in Table 1.

Table 1: Sequences of DNA and RNA oligonucleotides used. ('Comp' stands for complementary strand)

Oligo Name	Sequence
b-strand	CCTCGAGGGATCCGTCCTAGCAAGCCGCTGCTACCGGAAGCTTCTGGACC
h-strand	GGTCCAGAAGCTTCCGGTAGCAGCGAGAGCGGTGGTTGAATTCCTCGACG
r-b strand	CGTCGAGGAATTCAACCACCGCTCTCGCTGCTACCGGAAGCTTCTGGACC
Pre-cleaved r-b	1) CGCTGCTACCGGAAGCTTCTGGACC 2) CGTCGAGGAATTCAACCACCGCTCT
r-strand	CGTCGAGGAATTCAACCACCGCTCTTCTCAACTGCAGTCTAGACTCGAGC
x-strand	GCTCGAGTCTAGACTGCAGTTGAGAGCTTGCTAGGACGGATCCCTCGAGG
h-x strand	GGTCCAGAAGCTTCCGGTAGCAGCGGCTTGCTAGGACGGATCCCTCGAGG
b_m-strand	CCTGCAGTGATCCGTCCTAGCAAGCCGCTGCTACCGGAAGCTTCTGGACC
r_m-strand	CGTCGAGGAATTCAACCACCGCTCTTCTCACCGATAAGTACGACTCGAGC
x_m-strand	GCTCGAGTCGTACTTATCGGTGAGAGCTTGCTAGGACGGATCACTGCAGG
Ns/28Sd 25 bp	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTCATCTAATT

Comp ns/28Sd 25 bp	AATTAGATGACGAGGCATTTGGCTACCTTAAGCTACCGGAAGCTTCTGGA
Pre-cleaved comp ns/28Sd 25 bp	1) AATTAGATGACGAGGCATTTGGCTA 2) CCTTAAGCTACCGGAAGCTTCTGGA
x_m-b strand	GCTCGAGTCGTACTTATCGGTGAGACGCTGCTACCGGAAGCTTCTGGACC
R2 3' DNA/ns	TGGCATGATGATCCGGCGATGAAAACCTTAAGCTACCGGAAGCTTCTGGA
Comp 28Sd 25 bp / Comp R2 3'DNA	AATTAGATGACGAGGCATTTGGCTATCTCACCGATAAGTACGACTCGAGC
R2 3' DNA 25	TGGCATGATGATCCGGCGATGAAAA
R2 3' RNA 25	UGGCAUGAUGAUCCGGCGAUGAAAA
Comp R2 5'DNA/ comp 28Sd 25 bp	AAATTAAAATTATGCGTATCGCCCCCCTTAAGCTACCGGAAGCTTCTGGA
R2 5'RNA 25 bp	GGGGCGAUACGCAUAAUUUUAAUUU
R2 3'-5' DNA	TGGCATGATGATCCGGCGATGAAAAGGGGCGATACGCATAATTTTAATTT
R2 5'DNA 25 bp	GGGGCGATACGCATAATTTTAATTT
Ns/28Sd 47 bp	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTCATCTAATTA GTGACGCGCATGAATGGATTA
Comp 28Sd 47 bp/comp R2 3' RNA	TAATCCATTCATGCGCGTCACTAATTAGATGACGAGGCATTTGGCTATTTTC ATCGCCGGATCATCATGCCA
28Su 73 bp/ns	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGG GAGTAACTATGACTCTCTTAAGGTAGGGTCCAGAAGCTTCCGGTAGCAGCG AGAGCGG
Comp ns/ comp R2 3' RNA	CCGCTCTCGCTGCTACCGGAAGCTTCTGGACCCTATTTTCATCGCCGGATC ATCATGCCA
Comp R2 5' RNA/ Comp 28Su 73 bp	AAATTAAAATTATGCGTATCGCCCCCCTTAAGAGAGTCATAGTTACTCCCGC CGTTTACCCGCGCTTGCTTGAATTTCTTCACGTTGACATTCAGAGC
28Su 73bp/28Sd 47bp	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGG GAGTAACTATGACTCTCTTAAGGTAGCCAAATGCCTCGTCATCTAATTAGTG ACGCGCATGAATGGATTA

All the linear DNAs were 50 bp in length. Each arm of most of the three-way and four-way junctions were 25 bp in length except for junctions tested for cDNA synthesis, for which the 28S DNA arm lengths were strategically varied to observe second-strand syntheses products. Diagrams of the constructs are provided in the main figures. Oligos with 28Sd sequence contained either 25 bp or 47 bp of post R2 insertion site 28S rDNA. Seven base pairs of upstream sequence were also included in these "downstream" oligos to span the insertion site. Oligos with 28Su sequence contained 72 bp prior to the insertion site as well as 5 bp of post R2 insertion site 28S rDNA. The largest oligo contained 72 bp of upstream and 47 bp of downstream 28S rDNA. Several oligos incorporated 25 bp of sequence complementary to either the 3' or the 5' RNA. Shorter oligos (25 bp) of sequence corresponding to the first and last 25 bp of R2Bm were also used in many of the constructs. The sequence for the x, h, b, and r strands of the nonspecific 4-way junction were obtained from Middleton et al (*18*). The constructs were formed by annealing the component oligos procedure: 20 pmole of the labeled oligo was mixed with 66 pmol of each cold oligo. The primers were annealed in SSC buffer (15 mM sodium citrate and 0.15 M sodium chloride) for 2 minutes at 95°C, followed by 10 minutes at 65°C, 10 minutes at 37°C and finally 10 minutes at room temperature. One of component

oligos had been 5' ^{32}P end labeled, prior to annealing the other component oligos. The annealed junctions were purified by polyacrylamide gel electrophoresis, eluted in gel elution buffer (0.3 M Sodium acetate, 0.05% SDS and 0.5 mM EDTA pH 8.0), chloroform extracted, ethanol precipitated, and resuspended in Tris-EDTA. Junctions that shared a common labeled oligo were equalized by counts DNA, otherwise equal volumes of purified constructs were generally used in R2 reactions. R2 3' PBM RNA (249 nt), 5' PBM RNA (320 nt), and a non-specific RNA (180 nt) were generated by in vitro transcription as previously published (*16*).

R2Bm reactions and analysis

R2 protein and target DNA binding and cleavage reactions were performed largely as previously reported (*12, 14*). Briefly, each DNA construct was tested for its ability to bind to R2 protein and to undergo DNA cleavage in the presence and absence of 5' PBM RNA, 3' PBM RNA, and non-specific RNA. All the reactions contained excess cold competitor DNA, dIdC. The reactions were loaded onto electrophoretic mobility shifting assays (EMSA) gels and companion denaturing gels for analysis. The ability to bind to branched and linear DNA was obtained from the EMSA gels and the ability to cleave DNA, as well as cleavage position, were obtained from the denaturing urea gels. A+G ladders were run alongside the reactions in the denaturing gels to aid in mapping cleavages. Second-strand synthesis assay was performed by the addition of dNTPs to the DNA cleavage reactions in the absence of RNA. 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 so that the most intense bands were dark gray. Adjusted TIFF files were quantified using FIJI (*17*).

Results

R2 protein binds preferentially to a nonspecific 4-way junction DNA over nonspecific linear DNA

Assuming common ancestry with archaeal Holliday junction resolvases, it is possible that the R2 PD-(D/E)XK RLE, or more generally the R2 protein, may exhibit DNA binding and cleavage activities normally associated with Holliday junction resolvases. Holliday junction resolvases recognize DNA structure rather than sequence. They bind to and symmetrically cleave 4-way DNA junctions (Holliday

junctions) and resolve those junctions into two linear DNAs. R2 protein was tested for the two biochemical activities of a Holliday junction resolvase: DNA binding and DNA cleavage of 4-way junctions.

The potentiality of R2 protein to recognize and bind to a 4-way DNA branched structure was tested by comparing the relative ability of R2 protein to bind to nonspecific linear and nonspecific 4-way junction DNA—individually and in competition (Figure 2).

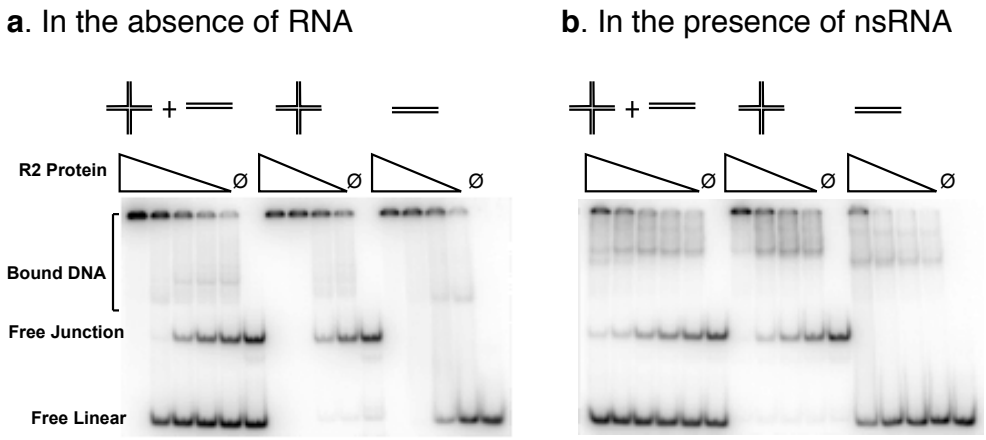


Figure 2: R2 binds preferentially to a 4-way junction. Competitive binding studies of a nonspecific 4-way junction DNA and nonspecific linear DNA analyzed by electrophoretic mobility shifting assays (EMSA). a) R2BM protein bound to nonspecific 4-way junction DNA and to nonspecific linear DNA, separately and in competition, in the absence of RNA. Triangles represent an R2Bm protein titration series. The DNA only (no R2Bm protein) lanes are marked with a Ø. The 4-way junction and linear DNAs shared a common DNA oligo. The shared DNA oligo was 5' end-labeled prior to formation and purification of the structures. b) Same as in a), except in the presence of nonspecific RNA (nsRNA).

The linear and junction DNAs were formed by annealing complementary oligos. The linear DNA and the junction DNA shared a common DNA strand that had been radioactively labeled prior to annealing. Sharing a common labeled DNA strand allowed radioactive decay counts to be a proxy for equalizing the DNA concentrations between the linear and junction DNAs and for similar sequence to be probed. In the absence of RNA (Figure 2a), the R2 protein bound to both nonspecific linear and nonspecific 4-way junction DNAs with roughly equal efficiency when individually examined across a protein concentration series. DNA binding was analyzed by EMSA. In competitive binding reactions, however, R2 protein had a clear preference for binding to the 4-way junction over the linear DNA. It should be noted that the junction DNA contained a greater number of total base pairs (100 bp; each arm being 25 bp) while the linear DNA was less (50 bp). It is unlikely, however, that the difference in DNA "length" had a significant effect on the observed binding affinity. The difference in binding affinity was much greater than two-fold; the R2 protein

did not bind to the linear DNA until most of the junction DNA had been bound. All the DNA binding reactions were carried out with an excess cold competitor DNA (dIdC).

The migration patterns for both linear and junction DNA were quite similar. A portion of the signal was stuck in the well with a smear that ran down from the well to faint protein-DNA complexes in the gel. The gel running protein-DNA complexes for the linear and junction DNAs migrated to roughly the same position within the gel. In the case of the linear DNA the smear continued from well all the way to the free DNA. The migration pattern, particularly that of R2 protein bound to junction DNA, was similar to that of R2 protein bound to its own target DNA in the absence of RNA prior to DNA cleavage (*19, 20*).

In the presence of nonspecific RNA (abbreviated as nsRNA, Figure 2b), R2 protein still bound preferentially to junction DNA as it had in the absence of RNA. Again, there was a smear running from the well to the major complex(es) in the gel. The junction and linear protein-RNA-DNA complexes migrated to similar but distinct positions within the gel. In the presence of R2 3' PBM RNA, R2 protein bound to junction DNA mostly as it did with nonspecific RNA and again 4-way junction DNA was preferred over non-specific linear DNA (Supplementary Figure 1). Interestingly, in the presence of 5' PBM RNA the behavior was different (see next section).

5' PBM RNA, but not 3' PBM RNA, is inhibitory to binding a nonspecific 4-way DNA junction

A direct comparison of R2 protein bound to 4-way junction DNA across a range of RNA concentrations (from ~18 pmole down to ~1 fmole) for nonspecific RNA, 3' PBM RNA, and 5' PBM RNA are reported in Figure 3.

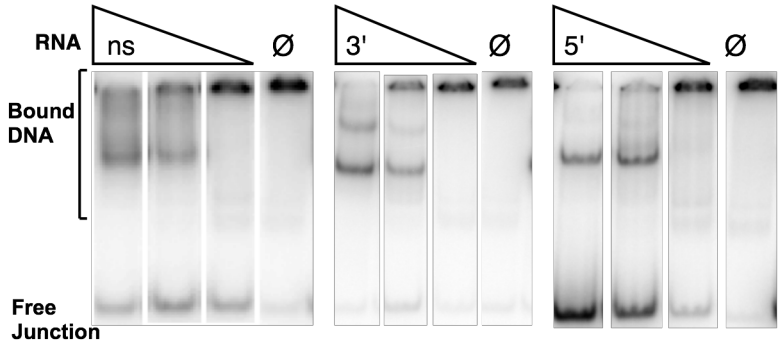


Figure 3. The R2 5' PBM RNA, but not 3' PBM RNA, is inhibitory to binding a nonspecific 4-way DNA junction. R2Bm protein bound to the nonspecific 4-way junction used in Figure 1, but in the presence of nsRNA, 3' PBM RNA, and 5' PBM RNA, were analyzed by EMSA. R2Bm protein was bound to 4-way junction DNA over

a large range of RNA concentrations (triangle) ranging from just over 18 pmole to just under 1 fmole. Only the 18 pmole of RNA, 675 fmole of RNA, 75 fmole of RNA, and a no RNA (Ø) lanes are shown here.

For each RNA titration set, the amount of protein used was sufficient to bind most of the junction DNA in the reaction that lacked RNA. In general, the addition of any of the three RNAs pulled material out of the well and into the gel migrating species. The R2 RNAs were more efficient at pulling material out of the well and into the gel. A similar phenomenon was observed when R2 protein is bound to its normal (linear) 28S target DNA in the presence of RNA (*20*). The addition of 5' PBM RNA, but not 3' PBM RNA or nonspecific RNA, greatly inhibited the binding of R2 protein to the 4-way junction DNA. Nonspecific RNA very minorly reduced binding to the 4-way junction and the reduction did not scale with RNA concentration. Binding to 3' PBM RNA also minorly reduced binding to the junction DNA, but only in the middle RNA concentrations. Only the presence of 5' PBM RNA greatly affected the binding of R2 protein to junction DNA. Inhibition scaled with increased 5' PBM RNA concentration. Binding to the (nonspecific) linear DNA was also similarly affected by the presence of 5' RNA. This inhibition is not observed when R2 protein is bound to its normal 28S target DNA (*14*).

R2 protein cleaves nonspecific linear and 4-way junction DNA at low levels and only in the absence of RNA

DNA from binding reactions in the presence and absence of RNA, similar to the high and no RNA reactions in Figure 3, were analyzed for DNA cleavage events by denaturing polyacrylamide gel electrophoresis with appropriate A+G ladders in order to map any DNA cleavages. Reactions with R2 protein bound to 4-way junction DNA (Figure 4a) and linear DNA (Figure 4b) substrates were analyzed independently for DNA cleavage. Each strand of the junction and linear DNA was sequentially radiolabeled on the 5' end and tracked in separate binding/cleavage reactions. A complicated pattern of random appearing cleavages occurred only in the absence of RNA. The cleavage locations were mapped with dots onto diagrammatic representations of the 4-way and linear DNA constructs used in the reactions (Figure 4c). The size of the dot is a rough indication of the relative efficiency of cleavage at that site. Changing the local sequence changed the location of cleavage sites (compare x/r to x_m/r_m in Figure 4c).

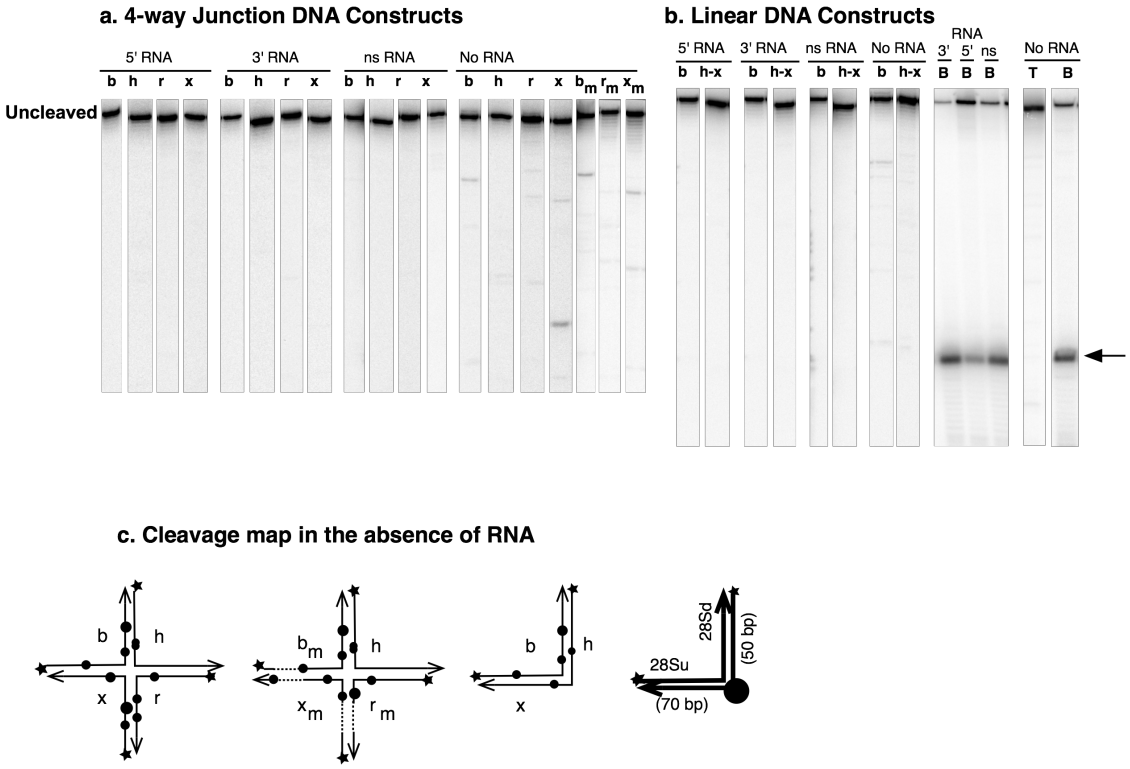


Figure 4: Non-specific 4-way junction exhibiting structure independent DNA cleavages. a) DNA cleavage reactions on non-specific 4-way junction DNA in the presence and absence of 5' PBM RNA, 3' PBM RNA, and nsRNA were analyzed by denaturing gel electrophoresis. The reactions were identical to the high and no RNA reactions analyzed by EMSA in Figure 3 except that each strand of the junction DNA was tracked for cleavage activity by sequentially labeling each of the four DNA strands. New junctions were built with local sequence modifications on the b, r and x strands, represented with a suffix m. b) DNA cleavage reactions on nonspecific linear DNA and on 28S rDNA in the presence and absence of 5' PBM RNA, 3' PBM RNA, and nsRNA were analyzed as above. Arrow represents cleavage at the R2 insertion site. c) Diagrams of 4-way junction and linear DNA constructs. Thin lines represent nonspecific DNA sequences and thick lines represent 28S rDNA sequences. Broken lines represent an area where local sequence was changed in the modified construct. See table 1 for sequence information. All arms are 25 bp unless otherwise noted. DNA cleavage locations and intensity are indicated by black circles with the diameter being a relative indication of cleavage efficiency at that site. The black star represents 5' 32P end-labeling. Other abbreviations: DNA sequence upstream of the R2 insertion site on 28S rDNA (28Su), DNA sequence downstream of the R2 insertion site on 28S rDNA (28Sd), top strand of 28S rDNA (T), bottom strand of 28S rDNA (B).

For a given sequence, the cleavages observed on linear DNA were like the cleavages observed on junction DNA. Thus, the observed DNA cleavages were not structure specific, but rather local sequence driven. Also, the cleavages on the junction DNA were not symmetrical and did not resolve the junction DNA into linear DNAs. The presence of any of the three RNAs (nonspecific, 3' PBM RNA, or 5' PBM RNA) abolished DNA cleavage. The lack of DNA cleavage in the presence of RNA was not due to RNA acting as a sink for Mg^{++} as R2 protein cleaved 28S target DNA very efficiently at the same RNA concentrations (see

Figure 4b). Additionally, reactions containing RNA but spiked with additional Mg^{++} did not cleave with any greater efficiency (data not shown).

Linear DNA and TPRT product are relatively poor substrates for second-strand cleavage

R2Bm inserts into a specific site in the 28S rDNA. In previous studies, it was determined two R2Bm protein subunits are involved in the integration reaction, one subunit bound upstream of the insertion site and one bound downstream of the insertion site (*12-14*). The downstream subunit provides the endonuclease involved in second-strand cleavage. For R2Bm, second-strand cleavage equates to cleavage of the "top" or sense strand of the 28S rRNA gene. It has always been tricky to get top-strand cleavage as it has required a narrow range of 5' PBM RNA, R2 protein, and DNA ratios. Too much or too little 5' PBM RNA can affect the top-strand cleavage. The data has been interpreted to indicate that first strand cleavage is required before the second-strand can be cleaved, that the downstream subunit must be bound to the DNA (which require 5' PBM RNA), and that the 5' PBM RNA must then dissociate from the downstream subunit for second-strand cleavage to occur. In vivo, with a full length R2 RNA, the process of TPRT would pull the 5' PBM RNA from the downstream subunit to initiate second-strand cleavage. TPRT is expected to generate a 3-way junction. In order to test various aspects of sequence and structure requirements for second-strand cleavage, linear and 3-way junctions were generated and tested in the presence and absence of downstream 28S target DNA, R2 sequences, and 5' PBM RNA.

5' PBM RNA in the reaction inhibited binding to nonspecific linear DNA (Figure 5a, EMSA gel with RNA titrated) and first-strand precleaved nonspecific DNA (Figure 5b, EMSA gel with RNA titrated). If 28S downstream sequence is added to the linear DNAs, binding is no longer inhibited in the presence of 5' PBM RNA (Figure 5c and 5d, EMSAs). The 28S DNA spanned from 7 bp upstream of the insertion site to 25 bp downstream of the insertion site. In cleavage reactions on linear DNA, regardless of the presence of downstream 28S DNA or the presence of a precleaved substrate, no cleavage activity was observed on the top-strand—the strand being tracked for second-strand cleavage (Figure 5a-5d, denaturing gels).

A nonspecific 3-way junction was also tested for binding in the presence/absence of 5' PBM RNA and for DNA cleavage in the absence of RNA. Like the nonspecific 4-way junction, R2 protein bound to a

nonspecific 3-way junction in the absence of RNA had minor cleavages (Figure 5e, denaturing gel). The presence of 5' PBM RNA inhibited binding of the R2 protein to the 3-way junction DNA (Figure 5e, EMSA).

A 3-way TPRT-like substrate based off the precleaved linear DNA containing downstream 28S DNA used above was generated (Figure 5g). The 3' end from the internal bottom-strand cleavage was covalently extended to be complementary to the final 25 bp of the 3' end of the R2 RNA. Annealed to the cDNA portion of the construct was either 25 bp of R2 RNA or a DNA version of the same 25 bp. The R2Bm protein was able to cleave the top-strand of these 28S DNA containing 3-way junctions at the R2 insertion site, albeit at a low level, in the absence of RNA (Figure 5g, denaturing gel). It did not matter if the R2 sequence containing arm was in the form of an RNA-DNA duplex or a DNA duplex. Interestingly, the efficiency of the top-strand cleavage at the insertion site was increased when the gap on the bottom-strand was eliminated (Figure 5f, denaturing gel). A fully covalently closed junction was a better substrate for 28S top-strand cleavage. The presence of 5' PBM RNA was not inhibitory to binding of the R2 protein when downstream 28S sequence was present (Figure 5f and 5g, EMSAs), but did eliminate DNA cleavage.

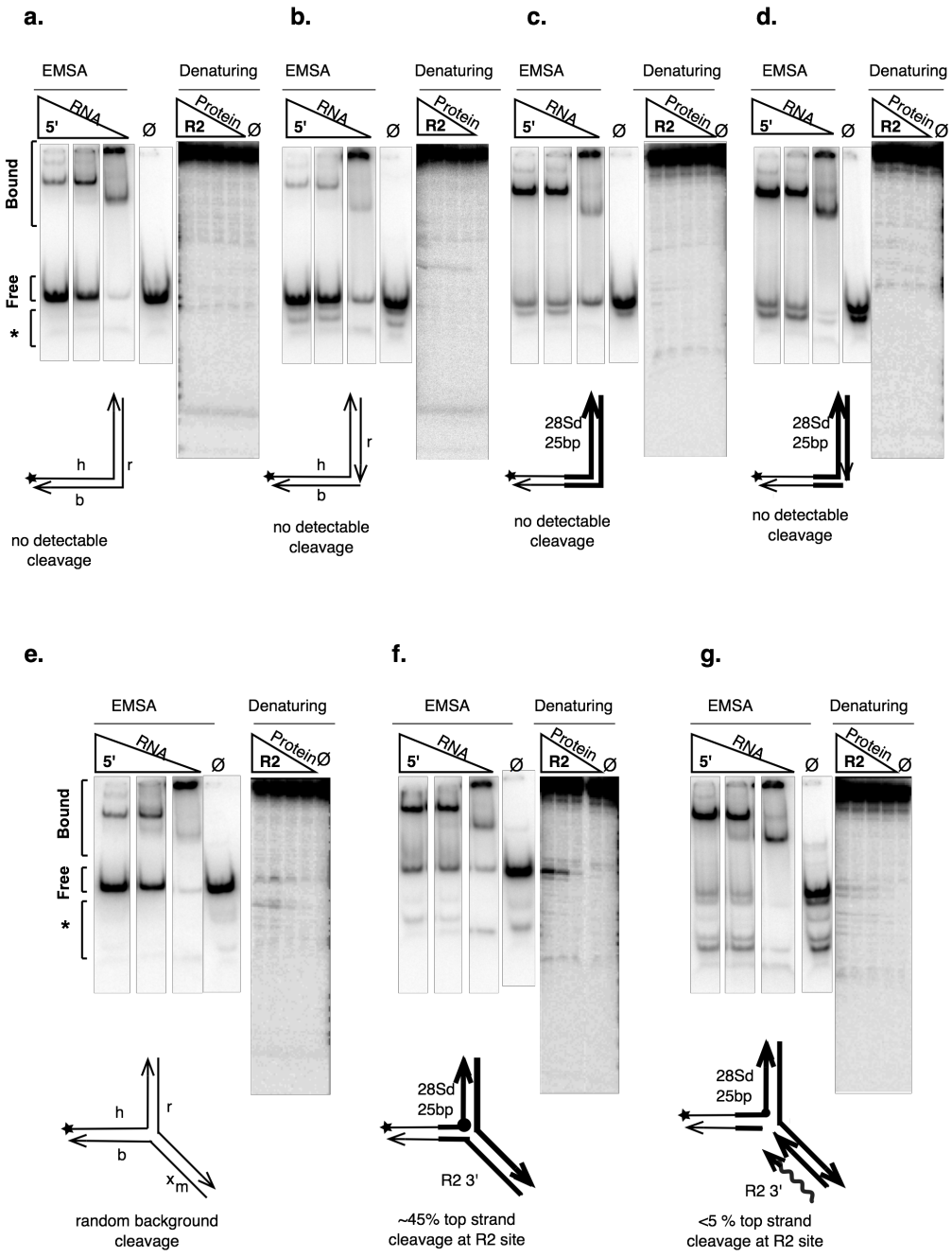


Figure 5. Linear DNA and TPRT products are relatively poor substrates for second strand cleavage. Panels a through g report the relative binding and cleavage activity of R2 protein bound to several linear and 3-way branched DNA constructs. A diagram of each construct is given. The diagram symbols are as in previous figures except that thick lines represent R2 sequences as well as 28S rDNA sequences. R2 RNA is a wavy line. The "arms" in each construct were 25 bp in length. The DNA binding reactions shown (see EMSAs) represent a titration series of 5' PBM RNA (triangle marked RNA) to no RNA (Ø). The amount of R2 protein used was just enough to shift most of the DNA in the absence of RNA. DNA cleavage reactions were performed in the absence of RNA and across a range of protein concentrations (triangle marked protein) to no protein (Ø). The asterisk in EMSA panels are partially formed junctions resulting from dissociation of one or more of the component oligos.

A 4-way junction that mimics a post-TPRT template jump is a substrate for second-strand cleavage

Existing genomic evidence indicates that, at least in some instances, template jumps occur during TPRT. It has been hypothesised that a template jump could occur at the end of first strand DNA synthesis from the RNA template to the upstream target DNA (*21, 22*). A 4-way junction was generated that would mimic a post TPRT template jump to the upstream 28S rDNA (Figure 6a). The 4-way junction generated was very similar to the downstream 28S-containing TPRT construct used in Figure 5g, except for the addition of a 25 bp RNA-DNA arm derived from the 5' end of the R2 RNA. The 5' R2 RNA-DNA arm was covalently joined to "upstream target DNA" arm as if a template jump has occurred. The upstream target DNA arm was nonspecific DNA except for 7 bp near the insertion site. Several variants of the 4-way junction depicted in Figure 6a were also generated. To test for the structural importance of the bottom-strand cleavage-remnant, the gap was covalently sealed in the 4-way junction in panel 6b. In order to test the relative importance of the 3' R2 RNA-DNA arm, the construct in panel 6c had the arm removed. The construct in panel 6d was similar, but consisted of a single stranded arm. The construct in 6d was the same construct that was used in Figure 5f and presented here for direct comparison purposes.

To varying degrees, each of the constructs tested in Figure 6 was also found to be a substrate for DNA cleavage in the absence of RNA. Finally, each construct, except perhaps 6d, cleaved almost exclusively at the R2 insertion site on the top-strand of the 28S rDNA. The 4-way junction that mimicked a post-TPRT template jump to the upstream 28S rDNA sequences exhibited the most robust second-strand DNA cleavage activity, edging out the activity observed on the covalently closed 3-way junction (Figure 6a, 6e/5f). Top-strand DNA cleavage was greatly reduced if the 4-way junction DNA was covalently closed instead of precleaved on the bottom 28S DNA strand (Figure 6b). Removal of the 3' R2 RNA-DNA arm from the junction also greatly reduced top-strand DNA cleavage (Figure 6c). Removing the RNA from the construct in panel 6c further reduced top-strand 28S DNA cleavage (Figure 6d). At this time, it is unclear whether the cleavage products are released or not upon cleavage (Figures 5f and 6f), due to the partially formed junctions resulting from dissociations are also running below intact free DNA.

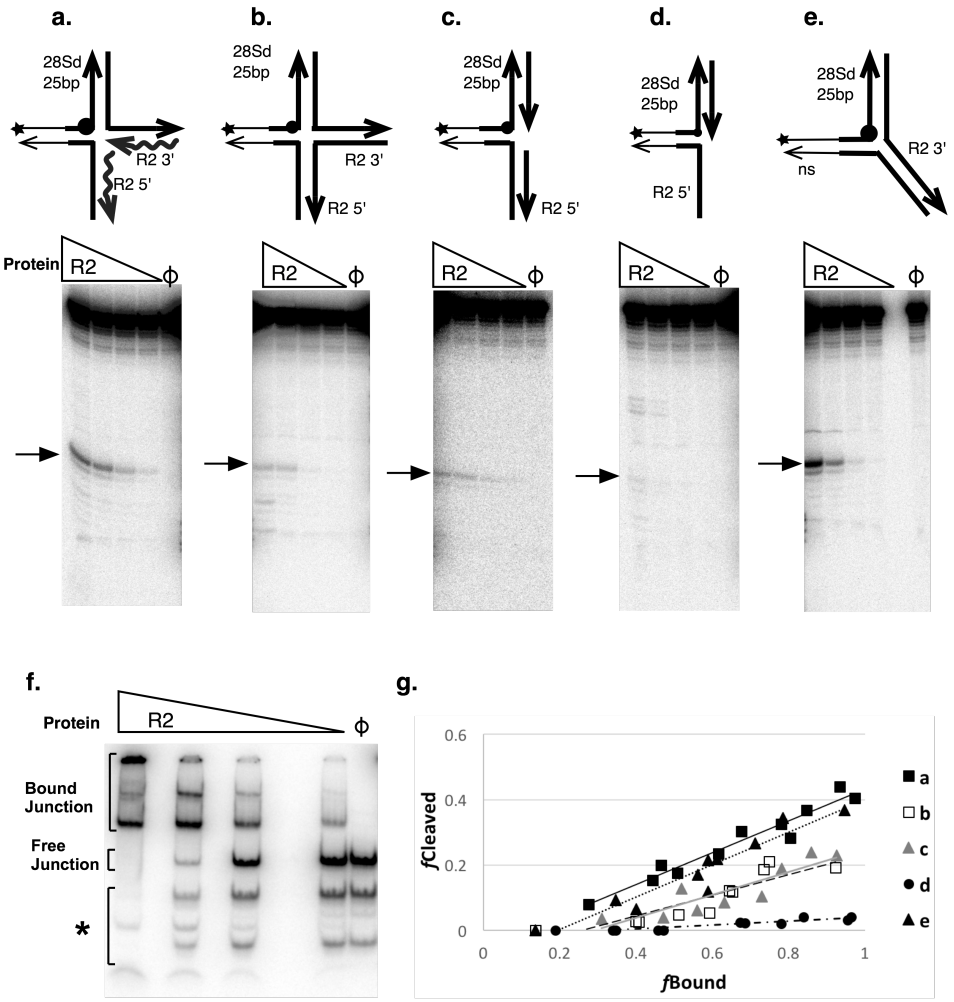


Figure 6: Second-strand cleavage at the R2-site on template-jumping related constructs. The construct tested in panel a is a construct that is expected to roughly mimic a bona fide integration intermediate, the post-TPRT template-jump intermediate. Panels b though d delete or change various parts of the construct in panel a. The diagram symbols are as in previous figures. The construct in panel e is the same construct used in Figure 5f. The arms of each construct are 25 bp. DNA cleavage reactions across a range of protein concentrations (triangle marked protein) were assayed by denaturing gel electrophoresis. The denaturing gel for each construct is presented. Panel f is the companion EMSA gel to the denaturing gel in panel a. The asterisk in panel f represents partially formed junctions resulting from dissociation of one or more of the component oligos. Panel g reports a graph of the fraction cleaved (*ƒ*Cleaved) as a function of the fraction of protein bound to DNA (*ƒ*Bound).

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

To further test for second-strand cleavage requirements, the construct used in Figure 6a was further modified. The first modification (Figure 7a) elongated the downstream 28S rDNA arm from 25 bp to 47 bp with additional 28S rDNA sequences. The other three arms of the junction remained 25 bp as before and the construct behaved identically. In the extended design, however, the cleaved 28S top-strand would, upon primer extension, generate a labeled strand of shorter overall length (50 bp) compared to the original

labeled strand length of 72 bp, assuming second-strand synthesis occurred. The construct used in Figure 7b was identical to 7a except that the two R2 arms lacked the RNA component of the RNA-DNA duplexes to simulate removal/degradation of the RNA template. The construct in Figure 7c and 7d replaced the original 25 bp nonspecific DNA arm with a 72 bp arm derived from upstream 28S rDNA sequences. The downstream 28S rDNA arm was largely replaced with nonspecific DNA in 7c and retained in 7d. Thus, the 7d construct contained the entire target site.

Each construct was subjected to DNA cleavage reactions across a range of R2Bm protein concentrations in the absence of RNA. The construct in 7a cleaved the top-strand of the 28S rDNA at the R2 insertion site the best followed by the constructs in 7b, 7c, and finally 7d. The lack of the RNA-DNA duplexes reduced cleavage efficiency per bound unit by about 50%. Surprisingly, the presence of the upstream DNA sequence also reduced DNA cleavage to under 50% per bound unit. The construct that had upstream 28S rDNA sequence but a truncated 28S rDNA downstream sequence (7c) was still able to cleave, but not as efficiently as when the full downstream sequence was present (7a).

These 4-way junctions, particularly the ones that cleaved well, are intriguing, not only because they are good substrates for top-strand cleavage, but also because these constructs, by their very nature are also potential substrates for second-strand synthesis. Upon top-strand DNA cleavage, a paired primer-template product is produced, ready to synthesize the second DNA strand of the inserting element.

In the presence of dNTPs, only construct 7a exhibited second-strand synthesis. A dNTP dependent band of the size expected (50 bp) was observed in the denaturing gel: 25 bp longer than the cleaved product and 22 bp shorter than the original labeled DNA strand. The background is high in these experiments due to the poor quality of the long oligos. The experiment will be repeated with gel purified DNA oligos that I purify instead of company gel purified oligos.

It is possible that construct 7b may have undergone second-strand synthesis, but that it was below detection level. It is also possible that either 7c and/or 7d may yet yield second-strand synthesis signal. Second-strand synthesis in these two constructs would generate a band that would run near the bottom of the dark smear of bands below the full-length oligo. The dark smear is the result of oligos that terminated prematurely during oligo synthesis that were not removed during gel purification of the oligos. The signal

above full length oligo in the presence of dNTPs results from the original full-length oligo being extended by R2. R2 can take almost any 3' end and extend it given a template in cis or in trans (*23*).

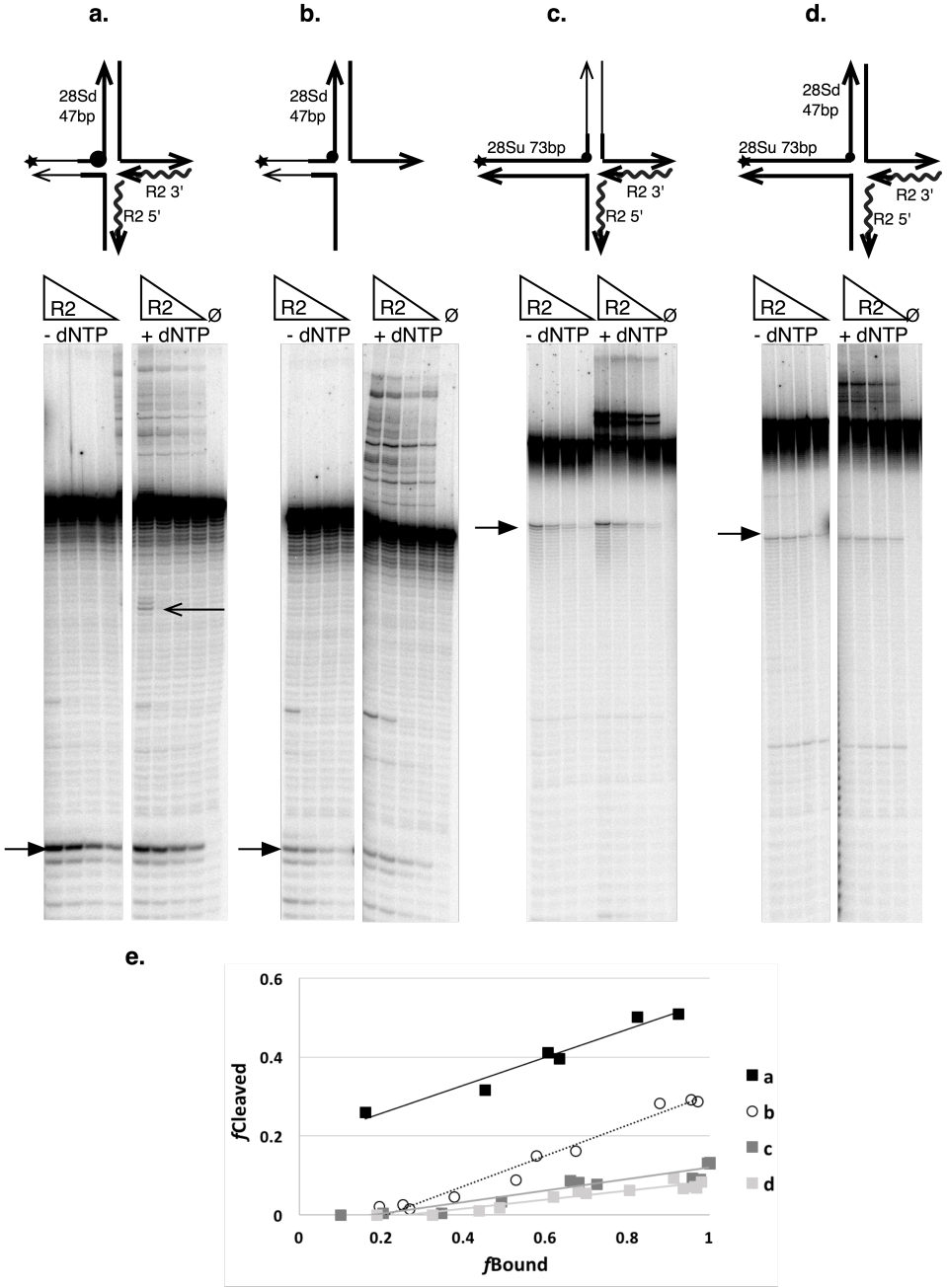


Figure 7: Second-strand cleavage and second-strand synthesis on template-jumping intermediate constructs. DNA cleavage reactions on various template-jumping derived constructs were tested for DNA cleavage (-dNTP) and second-strand synthesis (+dNTP) across a range of protein concentrations (triangle) and analyzed on denaturing gels (a-d). A diagram of each construct is given. Panel e reports a graph of the fraction cleaved as a function of the fraction bound for the -dNTP reaction set. Closed arrow indicates the R2 site cleavage and open arrow indicates the second strand cDNA. All other abbreviations and symbols are as in other figures.

Discussion

R2 protein was found to bind preferentially to nonspecific 4-way junctions over nonspecific linear DNA, but it was not able to subsequently resolve the junctions. Symmetrical DNA cleavages did not occur. The DNA cleavages that were observed were minor and were local sequence driven, rather than DNA structure driven. Therefore, R2 protein is not a Holliday junction resolvase, in the strictest sense. However, with the more specific 4-way junction containing 28S rDNA and R2 sequences, the top-strand 28S rDNA cleavage event was nearly symmetrical with the bottom-strand cleavage that had been engineered into the 4-way junction. This activity is very Holliday junction resolvase-like.

The fact that the presence of 5' PBM RNA inhibited R2 binding to branched DNA structures, but not if one of the branches contained downstream 28S rDNA sequences is interesting. In the presence of 5' PBM RNA, the R2 protein is probably binding to the downstream 28S rDNA sequence, when present, as it would on linear 28S rDNA in historical studies (*13, 14*). The new experiments indicate that the part of the R2 protein that recognizes branched DNA structure becomes masked by the association with the 5' PBM RNA. It is not yet understood what part of the R2 protein recognizes 4-way junctions. It is possible that the endonuclease is responsible. The conformation that the R2 protein adopts when associated with the 5' PBM RNA is such that it uses the ZF and Myb domains to bind to DNA sequences downstream of the insertion site on linear DNA; the DNA endonuclease is not active in this conformation (*12, 14*). The 5' PBM RNA could be sterically hindering the endonuclease, perhaps by competitive binding to a site that overlaps RLE R-box, or by allosteric interactions that sequester or otherwise inactivate the endonuclease.

The 4-way junction with the R2 derived arms, a full downstream 28S rDNA arm, a nonspecific upstream arm, and a template-jump from the 5' R2-arm to the 28S DNA arm cleaved the best. The presence of the template jump appeared to be one of the most important structural components for a second-strand cleavage substrate to have. Both 3-way and 4-way junctions that had this aspect were generally good cleavers. The junction missing the R2 3' arm, but retaining the RNA-DNA duplexed 5' arm cleaved, but not well. A variation of that same construct minus the RNA component, forming a linear DNA with a single stranded template-jump derived flap, failed to cleave. Thus the 3' arm is important so is the 5' arm duplex. That said, a "4-way" junction with the RNA components removed from both the R2 5' and 3' arms cleaved moderately well, so having an unpaired R2 5' arm (or 3' arm) is not catastrophic, substrate wise, in of itself.

The presence of a preexisting bottom-strand cleavage also appeared to be an important structural component. The break in the DNA backbone must be in the correct position relative to the template jump arm (the 5' R2-arm) as sealing the break reduces second strand cleavage. The 3-way TPRT-like product (no template jump) does not cleave well. The R2 protein appears to keep track of the overall structure as well as the specific structure of certain arms and the sequence of those arms.

The 4-way junctions with primarily upstream 28S DNA overlapped the primarily downstream 28S DNA containing junctions in the region spanning from 7 bp upstream to 5 bp downstream of the R2 insertion site. The 28S rDNA sequence within this overlap appeared to be at least somewhat important as both the primarily upstream and primarily downstream constructs were able to cleave the top-strand, to some degree. The presence of the full downstream 28S sequences greatly improved cleavability, but not if the full upstream 28S DNA sequence were also present. The presence of a full target site was inhibitory. There are several possible reasons for the reduced cleavage: 1) the upstream arm is competing for R2 protein subunits, 2) the presence of the upstream DNA causes protein subunits to adopt suboptimal position or conformation to cleave, 3) steric clash between subunits bound to both arms, and 4) the junction may need gap or a gap+flap on the upstream 28S rDNA arm (see Figure 8). Higher resolution experiments will need to be performed.

The 4-way junction that cleaved the best was also the only construct tested that had demonstrable second-strand synthesis activity in the presence of dNTPs. Putting all the data together leads to a new and deeper understanding of the second half of the insertion reaction and allows more detailed model to be put forth (Figure 8). The first half of the integration reaction is identical to steps 1 and 2 in Figure 1. After TPRT, however, the new model proposes a template-jump from the 5' end of the R2 RNA to the top-strand of the 28S rDNA upstream of the R2 insertion site. It is this step that, to date, does not efficiently occur in vitro. The reverse transcriptase of the upstream subunit is presumably the reverse transcriptase involved in the template jump occurs although it remains possible that the downstream subunit took over from the upstream subunit at some point during TPRT. The template jump would form a 5' flap and a 4-way junction. The 4-way junction is cleaved by either the downstream subunit or by the downstream subunit in the context of a dimer. Top-strand cleavage is followed by second-strand DNA-synthesis primed from the free 3'-OH liberated from the top-strand cleavage event. As the second strand synthesis signal observed was only at

the highest protein concentrations, it indicates that either our constructs are not yet fully optimized for efficient polymerization post top-strand cleavage, or that the upstream subunit performs second-strand synthesis in addition to TPRT. Future experiments will be needed to resolve these new questions.

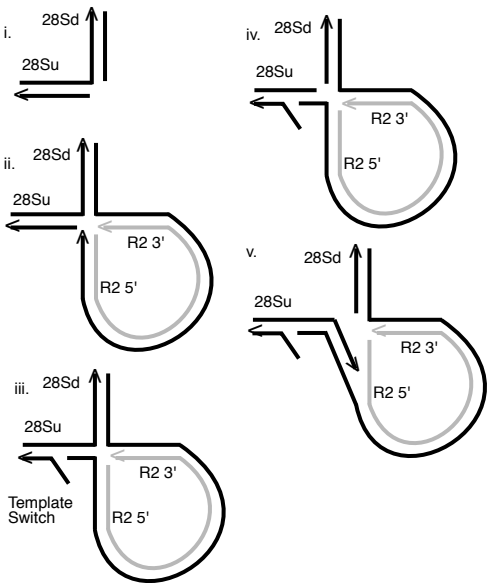
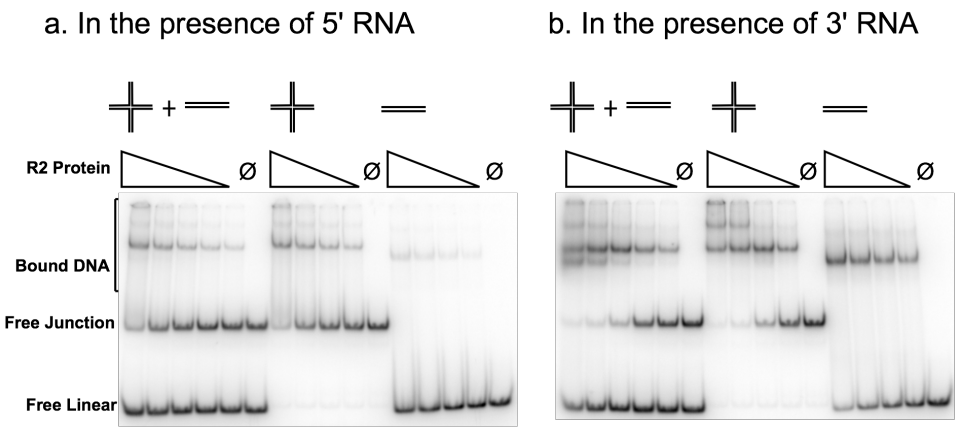


Figure 8. A new model of integration. The initial steps of the integration (i, ii) are as in Figure 1. The new integration intermediates to complete the integration reaction are presented (iii-v). Intermediate iii depicts template switching at or near the gap generated by first strand cleavage. Intermediate iv depicts top strand cleavage and the priming of second strand synthesis. Finally, the intermediate v shows the second strand synthesis using the correct template sequence and the RNA strand being displaced as the new cDNA advance. Black lines are DNA. Gray lines are RNA. Abbreviations: 28Su (28S rDNA just upstream of the R2 insertion site), and 28Sd (28S rDNA just downstream of the R2 insertion site).

Finally, the fact that R2 binds to 4-way junctions in the presence of 3' PBM RNA and in the absence of RNA leaves open the question as to whether R2, or by extension other RLE LINEs might target 4-way junctions for insertion, as part of their site-specificity, perhaps binding to cruciform structures that may form at larger inverted repeats. There is an imperfect inverted repeat near the R2 insertion site. Further studies focusing on the first strand cleavage and TPRT with cruciform structures are required to address this interesting question. The current study focused only on the second half of the reaction. The fact that, thus far, DNA cleavages were abolished in the presence of RNA, for both nonspecific and for target DNA containing 4-way junctions, would seem to be preliminary evidence that, at least in the case of R2, RLE LINEs may not target cruciform structures for the initial stages of integration. If not, DNA structure may yet

be crucial as the R2 insertion site is calculated to have an intrinsic bend to the DNA (Christensen, unpublished data) which likely phased a nucleosome and perhaps directs R2 binding.

Supplemental Files section:



Supplementary Figure 1: R2 binds preferentially to a 4-way junction. Competitive binding studies of a nonspecific 4-way junction DNA and nonspecific linear DNA analyzed by electrophoretic mobility shifting assays (EMSA). a) R2BM protein bound to nonspecific 4-way junction DNA and to nonspecific linear DNA, separately and in competition, in the presence of 5' RNA. Triangles represent an R2Bm protein titration series. The DNA only (no R2Bm protein) lanes are marked with a Ø. The 4-way junction and linear DNAs shared a common DNA oligo. The shared DNA oligo was 5' end-labeled prior to formation and purification of the structures. b) Same as in a), except in the presence of 3' RNA.

References

1. D. D. Luan, M. H. Korman, J. L. Jakubczak, T. H. Eickbush, 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 (1993).
2. G. J. Cost, Q. Feng, A. Jacquier, J. D. Boeke, Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910 (2002).
3. J. V. Moran *et al.*, High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).
4. W. D. Burke, H. S. Malik, J. P. Jones, T. H. Eickbush, The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol Biol Evol* **16**, 502-511 (1999).
5. J. Yang, H. S. Malik, T. H. Eickbush, Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 7847-7852 (1999).
6. K. K. Kojima, H. Fujiwara, Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol Biol Evol* **20**, 351-361 (2003).

7. H. S. Malik, W. D. Burke, T. H. Eickbush, The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**, 793-805 (1999).
8. J. L. Jakubczak, Y. Xiong, T. H. Eickbush, Type I (R1) and type II (R2) ribosomal DNA insertions of Drosophila melanogaster are retrotransposable elements closely related to those of Bombyx mori. *J Mol Biol* **212**, 37-52 (1990).
9. K. K. Kojima, K. Kuma, H. Toh, H. Fujiwara, Identification of rDNA-specific non-LTR retrotransposons in Cnidaria. *Mol Biol Evol* **23**, 1984-1993 (2006).
10. E. A. Gladyshev, I. R. Arkhipova, Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* **448**, 145-150 (2009).
11. D. G. Eickbush, D. D. Luan, T. H. Eickbush, Integration of Bombyx mori R2 sequences into the 28S ribosomal RNA genes of Drosophila melanogaster. *Molecular and cellular biology* **20**, 213-223 (2000).
12. S. M. Christensen, T. H. Eickbush, 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 (2005).
13. S. M. Christensen, A. Bibillo, T. H. Eickbush, Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468 (2005).
14. S. M. Christensen, J. Ye, T. H. Eickbush, 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 (2006).
15. A. Govindaraju, J. D. Cortez, B. Reveal, S. M. Christensen, Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic acids research*, (2016).
16. A. Govindaraju, J. D. Cortez, B. Reveal, S. M. Christensen, Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287 (2016).
17. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).
18. C. L. Middleton, J. L. Parker, D. J. Richard, M. F. White, C. S. Bond, Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res* **32**, 5442-5451 (2004).
19. S. Christensen, T. H. Eickbush, Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* **336**, 1035-1045 (2004).
20. J. Yang, T. H. Eickbush, RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol Cell Biol* **18**, 3455-3465 (1998).
21. A. Bibillo, T. H. Eickbush, The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol* **316**, 459-473 (2002).
22. A. Bibillo, T. H. Eickbush, End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* **279**, 14945-14953 (2004).
23. A. Kurzynska-Kokorniak, V. K. Jamburuthugoda, A. Bibillo, T. H. Eickbush, DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *Journal of molecular biology* **374**, 322-333 (2007).

CHAPTER 4

Conclusions

LINE elements replicate by reverse transcribing their RNA at element generated chromosomal nick, a process termed **T**arget-**P**rimed **R**everse **T**ranscription (TPRT) because the free 3' OH at the cleaved target site primes the reverse transcription (1). The cleavage and reverse transcription processes are performed by multifunctional protein encoded by the retrotransposons themselves. While initial steps of integration including first (bottom) strand cleavage and TPRT have been characterized, many aspects of the integration mechanism are not yet completely understood, including RNA-protein particle (RNP) formation, DNA target recognition, and DNA cleavage site choice. Bioinformatics and *in vivo* transposition studies have been of great help, but both lack the detail and control of *in vitro* based biochemical studies.

The more ancient LINE clades, which include R2, R4, NeSL and CRE-related elements, encode a $PDX_{12\text{-}14}(D/E)$ motif that is part of the catalytic core of the element's DNA endonuclease (2). The catalytic motif is located downstream of the reverse transcriptase. The boundaries of the endonuclease domain, however, have not been determined before. The PD-(D/E) motif and thus the endonuclease itself appear to be a member of larger endonuclease family, the PD-(D/E)XK endonucleases. The first PD-(D/E)XK DNA endonucleases to be identified were the type II restriction enzymes (e.g. EcoRI, BamHI, and FokI). Now, however, the PD-(D/E)XK superfamily comprises a large group of diverse proteins that are involved in various functions including, but not limited to, DNA repair (3, 4), Holliday junction resolution (5-7), and RNA processing (8, 9). The PD-(D/E)XK superfamily proteins share little to no sequence homology, however they all share a structurally conserved core. The consensus core consists of a four-stranded mixed β-sheet flanked by α-helix on each side (10).

A crystallographic structure of R2Bm protein is not yet available. Blast searches and protein threading algorithms have been used to model the R2 endonuclease. The most recent model of the R2Bm endonuclease was generated that relied heavily on the crystal structure of Holliday junction resolvase C (Hjc) and to a lesser extent the restriction endonuclease FokI (as well as a few other lesser known proteins with solved structures) (see chapter 2 figure 6 and (11, 12)). The structural model of R2 endonuclease presented in chapter 2 Figure 6 indicates that the RH, RE, KRNK, and KY residues are on the same face protein, near the endonuclease active site residues. The active site residues appear to have the proper geometry: a good indication that the model may be valid. Our mutational studies revealed the catalytic

lysine residue (K of PD-(D/E)XK) in R2 being located in α-helix 2 instead of the canonical position in β-strand 3. It is notable that the R2 endonuclease domain share the DNA binding 'R-box' sequence and location with some of the restriction endonucleases in the beginning of the domain structure and shares DNA binding basic residues with some of the Holliday junction resolvases towards the end of the structure (12).

The sequence being cleaved by the R2 endonuclease is different on the top and bottom strands. It is unclear how two different specific cleavages can be made by a single endonuclease. Second (i.e., "top strand") target DNA cleavage has been over particular and is not as robust as observed first strand cleavage. It can only be scantily achieved with a narrow amount protein, DNA and RNA ratio. Even then, second strand DNA synthesis does not occur *in vitro* following the top strand cleavage. Several possible reasons could account for the relative inefficiency of second strand cleavage in our *in vitro* reactions, including: (1) it is an intrinsically inefficient reaction; (2) a host factor is required; or (3) a step before second strand cleavage in not reached in our reactions i.e. the correct DNA structure has not been formed. To address the third possibility, several branched DNA structure-sequence substrates for second strand cleavage have been tested. It is possible that second-strand cleavage and synthesis might depend on structural aspect of the substrate, as TPRT intermediate product forms a three-way branched structure and second strand synthesis transiently forms a four-way junction.

Holliday junction resolvases are DNA structure selective endonucleases that cleave Holliday junctions (four-way branched DNA junctions) formed during homologous recombination events to generate two resolved dsDNAs. The fact that the R2 endonuclease is being modeled, in large part, based on the structure of a Holliday junction resolvase begs the question if R2 itself has/retains Holliday junction binding and cleavage activities. The similarity of R2 endonuclease catalytic core to that of Holliday junction resolvases might provide the key to understanding and investigating second strand DNA cleavage as three-way and four-way branched DNA structures form during the R2 integration reaction.

For the first time R2Bm is shown to bind preferentially to a four-way synthetic junction with non-R2 sequences over a linear DNA with non-R2 sequences, but it doesn't cleave the branched structure symmetrically, the way Holliday junction resolvases resolve. Nevertheless, it was intriguing that R2 can bind branched structure and it is possible to probe second strand cleavage and synthesis if we allow R2 to

react with an appropriate branched nucleic acid structure with sequences from R2 RNA, TPRT product and R2 target DNA itself. Several junctions were tested with different combination of structure and sequences. Out of those, a covalently closed 3-way junction with downstream R2 target sequence in top strand and R2 TPRT sequences was the one to be cleaved most at R2 insertion site. Currently, the most interesting of these new constructs in terms of second strand cleavage is the possible intermediate that would result from template switching from the RNA post TPRT to the target DNA. DNA construct that mimics this template switching event is a four-way branched DNA structure, with a nick at 3' RNA priming and the first cDNA extended to base pair with upstream insertion site sequences. This construct contains downstream insertion site sequences in addition to 3' and 5' RNA sequences and resulted in the most successful top-strand cleavage. Interestingly, following second strand cleavage on this intermediate with a high protein concentration, second strand synthesis was identified in the presence of dNTPs.

The findings from my dissertation research lead to a new updated model of R2 integration addressing the possibility of target DNA intermediate that can support better second strand cleavage and second strand synthesis. Characterization of second strand cleavage and synthesis is essential for the better understanding of the replication mechanism. This study will resolve some of the lingering questions in the mechanism of integration, not only for RLE encoding LINEs but also across APE LINEs, as they are functionally expected to have similar integration mechanism.

Limitations

Non-covalently closed junctions tend to melt during reactions, which appears to form a trail of low molecular weight bands below the fully annealed unbound or free junction in EMSA gels. It is imperative to take these "partial" junctions into consideration and understand their formation during reactions so that the distinction between partial junctions and released cleavage products if any can be made unambiguously. It is unclear yet whether the cleaved products are released or still bound by R2 protein.

Our current knowledge is inadequate to answer why top-strand cleavage and synthesis is inhibitory in the presence of upstream insertion site sequence, especially, when both upstream and downstream sequences are present in the same junction that mimics template switch. Chapter 3 has discussed the following reasoning: 1) the upstream arm is competing for R2 protein subunits, 2) the presence of the

upstream DNA causes protein subunits to adopt suboptimal position or conformation to cleave, 3) steric clash between subunits bound to both arms, and 4) the junction may need gap or a gap+flap on the upstream 28S rDNA arm (see Figure 8).

The oligonucleotides being used for second strand cDNA synthesis is not pure as we observed several lower molecular weight bands appearing below the full length oligos in denaturing urea gels. Although oligos have been purchased as PAGE purified, the level of the background is high enough to undermine the signal for second strand cleavage and synthesis.

Finally, it is not known what part of the protein is recognizing the branched DNA structure. The presence of 5'RNA is inhibitory to binding branched DNA structure with non-specific sequences, which could possibly imply that the domain that sees the structure of the DNA is also the domain that interacts with 5' PBM RNA; however, this inhibition is reversed by having target downstream sequence in the structure. R2 protein when associated with the 5' PBM RNA adopts a conformation such that it uses the ZF and Myb domains to bind to DNA sequences downstream of the insertion site. If the endonuclease is the domain responsible for binding DNA junctions, the 5' PBM RNA could be sterically hindering the endonuclease, perhaps by competitive binding to a site that overlaps RLE R-box, or by allosteric interactions that sequester or otherwise inactivate the endonuclease.

Future directions

The immediate follow-up for chapter 3 is to work on the limitations mentioned above. First and foremost, the reactions for second strand cDNA synthesis will be repeated for all the candidate junctions mentioned in Chapter 3, Figure 7, but with the oligos that will be PAGE purified more stringently in our lab rather than buying PAGE purified oligos. Additionally, one more junction with an inclusion of strand displacement as the result of template jump (see Chapter 3, Figure 8iii) will be tested for second strand cleavage and synthesis. This new junction will be built strategically to have a nick between the cDNA region that pairs with upstream target DNA sequence and the displaced strand. Presence of this nick could possibly make the junction more flexible upon R2 binding especially at the upstream sequence whose presence was in fact discouraging the second strand cleavage and synthesis in the earlier constructs.

Identification of partially formed junctions will be carried out by cutting each of those individual bands that runs below the free junction DNA and analyzing them on denaturing gel. This will help us differentiate cleaved DNA from partially formed junction DNA. A series of junctions will also be built strategically by excluding the pairing oligo components one by one and run it in the EMSA gel to map the running distance of partial junctions.

Finally, to identify the regions of protein that contacts the DNA branched structure, mass spectrometry analysis will be undertaken. Biotinylation of R2 protein and R2 protein bound to a DNA junction will reveal the responsible domains. R2Bm endonuclease domain alone will be cloned and expressed, followed by mutational and functional analysis to directly find its role in binding junctions, especially its role in structure recognition. Additionally, DNA foot printing studies will be performed with an emphasis on how the presence/absence of target R2 sequence will affect protein binding and what part of the junction is contacted by protein.

The function of bent DNA on R2 integration, particularly with respect to first strand cleavage will be examined. The intrinsic bend might facilitate nucleosome phasing at R2 insertion site at 28S rDNA and it would be interesting to find whether R2 integration is dependent by any form on this DNA bend. R9, a member of R2-A subclade targets 28S rDNA, in a location different than that of R2 (13). It is suggested that the members of R2-A and R2-D subclades use different targeting mechanisms (14, 15). It would be interesting to clone and express R9 element and test its activity on branched DNA substrates.

Last, but not least the experimental findings from in vitro analysis of R2 will be used as a foundation for understanding LINE-1 integration in human genome. By understanding their respective endonuclease and targeting mechanisms, we can learn their propagation in terms of evolution as well as pathogenicity.

References

1. Luan DD, Korman MH, Jakubczak JL, & Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72(4):595-605.

2. Yang J, Malik HS, & Eickbush TH (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(14):7847-7852.
3. Ban C & Yang W (1998) Structural basis for MutH activation in E.coli mismatch repair and relationship of MutH to restriction endonucleases. *EMBO J* 17(5):1526-1534.
4. Tsutakawa SE, Jingami H, & 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(6):615-623.
5. Hadden JM, Convery MA, Declais AC, Lilley DM, & Phillips SE (2001) Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I. *Nat Struct Biol* 8(1):62-67.
6. Nishino T, Komori K, Tsuchiya D, Ishino Y, & Morikawa K (2001) Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition. *Structure* 9(3):197-204.
7. Middleton CL, Parker JL, Richard DJ, White MF, & Bond CS (2004) Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res* 32(18):5442-5451.
8. Dias A*, et al.* (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* 458(7240):914-918.
9. Yuan P*, et al.* (2009) Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature* 458(7240):909-913.
10. Feder M & Bujnicki JM (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.
11. Mukha DV, Pasyukova EG, Kapelinskaya TV, & Kagramanova AS (2013) Endonuclease domain of the Drosophila melanogaster R2 non-LTR retrotransposon and related retroelements: a new model for transposition. *Front Genet* 4:63.
12. Govindaraju A, Cortez JD, Reveal B, & Christensen SM (2016) Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* 44(7):3276-3287.
13. Gladyshev EA & Arkhipova IR (2009) Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* 448(2):145-150.
14. Shivram H, Cawley D, & Christensen SM (2011) Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements* 1(3):169-178.
15. Thompson BK & Christensen SM (2011) Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: Plasticity of integration mechanism. *Mob Genet Elements* 1(1):29-37.