INSERTION MECHANISM OF LONG INTERSPERSED ELEMENTS (LINEs) AND THE ROLE OF THE LINKER DOMAIN

by

MONIKA PRADHAN

DISSERTATION

Submitted in partial fulfillment of the requirements

For the degree of Doctor of Philosophy at

The University of Texas at Arlington

August, 2018

Arlington, Texas

Supervising Committee:

Dr. Shawn M. Christensen, Supervising Professor

Dr. Esther Betran

Dr. Todd Castoe

Dr. Matthew Fujita

Dr. Mark Pellegrino

Abstract

INSERTION MECHANISM OF LONG INTERSPERSED ELEMENTS (LINEs) AND THE ROLE OF THE LINKER DOMAIN

Monika Pradhan, PhD

The University of Texas at Arlington, 2018

Supervising Professor: Shawn M. Christensen

Long Interspersed Elements (LINEs), also known as non-Long Terminal Repeat (non-LTR) retrotransposons are major group of transposable elements ubiquitous in eukaryotic genomes and are known to altogether influence the structure and function of the host genome. LINEs encode a multifunctional protein that reverse transcribes its mRNA to DNA at the insertion site by a process called Target Primed Reverse Transcription (TPRT) which involves two half reactions involving DNA cleavage followed by DNA synthesis. TPRT is the first half of the integration reaction. The second half of LINE integration, second strand cleavage and second strand synthesis, has remained poorly understood. Also, poorly understood is the role of the nearly universally conserved linker domain of the LINE encoded protein. The unknown aspects of the integration mechanism and the protein domains were studied *in vitro* by using a site-specific R2 LINE from *Bombyx mori* (R2Bm). A Holliday junction-like 4-way target DNA structure was identified to be an essential integration intermediate and the gateway into the second half of the integration reaction. The 4-way junction cleaved at the proper second-strand cleavage site and upon cleavage created a primer-template that led to second-strand DNA synthesis. The Linker region of the R2Bm protein was found to be important for recognizing the 4-way junction and for positioning the DNA relative to the endonuclease and the reverse transcriptase active sites. The linker region is located just after the reverse transcriptase and harbors a conserved set of predicted α-helices, thought to be an α-finger, followed by gag-like zinc knuckle. In addition to binding and positioning the 4-way junction, the α-finger residues were found to control target DNA cleavages and new strand synthesis at every step of the integration mechanism. The zinc knuckle residues also showed similar function but were more prominent for the second half of integration. Finally, the role of specific residues in the N-terminal RT-1 region and the endonuclease R-box region were also explored. A unique residue has been identified in the RT-1 region that is able to distinguish between 3’ PBM RNA and 5’ PBM RNA. The endonuclease R-box region appeared to be important for second strand DNA cleavage.

Dedication

To my parents for their unconditional love and support. I appreciate their sacrifices and
I wouldn’t have been able to pursue a doctorate degree
without their belief in me.

To my husband, Dr. Surendra for his unending support, level-headedness, love, and humor.
I can hardly describe his compromises and relentless efforts to
keep my motivation high through the times of
perseverance and hardships.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Shawn M. Christensen, for being a tremendous mentor and a constant source of guidance and encouragement throughout my Ph.D. journey. His impressive scientific knowledge has helped me through many roadblocks. I could not have imagined having a better advisor and a mentor. Thank you, Dr. Shawn, for everything. You are the best!

I would also like to give special thanks to my committee members, Dr. Esther Betran, Dr. Todd Castoe, Dr. Matthew Fujita, and Dr. Trey Fondon. I am also thankful to Dr. Mark Pellegrino for agreeing to be in my committee this year. Their constructive suggestions and insightful comments have helped me widen my research perspective, for which I have the utmost appreciation.

It has been such a pleasure to work in a research group of remarkable people. I am thankful to Dr. Aruna Govindaraju, Dr. Murshida Madbub, Eyad Shihabeddin, Brijesh Khadgi, Dr. Kim Bowles, Jeremy Cortez, Santosh Dhamala, Neeti Maan, Cora Hernandez, Zainab Adeola Omisade, and all previous and current undergraduates for all their help and support during this journey.

I am grateful to Rachel Wostl, my teaching lab coordinator for her support and for making my teaching experience a positive one. I would like to take this opportunity to thank all administrative staffs of Biology Department for their prompt help whenever needed. Thanks are also due to UTA Graduate School, National Science Foundation, and Phi Sigma Biology Graduate Student Society for their financial support.

Thank you very much, everyone!

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Long INterspersed Elements (LINEs)

LINEs, also known as non-Long Terminal Repeat (non-LTR) retrotransposons are an important group of retrotransposable elements. The major clades of LINEs date back at least to the Pre-Cambrian era, approximately 600 million years ago, putting the origin of LINEs to around the origin of eukaryotes [1,2]. LINEs get transcribed into mRNA which then gets translated into a multifunctional protein with DNA endonuclease, reverse transcriptase, and nucleic acid binding activity. The endonuclease is used to nick the target DNA. The liberated chromosomal 3'-OH is used by the reverse transcriptase to prime cDNA synthesis in a process called Target Primed Reverse Transcription (TPRT) [1,3]. Short Interspersed Elements (SINEs) are a type of non-autonomous retrotransposons that are known as parasites of LINEs. They do not encode their own protein but highjack the protein machinery of LINEs for their reverse transcription and integration into a new genomic site [4,5].

1.2 Groups of LINEs

Based on structural features and phylogeny of reverse transcriptase, LINEs are classified into two major groups [6]. The early branching group is characterized by single open reading frame (ORF) encoding restriction-like endonuclease (RLE) (Figure 1-1), and most clades under this group insert into a specific site during integration (reviewed in [7–9]). The late branching group is hallmarked with two ORFs, the second of which encodes Apurinic-Apyrimidinic like endonuclease (APE) (Figure 1-2) (reviewed in [9–11]). Majority of APE bearing LINEs retrotranspose site- nonspecifically [12,13]. The early branching group of LINEs are also called RLE bearing LINEs and late branching group of LINEs are called APE bearing LINEs.

RLE and APE bearing LINEs are divided into 6 groups which are further sub-divided into more than 28 clades (Figure 1-1) [14]. The 6 groups of LINEs are R2, RandI, L1, RTE, I, and Jockey. RLE bearing LINEs are composed of the most ancient R2 group and includes Genie, CRE, NeSL, R2, Hero, and R4 clades. APE bearing LINEs are composed of L1, RTE, I, and Jockey groups and includes all other clades except for RandI (Figure 1-2).

RandI (Dualen) group of LINEs encode both APE and RLE endonuclease flanking on either side of RT domain on a single ORF and are considered as progenitors of late branching APE bearing LINEs [15,16].

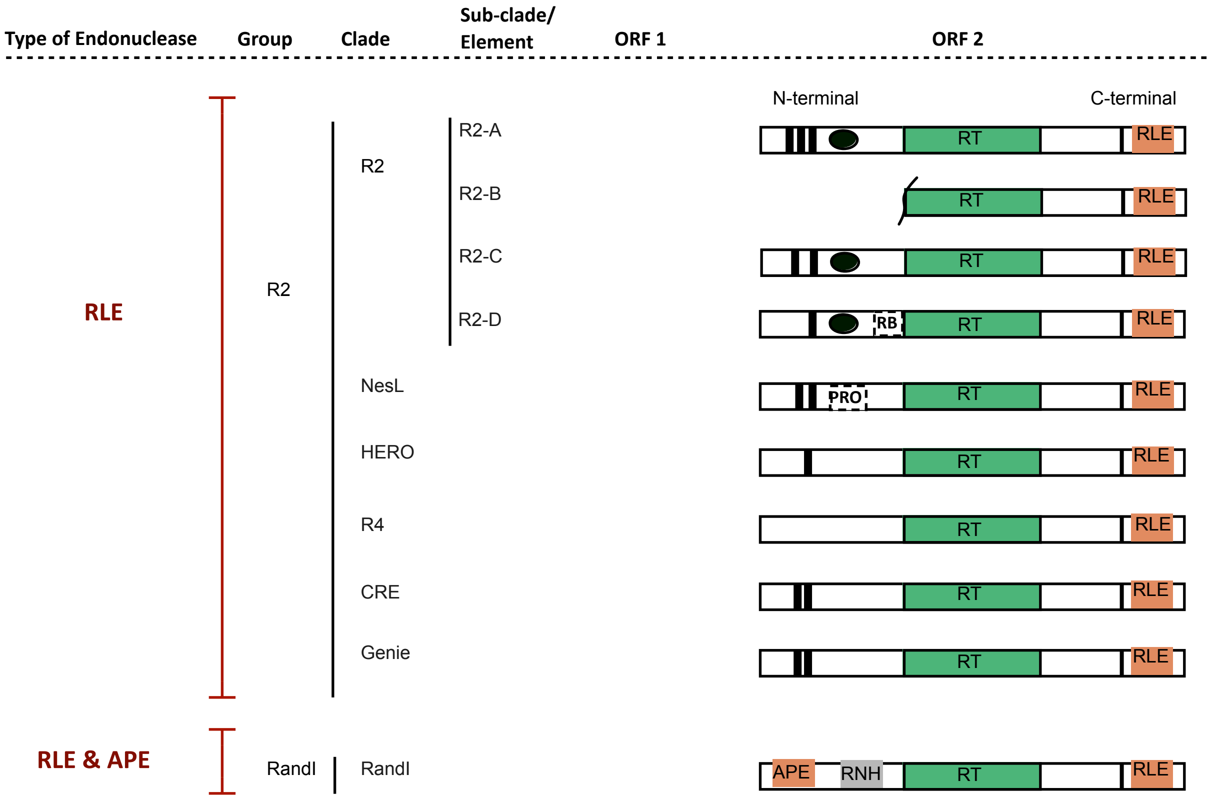


Figure 1-1: Open reading frame (ORF) structure of RLE bearing LINEs from different clades. Open boxes represent ORF and shaded boxes represents the element encoded enzymatic domains. All elements have centrally located reverse transcriptase (RT) and a C-terminal region with zinc knuckle CCHC motif (thin vertical line) and a restriction like endonuclease (RLE). N-terminal region has variable number of zinc finger (ZF) CCHH motif (thick vertical lines), a Myb-like (Myb) domain (black oval), a RNA binding domain (RB), and a cysteine protease domain (PRO) [17–19]. This figure is adapted from [8].

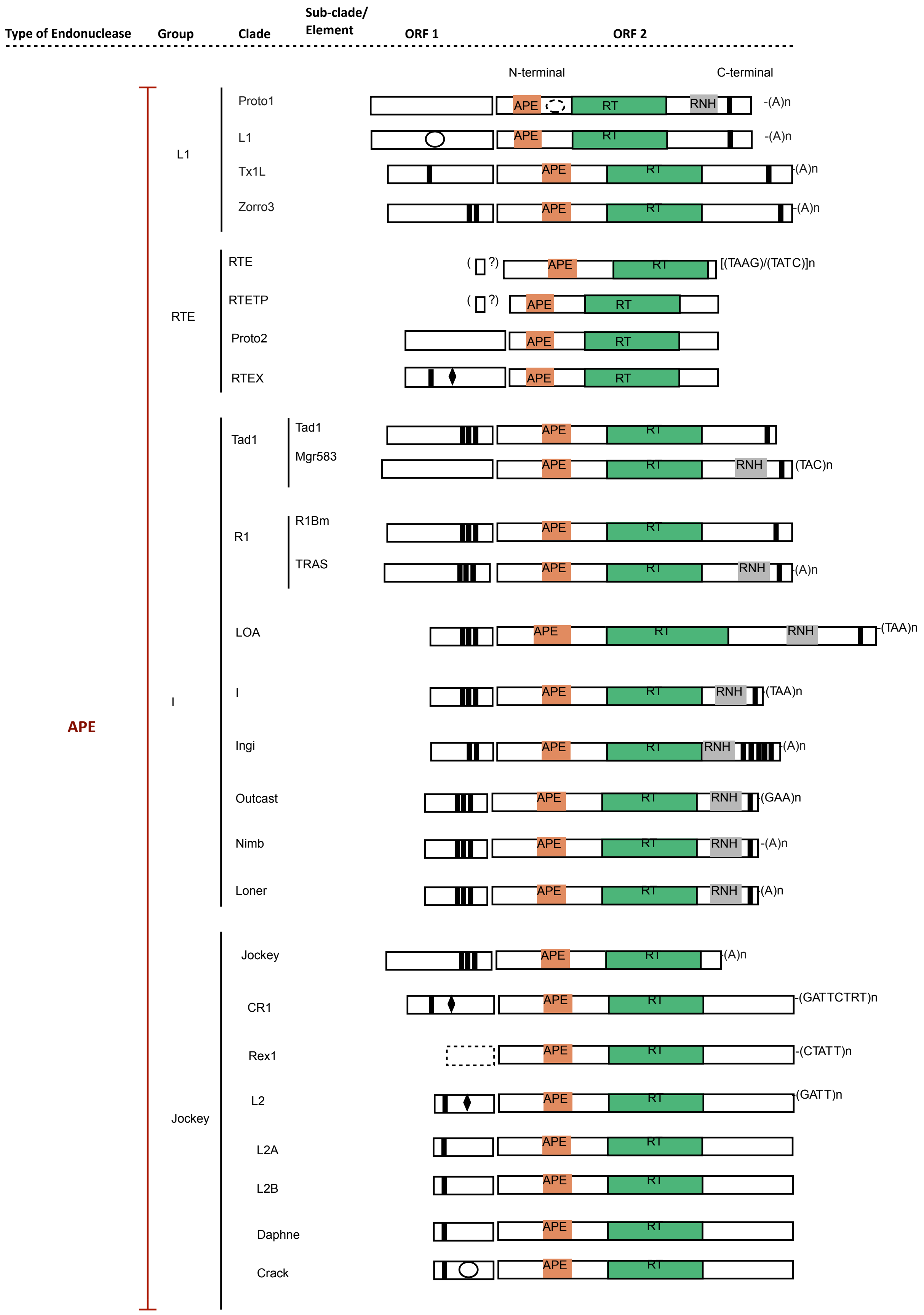

Type of Endonuclease
Group
Clade
Sub-clade/ Element
ORF 1
ORF 2
N-terminal
C-terminal
APE
L1
Proto1
APE
RT
RNH
-(A)n
L1
APE
RT
-(A)n
Tx1L
APE
RT
(A)n
Zorro3
APE
RT
(A)n
RTE
RTE
(?)
APE
RT
[(TAAG)/(TATC)]n
RTETP
(?)
APE
RT
Proto2
APE
RT
RTEX
APE
RT
I
Tad1
Tad1
APE
RT
Mgr583
APE
RT
RNH
(TAC)n
R1
R1Bm
APE
RT
TRAS
APE
RT
RNH
(A)n
LOA
APE
RT
RNH
(TAA)n
I
APE
RT
RNH
(TAA)n
Ingi
APE
RT
RNH
(A)n
Outcast
APE
RT
RNH
(GAA)n
Nimb
APE
RT
RNH
-(A)n
Loner
APE
RT
RNH
(A)n
Jockey
Jockey
APE
RT
(A)n
CR1
APE
RT
-(GATTCTRT)n
Rex1
APE
RT
-(CTATT)n
L2
APE
RT
(GATT)n
L2A
APE
RT
L2B
APE
RT
Daphne
APE
RT
Crack
APE
RT

Figure 1-2: Schematic ORF(s) structure of APE bearing LINEs from different clades. The ORF1 and ORF2 are shown as short and long open boxes, respectively. ORF1 have variable number of zinc finger motif (thick vertical lines), RNA recognition motif (RRM) and esterase (diamond and open ovals). ORF2 have Apurinic-apyrimidinic endonuclease (APE), bioinformatically predicted myb domain (black oval), PCNA interacting protein domain (PIPbox) (dotted oval), Reverse transcriptase (RT), RNase H (RNH), and Zinc knuckle CCHC motif (think vertical lines). The ORF1 for RTE element is putative and only 43 amino acids. The dotted open box in Rex1 indicates that the 5' end of this element has not been identified yet. The ORFs are not drawn to scale. This figure is adapted from [14,20].

1.3 Lifecycle of LINE and basic integration mechanism

LINE retrotransposition cycle is shown in Figure 1-3. The 5' untranslated region (UTR) of some LINEs encode an internal promoter that initiates transcription by RNA polymerase II to generate mRNA transcript [21–23]. There are also LINEs with no endogenous internal promoter like R2 elements that are cotranscribed with their host rRNA genes [24]. Following transcription, nuclear export of mRNA into the cytoplasm occurs where the LINE ORF(s) are translated into protein [25,26]. ORF1 of elements transcribed by RNA Polymerase II, like human L1 elements, is translated in a cap-dependent manner [27]. RNA Polymerase I transcribed elements like R2 are most probably translated via Internal ribosomal entry site (IRES) [24]. LINE encoded protein(s) show *cis*-preference and binds with the mRNA from which it is translated to form a ribonucleoprotein (RNP) complex [25,28]. The RNP complex is transported back into the nucleus where they bind the chromosomal target site and get integrated by a mechanism termed as target primed reverse transcription (TPRT) [6,29]. Integration by TPRT mechanism into a new site occurs in four steps (Figure 1-3): i) First strand DNA cleavage, ii) First strand synthesis by TPRT, iii) Second strand cleavage, iv) Second strand synthesis [30,31]. First and second strand cleavages are catalyzed by LINE encoded endonuclease, whereas first and second strand synthesis are catalyzed by LINE encoded reverse transcriptase [9,31–33]. Detailed integration mechanism is discussed later.

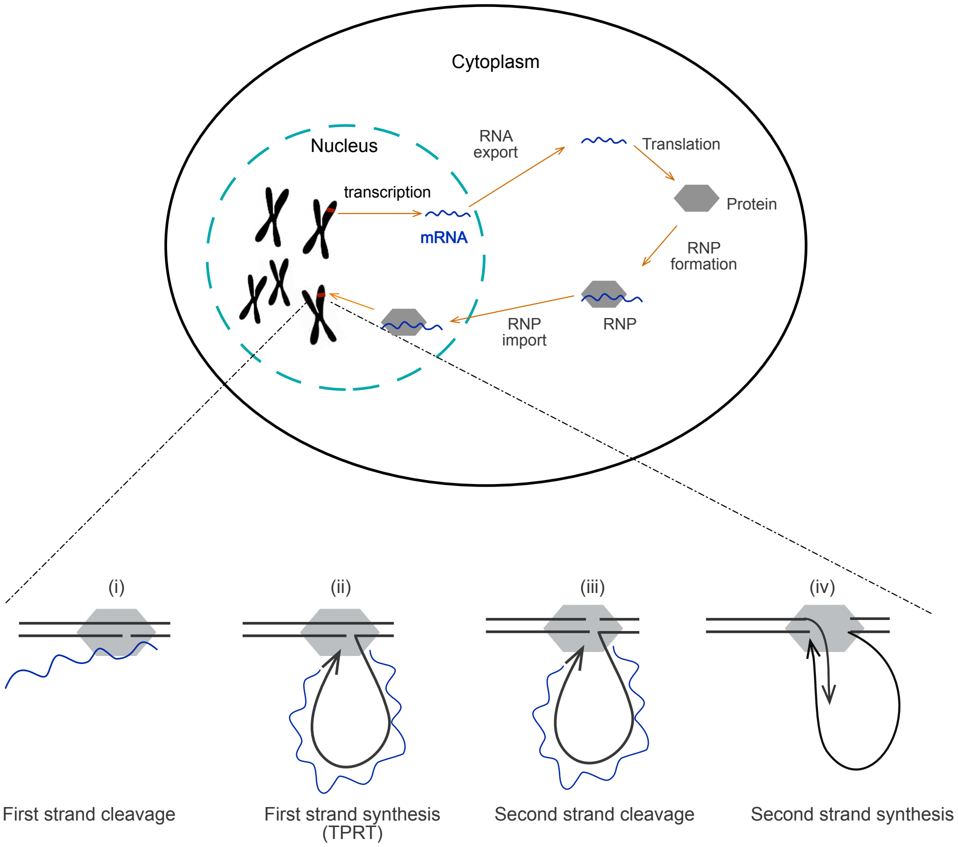


Figure 1-3: Schematic representation of LINE retrotransposition cycle. During a complete life cycle, an LINE element is transcribed and the resulting mRNA is exported out of the nucleus into the cytoplasm. Then, the RNA is translated into protein(s), which bind the element RNA in cis to form ribonucleoprotein (RNP) particles. Finally, RNP complexes are imported into the nucleus where the RNA can be integrated at a new genomic location by the process of TPRT. Integration mechanism is zoomed in: (i) LINE encoded endonuclease cleaves the first (antisense) target DNA strand to expose 3'-OH group. (ii) The 3'-OH group is used to prime first strand synthesis by element encoded reverse transcriptase (TPRT). (iii) LINE encoded endonuclease make another cleavage on the second (sense) DNA strand. (iv) Reverse transcriptase catalyzes the second strand synthesis thus completing element integration. Adapted from [30,34].

1.4 ORF structure of RLE LINEs

ORF structure of all RLE bearing R2 group of LINEs encode one to three Zinc finger(s) (ZF(s)) at the N-terminal end (except R4 clade); a centrally located RT domain; a linker region containing presumptive α -finger (H/RINALP residues in R2 elements), a zinc knuckle like cysteine/histidine rich (CCHC) motif; and an RLE domain at the C-terminal end of ORF (Figure 1-1). The major differences between the different clades of RLE bearing LINEs lie on the N-terminal region of ORF. Specifically, R2 clade of LINEs encodes an additional Myb domain located in between ZF and RT domain. RNA binding (RB) motif preceding the reverse transcriptase domain has been identified in R2 elements from R2-D clade. NeSL clade contains *Ulp1*-like cysteine protease (PRO) domain in the N-terminal region between ZF and RT domain [19]. R4 clade lack both ZFs and Myb motifs in the N-terminal region. The most extensively studied RLE bearing elements are the R2 clade elements, especially the R2 element from *Bombyx mori* (R2Bm). The protein encoded by R2Bm RLE LINEs has been identified to form two globular-domains: the reverse transcriptase, linker, and RLE together form a single large globular-domain and the N-terminal region forms a smaller globular domain [35].

1.4.1 Zinc fingers and Myb motif

The N-terminal ZF and Myb domain of site-specific RLE LINEs has been identified to contribute to target DNA binding, i.e., to load the LINE protein onto the target DNA. They interact with target DNA by binding to either upstream or the downstream sequences relative to the insertion site. The ZF of R2 clades is comparable to that of the DNA binding cysteine-histidine ZF motif of a eukaryotic transcription factor, TFIIIa [36]. The ZF motif of TFIIIa forms an alpha-helix initiating from the second cysteine to first histidine residue of the motif. Three amino acids within the helix contacts three consecutive bases of the target DNA, thus providing binding in a sequence-specific manner [36]. The Myb motif of R2 clade is comparable to the DNA binding region of c-Myb protein which consists of three tandem repeats. The second and third repeats each contain three helices, the third of which are called recognition helices, and package in the major groove to directly contact specific DNA bases [37]. R2 clades have conserved amino acid sequences that correspond to the first and the third helix.

Elements belonging to NeSL, CRE, and Genie clades typically encode two N-terminal ZFs while HERO clade encodes one [38]. R2 clade encodes a variable number of ZFs along with a Myb domain [17,38,39]. Elements from R4 clade lack both ZFs and Myb domain. R2 clade can be further divided into four sub-clades: R2-A, R2-B, R2-C, and R2-D—based on reverse transcriptase phylogeny [17]. R2-A clade encodes three N-terminal ZFs (ZF1, ZF2, and ZF3), the first one being just upstream to the Myb domain, and with a consensus cysteine and histidine spacing of $CX_2CX_3FXT/SX_2GX_3HX_4H$, $CX_2CX_{12}HX_3C$ and $CX_2CX_{12}HX_4H$, respectively [17,38,39]. The amino-terminal structure of R2-B subclade has not yet been determined [17]. R2-C subclade encodes two ZFs, while it lacks the ZF2 corresponding to R2-A subclade [17]. R2-D encodes only one ZF while it lacks ZF2 and ZF3 corresponding to R2-A subclade.

There appears to be a variability on how ZFs and Myb motifs bind to target DNA. R2-D elements (R2Bm) uses the ZF and Myb-motifs to bind downstream of the target insertion site while R2-A elements (R2Lp, R9Av) uses these motifs to bind upstream of the target insertion site [40]. NeSL-1 has been reported to use its two ZFs to perform the upstream target DNA binding implying its similar role in CRE and Genie clades since they all have two ZFs [19]. In some cases, ZF was closest to the insertion site and in other cases, Myb was closest to the insertion site. The variability in binding modes suggests plasticity in how binding domain functions are wired into integration mechanism [41]. Between the ZF and Myb motif, the latter is used to attain major specificity in the target binding and DNA contact. In R2Bm*,* ZF motif was found to make contacts with 1 to 3 bases upstream of the insertion site and Myb motif makes contact with the 10 to 15 bases downstream relative to the insertion site [42].

1.4.2 Cysteine Protease domain

Only elements of NeSL clade contain Ulp1-like cysteine protease domain immediately after N-terminal ZFs [19]. The best characterized members of cysteine proteases related to NeSL-1 domain are Smt-3 and -4 proteases of yeast that removes ubiquitin-like linkages from proteins [43]. The exact function of this domain is unknown; however, two possible roles have been postulated [19]. The protease could possibly process RT domain of protein from DNA binding and/or RLE domain, and this domain would be like the RNase H domain found in some LINEs. It has also been speculated that the protease domain might not be required for NeSL-1 activity but for one or more cellular functions of the host.

1.4.3 RNA binding domain

The protein encoded by R2 elements are known to bind two distinct folded structures located at the 5' and 3' untranslated regions of R2 mRNA [24,44]. Very recently, in R2 elements, two conserved motifs in RT-1 and RT0 domains located immediately upstream to the RT domain have been identified to be involved in RNA binding [18]. These motifs are termed as 0 and -1 because they are encoded just before the motifs 1 to 7 domains of the RT. Mutations created in these motifs greatly reduced the ability of the protein to bind both 3' and 5' regions of R2 RNA [18]. In addition, the ability of the protein to synthesize the first strand using RNA as a template was significantly reduced, so was the ability of the protein to cleave the second strand in the presence of 5' region of R2 RNA [18]. These motifs are known to be involved in all activities of the protein that require specific RNA binding. Sequence similarity in 0 motif has been found in all lineages of LINEs [10,45,46], implying a possibility of similar RNA binding activity.

1.4.4 RT domain

The centrally located RT domain is functionally the most important and universally present domain in the protein encoded by all LINEs [1]. Classification of LINEs into different clades have reliably based upon the phylogeny of RT domain. There are seven blocks of amino acids that bear high sequence similarity among all LINEs and are named as RT1 to RT7 domains [45]. Detailed enzymatic studies of RT have been conducted for R2Bm elements. R2 RT is characterized by several enzymatic activities that differentiate it from RTs encoded by Long terminal repeat (LTR) retrotransposons and retroviruses. They show higher processivity on DNA and RNA templates generating longer product lengths [47]. R2 RT can end-to-end template jump from 5' end of one RNA template to the 3' end of the second RNA template in the absence of sequence identity between the two templates [48,49]. They do so by adding non-templated (overhanging) nucleotides to the cDNA when it reaches the end of the donor RNA and annealing the overhanging nucleotides to the sequences at the 3' end of the acceptor RNA [48]. Also, a distinct property of R2 RT is the ability to use the 3' end of RNA or single-stranded DNA to prime reverse transcription [49]. There is no requirement of sequence complementarity between the template and the primer. The RT can displace an annealed RNA strand while using a DNA strand as a template. In R2Bm, the synthesis reaction at the 28S target site is more efficient if the RNA contains 3' UTR as a template [50]. If this RNA template ends at the precise 3' end of the R2 element, then non-templated nucleotides are added to the target before cDNA synthesis [51]. The presence of downstream 5 to 10 nucleotides of the 28S sequence are used most efficiently *in vitro* TPRT reaction [51]. The R2Bm RT domain has recently been 3D modeled and it assumes a canonical hand-like configuration, with fingers, palm, and thumb regions, and was overall similar to RNA dependent RNA polymerase (RdRP) [52–54]. The index finger of the RT and the -1 regions were the most accessible areas for proteolytic cleavages to occur, and the thumb region of RT is found to be highly protected from cleavages [35].

1.4.5 Linker region with conserved predicted α-helices and zinc knuckle

The linker region located between the RT and EN domains of RLE LINEs remain as one of the less explored areas. The linker region harbors an important zinc-knuckle like cysteine/histidine rich (CCHC) motif with a spacing of $CX_{2\text{-}3}CX_{7\text{-}8}HX_4C$ [35,55]. The zinc knuckle like CCHC motif is found to be universally conserved, and hence common to both RLE and APE LINEs, however, its role in LINE mobilization remains ambiguous and understudied. In a previous *in vivo* study using APE bearing human LINE-1 elements, mutating the first two cysteines to serine in the CCHC motif significantly reduced LINE-1 retrotransposition [56]. Mutating the third cysteine to glycine in zinc knuckle of APE bearing LINEs called SART1 from Silk moth lost the ability to retrotranspose *in vivo* [57]. These studies called attention to the significance of the zinc knuckle for LINE retrotransposition. Diminished amount of protein bound RNP complex formation was observed *in vivo* with human LINE-1 elements where the first two cysteines were substituted to serine, implying its possible role in RNA binding [58]. When the zinc knuckle structure was altered by substituting four cysteines (three cysteines in CCHC motif and one immediately before the motif) into serine, no reduction in RNA binding activity was reported for human LINE-1 elements *in vitro* [59]. Although the cysteine-histidine rich zinc knuckle of the linker has shown to be indispensable for retrotransposition, and formation of RNP complexes, the actual biochemical role of this region is yet to be dissected. Our previous study has discussed structural similarity of zinc knuckle of LINEs to the ββα structure formed by non-zinc knuckle found in the linker region of Prp8, a eukaryotic splicing factor [35]. Like RLE LINEs, Prp8 protein also has very similar RT and RLE domains connected by a long linker region, which suggest a possibility of RLE LINEs and Prp8 sharing a common ancestor during their evolution. The amino acid sequences corresponding to the ββα structure of Prp8's non-zinc knuckle aligns well with that of LINEs. In Prp8, the residues in the knuckle form a part of the pocket to sequester G nucleotide of the intron 5' splice site [60]. Also, two residues in this region are known to coordinate G nucleotide via putative hydrogen bonds [61]. In addition to being structurally equivalent to the Prp8 knuckle, the zinc knuckle is also speculated to be functionally similar.

Immediately upstream of the zinc knuckle like CCHC motif, lies a set of α-helices predicted to be conserved in both RLE and APE LINEs. This region contains highly-conserved residues in RLE LINEs (R/HINALP motif in R2 clade elements) which aligns very well in sequence with the APE bearing LINEs in multiple sequence alignments [35]. While this presumptive α-helices shows sequence and structure conservation among LINEs, this region remains totally unexplored. Mutating HMKK and SSS residues located 14 and 10 amino acids upstream of the RYHLTP sequences in human LINE-1 that aligns well to R2Bm R/HINALP residues, was reported to reduce LINE-1 retrotransposition *in vivo* [56]. In multiple sequence alignment and 2D structural alignments, the presumptive α-helices of LINEs aligns very well with an important motif in Prp8 called the 1585 loop and helices. The 1585 loop and helices also found in the linker region of Prp8 was found to be dynamic, and interacts with nucleic acid and other proteins in the spliceosome complex [35,62–65]. The conserved helical region of LINEs could be structurally, and if possible functionally equivalent to the 1585 loop and helices region of Prp8. This makes the residues in the α-helices of LINEs prime candidates to be investigated for their possible role in integration mechanism of LINEs.

1.4.6 R-box

In a subset of type II restriction type endonuclease enzyme (e.g. EcoRII, DpnII, MboI, PspGI and Sso II), there is an arginine-rich region consisting of RXXR or NXRXXR found within or in front of the first α -helix of the endonuclease fold [66–68]. This region is named as "R-box" in the restriction enzyme and is known recognize DNA, and correctly position the DNA at the active site for cleavage. In a PD- (D/E)XK family of Holliday junction resolvases (Hjc, Hje), similar DNA binding residues were found at the beginning of the resolvase [69–71]. RLE LINEs also have similar R-box region with conserved "RH" residue at the beginning of α-helix1 of the RLE domain. Mutating both the "RH" residues to alanine reduced the ability of the protein to bind to the target DNA at the upstream and downstream sequences from the insertion site by ~40% and 30%, respectively [72]. The binding activity of RH motif does not seem to be site or sequence specific. In addition, the ability of the protein to synthesize the first strand by TPRT was also reduced. Reduction in DNA binding was not observed when only the H residue was mutated [72]. There is another arginine preceding the RH residue which is also speculated to be involved in the DNA binding, however, further analysis is required.

1.4.7 RLE domain

RLE encoded by members of R2 clade belongs to a superfamily of the PD-(D/E)XK endonucleases and is shown to have PD-(D/E) restriction endonuclease [73]. R2 RLE includes highly conserved residues of PD-$X_{12\text{-}13}$-D-$X_{16\text{-}18}$-K. The first of PD-(D/E)XK endonucleases to be identified were the type II restriction enzymes (e.g. EcoRI, BamHI, and FokI) [74]. R2 RLE bears a conserved core of four-stranded mixed β-sheet flanked by α-helix on each side (αβββαβ) [72]. Mutating the two catalytic D residues abolished the cleavage activity, but not binding or the subsequent TPRT process [73]. The catalytic residues D-(D/E) are at the beginning of β-strand 2 and within β-strand 3, respectively. The catalytic K has been identified to be located in a non-canonical position in the α-helix 2 [72]. There is a second K residue located within α-helix 2 immediately after the first catalytic K which could also be catalytic. R-box residues that are involved in DNA binding are located at the beginning of α-helix 1. Mutating the residues located beyond the β-strand 4 (R/AG.W) was shown to reduce DNA binding and cleavage [72].

1.5 ORF structure of APE LINEs

APE bearing LINEs typically have two open reading frames that encode for ORF1 and ORF2 proteins both of which are required for L1 retrotransposition [56,75]. The ORF1 usually encodes one to three zinc fingers, and ORF 2 typically encodes APE domain always preceding the RT domain followed by a cysteine/histidine rich gag knuckle (CCHC motif). Some of the clades of APE LINEs (Tad1. R1, LoA, I, and Ingi) encode ribonuclease H (RNH) domain in the C-terminal domain of ORF 2 preceding the CCHC motif. The most well-studied elements of APE bearing LINEs are the L1 and I elements. An active human L1 element is ~6 kb in length and contains two ORFs flanked by 5' UTR and 3' UTR. The ~910 bp long 5' UTR contains RNA polymerase II sense strand promoter and antisense promoter along with *cis*-acting binding sites for few transcription factors necessary for LINE-1 transcription (57). The 3' UTR includes a functional RNA polymerase II polyadenylation signal [76].

1.5.1 ORF1

Human L1 ORF1 encodes a 40 kDa protein (ORF1p) that is translated from mRNA in a cap-dependent mechanism [27]. Biochemical studies have that mouse and human L1 ORF1p resides in cytoplasmic RNP and binds to single-stranded DNA [25]. Structural studies have identified a coiled-coil domain at the N-terminal region that assists in the trimerization of ORF1p [77]. Centrally located is an RNA recognition motif (RRM) followed by a basic C-terminal domain. The central RRM along with the C-terminal region of ORF1p is required for ORF1p RNA binding in human L1 elements [77–79]. ORF1p of many APE bearing LINEs encode one to three CCHC motifs generally represented by $CX_2CX_4HX_4C$ [36,80]. The spacing of the second CCHC motif in some and the third motif in most APE LINEs is less conserved. In SART1 elements, each of the three CCHC motifs was important for its reptrotransposition, and packaging on ORF1p into RNP in a sequence-specific manner [80]. Similar CCHC motifs in the C-terminal region are also found in ORF1 of many other APE bearing LINEs [14]. In addition, the 13 amino acids, VARIGECPPDIVK found just upstream of CCHC motifs in the ORF1 of SART1, was found to be crucial for ORF1p-ORF1p and ORF1p-ORF2p interactions [80]. Human, mouse and non-mammalian ORF1p also has nucleic acid chaperone activity and can facilitate re-annealing of single-strand DNAs *in vitro* [81]. Nucleic acid chaperone activity has also been hypothesized to facilitate initial steps of L1 integration *in vivo*. However, deleting ORF1 does not abolish ZfL2-1 retrotransposition activity in cultured human cells [82]. Also, Alu retrotransposition only requires protein encoded by L1 ORF2 [83]. These data raise the question as to how ORF1 nucleic acid chaperone activity plays a part in LINE retrotransposition.

1.5.2 ORF2

The ORF2 of members of the L1, RTE, I, and Jockey groups encode the apurinic-apyrimidinic endonuclease (APE), which is always N terminal to the RT domain [45]. Although the ORF2 APE shares similarity with apurinic-apyrimidinic endonuclease, the APE has lost its ability to cleave apurinic/apyrimidinic site except for endonuclease of L1Tc from *Trypanosoma cruzi* [84,85] Human L1 ORF2 encodes a 149 kDa multifunctional protein (ORF2p) critical for L1 retrotransposition [56,85,86]. The human L1 APE makes a nick at the consensus sequence of 5′-TTTT/A-3’ in genomic DNA exposing a 3’ hydroxyl group at the cleaved site [85].

A c-myb like three helix motif has been predicted in an area between APE and RT in TRAS, R1Bm, SART1, RT1Ag, TARTDm, and L1Hs element using secondary structure prediction program, implying its possible role in binding the target DNA during integration [87,88].

PCNA, which is the sliding clamp protein essential for DNA replication was found to be directly interacting with a conserved sequence in the ORF2p [89]. The interaction was via a PCNA interaction protein domain called PIP box was found in a region between the APE and RT domain of L1. Mutating the residues in PIP box abolished L1 retrotransposition [89], however, the exact role of PIP box in L1 retrotransposition is unknown.

Centrally located in ORF2p is the RT domain that shares sequence similarity with RT encoded by telomerase, Penelope-like retrotransposons, group II introns, other LINEs, LTR retrotransposons, and retroviruses [45,90,91]. The RT exhibit both RNA dependent and DNA dependent polymerase activity [92]. Like RT encoded by R2Bm, L1 RT shows high processivity and lacks RNase H activity [32,47]. The ORF2p exhibits *cis*-preference and preferentially reverse transcribes its own mRNA [93]. Point mutation on the ORF1 and APE domain adversely affects L1 retrotransposition but the ability of the protein to reverse transcribe is not affected [93]. The RT can also extend terminally mismatched primer-template complexes [93,94]. ORF2p has been shown to co-localize with ORF1p and mRNA in cytoplasmic foci [58]. Although ORF1p is found to be in abundance compared to ORF2p in L1 RNPs [58], the exact stoichiometry of ORF1p and ORF2p bound to a single mRNA is not yet elucidated.

The C-terminal region of ORF2p in most APE LINEs contain at least one zinc knuckle like cysteine/histidine rich (CCHC) motif similar in spacing to that found in the C-terminal region in the protein encoded by RLE LINEs (mostly $CX_{1–3}$ $CX_{7–8}$ HX_4 C or CX_2 CX_{12} $HX_{3–5}$ H) [45,95]. Ingi element being a special case carry five degenerate cysteine/histidine rich regions. Elements of the Jockey group lack a zinc-finger domain in ORF2p [96]. Mutating the cysteine residues to serine in human L1 affects the ability of the protein to form RNP, and adversely affects L1 retrotransposition in cultured cells [56,58]. Studies using the recombinant protein containing 180 amino acids at the C-terminal end of ORF2 have shown non-sequence specific RNA binding *in vitro*, however, mutating the cysteines to serine do not adversely affect RNA binding [59]. The exact function of this highly conserved CCHC motif found in almost all LINEs remains to be elucidated.

ORF1 and/or ORF2 protein encoded by APE LINEs have shown to act in *trans* to mobilize non-autonomous retrotransposons like human Alu and SVA elements, and mouse B1 and B2 elements [83,97,98]. They are also capable of mobilizing cellular mRNAs to generate processed pseudogenes [99].

1.6 Target Site specificity of LINEs

The target insertion sites for most of the target site-specific LINEs are repetitive sequences found in the host genome. Most of the RLE LINE elements from R2, NeSL, and R4 clades, and some elements from HERO, CRE, and Genie clades are found to be site-specific during insertion. Elements from R2 cades were found to be inserted at a specific sequence in the 28S rDNA locus of many vertebrates and invertebrates [17,100,101]. Elements from NeSL clades like NeSL-1 in nematodes were found to target the spliced leader exons, while NeSL-1_Aca form the amoeba *Acanthamoeba castellanii were found to target* U2 snRNA gene [19,102]. Dong and R4 elements that belong to R4 clade were found in microsatellite TAA repeats, and another location of 28S rDNA, respectively [103,104] While most of the elements belonging to HERO clade is found to be non-specific, HERO-1_HR was recently found in a microsatellite, $(ATT)_n$ repeats [105]. Site-specific elements in CRE clades are CRE1, CRE2, SLACS, and CZAR in trypanosomes that were also found in the spliced leader exons [106,107] Genie-1 elements of Genie clade was found to be inserted into the 771-bp repeat near the telomere, however, Genie-2 elements were found to be non-specific in insertion [108]. Some of the representative elements of each if the RLE bearing LINE group with their specific site of insertion is presented in Table 1.

Out of more than 20 clades of APE LINEs, only Tx1 and R1 clades contain elements that are site specific during integration (Table 1). Site-specific elements belonging to Tx1 clade targets rRNA genes (Mutsu), tRNA genes (Dewa), snRNA genes (Keno), telomeric repeats (Tx1-1_ACar), and microsatellites (Kibi and Koshi) [109]. Tx1 and Tx2 elements have been in other retrotransposons Tx-1D and Tx-2D [110]. R1 clade elements integrate in a sequence-specific manner and target rDNA (R1/R6/R7/RT), telomeric repeats (TRAS/SART), and microsatellites (ACAY or AC repeats) (Waldo) [111]. Phylogenetic analysis of target sequences of R1 clade elements has shown that the target specificity has been altered several times independently over the course of evolution [111]. Some of the site-specific R1 elements like HOPE, Hal, Hida, Kaga, and Noto have now lost their specificity of integration [111]. Remaining of the clades of the APE LINEs do not integrate non-specifically, however, they show weak sequence specificity (human LINE-1 show specificity for TAAA repeats) [12,13].

Table 1- 1: Target site specificity of RLE and APE bearing LINEs. Adapted from [8,41,112].

Group	*Clades*	*Representative elements*	*Target insertion site*
RLE LINEs	R2-A	R2Lp, R8Hm, R9Av	rRNA gene (28S R2, 18S, 28S R9)
	R2-D	R2Bm	rRNA gene (28S R2, 18S, 28S R9)
	NeSL	R5, NeSL-1Ce, R5-2_SM	rRNA gene, Spliced leader, Transposon
	R4	R4Al, R4-2_Sra, Dong	rRNA gene (28S R4), tRNA-Asp gene, Microsatellite
	HERO	HERO-1_HR	Microsatellite
	CRE	CRE2Cf, MoTeR, CRE-1_NV	Spliced leader, Telomeric repeat, Microsatellite
	Genie	Genie-1Gl	771 bp repeat near telomere
APE LINEs	Tx1	Mutsu, Dewa, Keno, Tx1-1_ACar, Kibi, Koshi, Tx1, Tx2	rRNA (5S), tRNA, snRNA, Telomeric repeats, microsatellites, other transposons (Tx1D, Tx2D).
	R1	R1/R6/R7/RT, TRAS/SART, Waldo	rRNA genes, Telomeric repeats, microsatellites,

Abbreviations: R9 element from *Adineta vaga* (R9Av), R8 from *Hydra magnipapillata* (R8Hm), R2 from *Limulus polyphemus* (R2Lp), R2 from *Bombyx mori* (R2Bm), R4 from *Ascaris lum- bricoides* (R4Al), R4 from *Strongyloides ratti* (R4-2_Sra), NeSL from *Caenorhabditis elegans* (NeSL-1Ce), NeSL from *Schmidtea mediterranea* (R5-2_SM) CRE2 from *Crithidia fasciculata* (CRE2Cf), CRE from *Nematostella vectensis* (CRE-1_NV), HERO from *Helobdella robusta* (HERO-1_HR), Genie-1 from *Giardia lamblia* (Genie-1Gl).

1.7 Transcription of LINEs

LINEs are transcribed into mRNA either by an internal promoter or they rely upon their upstream cellular promoter to be transcribed as a co-transcript which later gets processed into LINE mRNA in the nucleolus. Human L1 elements, which is an APE bearing LINE, encode an internal RNA polymerase II promoter in the 5’ UTR [22,27]. R2 elements, which is a site-specific RLE bearing LINE, rely on the transcription of host rRNA gene using RNA polymerase I to be transcribed as 28S/R2 co-transcript [113]. The conserved 5' end of R2 RNA can be folded into a double pseudoknot structure encodes an autocatalytic self-cleaving ribozyme which is similar in sequence and structure of Hepatitis Delta Virus (HDV) ribozyme [114,115]. The HDV ribozyme is a self-cleaving RNA that folds into a conserved double pseudoknot structure and catalyze transesterification reaction leading to cleavage of RNA sugar-phosphate backbone [114]. The 5' ribozyme encoded by R2 RNA enables processing of R2 RNA from the 28S co-transcript. In contrast to most eukaryotic mRNAs, the self-scission of ribozyme yields a transcript with 5'- OH group that cannot be capped at the 5' end. HDV like self-cleaving ribozymes has also been reported in APE bearing LINEs [115,116]. Structure-based bioinformatics searches have identified the presence of HDV like 5' ribozyme in other promoterless RLE LINE including rDNA-specific R4 elements, and also in APE LINEs [116]. The site of self-cleavage by the *Drosophila simulans* R2 ribozyme differs from that of the other diverse animals. In *Drosophila simulans,* the cleavage occurs precisely at the 28S/R2 5' junction while in others the cleavage occurs in the GC-rich area 13 or 28 nucleotides upstream of the 5' end and the R2 transcript includes the upstream 28S sequences [7,117]. The upstream 28S sequence co-transcribed at the 5' end in R2 transcript has been suggested to play a role in priming second strand synthesis [117]. It is not known how the 3' end of the R2 transcript is processed from the 28S rRNA co-transcript. The ~250nt long 3' UTR only is necessary and sufficient for TPRT but appears to add non-templated nucleotides before cDNA [51]. The presence of 5 to 10 nucleotides from the downstream 28S sequence was found to be used most efficiently in TPRT assays *in vitro* and help initiate TPRT at an accurate position *in vivo* [51]. The requirement for downstream 28S rRNA sequences probably explains why R2 RNA does not contain an A-rich repeat at their 3' junction with the target DNA.

Transcription of APE bearing human L1 is controlled by an internal RNA polymerase II promoter located within the first 100-150 bp at the 5' UTR which initiates their transcription within a small region of -9 to +4 [22,23,118]. The L1 transcript starts with ~910 bp 5' UTR followed by ORF1, a 63-nt spacer, ORF2, ~210 bp 3'UTR and ends with a poly(A) tail [119]. The ORF1 and ORF2 of the bicistronic L1 transcript of is separated by a 63-nt inter ORF spacer which contains two in-frame stop codons and one out of frame AUG codon [120]. Most L1 transcripts contain a 5' end 7-methyl guanosine cap structure which facilitates their translation [27]. A binding site for a Ying Yang-1(YY1) transcription factor was identified in the 5' UTR (+13 to +21) of the human L1 element, and mutational studies have identified that YY1 binding sites direct precise transcript initiation at the +1 site of the L1 element [121]. Additional cis-acting transcription factor binding sites for Runx3, Sp1, and SRY-related (Sox) proteins have been identified in the 5'

UTR that must be involved in L1 transcription [22,121–123]. SART-1 which belongs to R1 clade of APE LINEs requires both 3' UTR and poly (A) tail for its retrotransposition while in human L1, the 3' UTR seems to be dispensable [56,57]. R1 elements which belong to R1 clade of APE LINE, do not harbor an A- rich tract at its 3' end but requires a 3' structure for its retrotransposition [124]. The APE LINEs that lack RNA polymerase promoter encodes a HDV like ribozyme at the 5' end of their RNA transcript like the ones found in rDNA specific R6, and telomere-specific SART, Baggins, and RTE elements [116]. L1 elements from *Trypanosoma cruzi* also contains HDV like ribozyme at the 5' terminus along with an internal promoter that generates L1 transcript [116]. However, it is shown that L1 ribozyme was lost from a lineage of L1 called L1PA suggesting that L1 ribozyme is not beneficial for retrotransposition (64).

1.8 Translation of LINEs

After transcription of LINEs, the RNA transcript is transported to the cytoplasm where it gets translated into protein. The self-cleaving HDV like ribozyme found in 5' end of LINE RNA adopts a complex pseudoknot structure which is also integral to the activity of some Internal ribosome entry sites (IRESes) seen in viruses and some cellular mRNAs that can interact with translational machinery to initiate protein synthesis. R2 ribozyme has shown to act like an IRES both *in vitro* and *in vivo* [116]. In addition to liberating the 5' terminus of the LINE transcript via self-scission, the ribozyme also acts similarly to an IRES and presumably binds the translation machinery, allowing translation of the downstream ORF to occur [24,125,126]. This mechanism bypasses the need for 5'-methylguanosine cap on the RNA and explains the absence of AUG codon in the R2 LINEs [116].

The ORF1 of the APE LINEs bearing internal RNA polymerase II promoter is translated in a cap-dependent manner [27]. In human L1 elements, two in-frame stop codons are located between ORF1 and ORF2 that are involved in the termination of ORF1 translation and re-initiation of ORF2 translation. L1 ORF2 was translated independently not as ORF1-ORF2 fusion protein [119,127]. In mouse, the IRES located at the 3' end of L1 ORF1 has shown to participate in ORF2 translation initiation [128]. However, in human L1, ORF2 has been proposed to be translated by an unconventional termination/ re-initiation mechanism, and it is possible that host translation machinery is used [127]. Also, ORF1 or ORF1 protein is not required for ORF2 translation. According to the unconventional mechanism, when the ribosome reaches the stop codon in the inter-ORF region, ORF1 protein is released, and the ribosome gets dissociated. The 40S subunit remains associated with the L1 RNA and scans the inter-ORF region until it reaches the first in-frame AUG in ORF2. The ribosome then gets reassembled for ORF2 translation [127]. Experiments have shown that a non-specific translatable upstream ORF is required for ORF2 translation, and ORF2 translation can initiate from a non-AUG codon [127]. It has also been speculated that binding of L1 RNA to ORF2 (or even ORF1) protein inhibits ORF2 translation which would lead to greater amounts of ORF1 protein to coat the transcript. In SART1 elements from the silkworm, mutagenesis studies on the UAAUG overlapping stop-start codon revealed that they follow translational coupling where ORF2 is translated exclusively by the ribosome that translates ORF1 as observed in prokaryotes and virus [129]. In addition, the downstream RNA secondary structure is found to be necessary for efficient ORF2 translation in SART1 [129].

1.9 RNA binding and ribonucleoprotein formation

The protein translated from LINE elements have a strong *cis* preference and binds to the mRNA from which they were translated to form a ribonucleoprotein (RNP) complex essential for TPRT. The *cis* preference of R2 protein for R2 RNA has been well established. Binding of R2 protein to R2 RNA increases the ability of R2 protein to find the target insertion site by ~150 fold [33]. The 5' and 3' UTRs of R2 mRNA can be folded into precise structures that are responsible for binding R2 protein and are named as 5' and 3' protein binding motifs (PBM). 5' PBM is formed by a 300 nt segment that starts within the 5' UTR and ends just before the N-terminal ZF, and adopts a distinctive pseudoknot structure found to be conserved across silk moths [44,126]. 3' PBM constitute the 250 nt of the 3' UTR of R2 transcript. The RNA motif (3' PBM or 5' PBM) bound by the R2 protein determines its function in the integration mechanism [30,33]. R2 protein bound to 3' PBM adopts a conformation that allows it to bind to the 28S gene upstream of the insertion site, and this upstream subunit is responsible for first strand cleavage and first strand synthesis by TPRT. R2 protein from *Bombyx mori* could recognize the R2 RNA 3' UTR from *Drosophila melanogaster* and other distantly related arthropods that has minimum primary nucleotide identity [50,130]. This indicated that R2 protein binding to RNA was not sequence-specific, rather it is mediated by the secondary and tertiary structures at the 3' UTR of the transcript. The secondary structure shared by the two RNA (*Bombyx mori* and *Drosophila melanogaster*) are the three helical regions and the sequence AAC/UAUC in the loop generated by one of these helixes [131]. RNA is considered to contact R2 protein via this conserved region of the transcript. R2 protein bound to 5' PBM binds downstream of the 28S rDNA and this downstream subunit is responsible for second strand cleavage and possibly second strand synthesis to complete LINE integration [31]. The stoichiometry of the reaction suggests a single subunit is involved in either upstream or downstream binding.

Recently, RNA binding motifs in R2 protein has been mapped. One motif in RT-1 region and another motif in RT0 region located immediately N- terminal to the RT domain, have been identified to have both 3' and 5' PBM binding ability [18]. Sequence similarity in 0 motifs has been found in all lineages of LINEs [10,45,46], implying a similar RNA binding domain.

Human L1 ORF1 and ORF2 protein act in *cis* to bind to their mRNA to form RNP complex that is essential for L1 retrotransposition [25,132]. The exact stoichiometry of ORF1: ORF2 protein bond to a single mRNA is not yet elucidated. However, it has been postulated that the RNP complex contains single L1 mRNA bound to multiple ORF1 trimers, and at least one ORF2 protein [58,127]. In addition, it is likely that the RNP includes other cellular proteins and RNAs [89,133,134]. Immunofluorescence microscopy studies have shown that L1 RNA, ORF1, and ORF2 protein accumulate in dense cytoplasmic foci which are closely associated with stress granule proteins [58,89,135]. In another study with yeast retrotransposon, localization of transposon-encoded protein in cytoplasmic foci called processing bodies (P-bodies) is important for RNP assembly, and possibly represents a host mechanism that regulates retrotransposition [136–138]. In TART and HeT-A, ORF1p has been implicated in intracellular targeting (e.g., localization in Het dots) [139,140]. How L1 RNP enters the nucleus is not fully understood.

Biochemical studies have shown that L1 ORF1 protein binds the L1 mRNA in a sequence-dependent manner [25]. A coiled-coil domain located at the N-terminus of the ORF1 protein has shown to facilitate trimerization of ORF1 protein [77,141,142] In telomere-specific SART1 elements, that amino acid residues 555 to 567 and 285 to 567 in the ORF1 protein are crucial for the ORF1-ORF1 protein and ORF1-ORF2 protein interactions, respectively [80]. The central region of ORF1 protein encodes a canonical RNA recognition motif, which along with the basic C- terminal domain (CTD) is involved in binding L1 mRNA [59,77,79]. Recent studies done with a recombinant protein containing the last 180 amino acid (aa) of ORF2 protein has shown to bind non sequence-specific RNA *in vitro* [59]. When cysteines were mutated to serine in the cysteine/histidine rich (CCHC) motif that has been suggested to function like zinc knuckle like domain, RNA binding ability was not severely affected [59]. The domain of L1 ORF2 protein that is involved in specific L1 RNA binding is yet to be elucidated.

1.10 DNA recognition and binding

The active site of the endonuclease encoded by all RLE LINEs shares similarity to that of the type II restriction like endonuclease. The active site and DNA binding site of type II restriction endonuclease are located apart from each other, and hence the enzyme binds the DNA few distances from the cleavage site [73]. Protein footprint analysis on R2 RLE LINE from *Bombyx mori* (R2Bm) has shown similar separation of DNA binding site from the cleavage site, as the protein was found to bind the upstream and downstream sequences from the insertion site [30,143]. The N-terminal region of RLE LINEs encodes a CCHH ZFs and a Myb nucleic acid binding domain. DNA binding and DNAase footprint analysis of mutant polypeptide containing 150 amino acid at the N-terminal end have shown that the ZF motif binds the target DNA 1 to 3 base pairs upstream of the cleavage site and Myb motif binds 10 to 15 base pairs downstream of the insertion site [42]. Complete R2 protein protects the target DNA 10-14 base pairs upstream of the cleavage site, however, the domain that binds these upstream sequences remains unknown [42]. The C-terminal end of the of almost all LINEs encode a zinc knuckle like CCHC motif which has been shown to be indispensable for human L1 retrotransposition and is speculated to be involved in the upstream target DNA binding [30,56]. Analysis of the N-terminal DNA binding motifs of RLE LINEs has shown considerable flexibility in their binding to the target site. While, R2-D elements (R2Bm) uses the DNA binding ZF and Myb-motifs to bind downstream of the target insertion site while R2-A elements (R2 from *Limulus Polyphemus* (R2Lp), R9 from *Adineta vaga* (R9Av)) uses these motifs to bind upstream of the target insertion site [40,42]. NeSL-1 has been reported to use its two ZFs to perform the upstream target DNA binding implying its similar role in CRE and Genie clades since they all have two ZFs [41]. Between the ZF and Myb motif, the latter is used to attain major specificity in the target binding and DNA contact.

Among the APE bearing LINEs that are not site-specific, human L1 inserts randomly with a slight preference for 5'-TTTT/ AA-3' (slash indicates the scissile phosphate), exposing a 3’ hydroxyl and a 5’ phosphate group [85]. The L1 APE has specificity for the DNA structural features found at the TpA junction of T_nA_n homopolymeric stretches [144]. The crystal structure has shown that the L1 APE encoded Bβ6- Bβ5 forms a DNA contacting hairpin loop that inserts into the wide minor groove at the TpA junction [145]. L1 APE recognizes the extrahelical flipped adenine residue 3' of the scissile bond to mediate the cleavage [145]. APE LINEs like I and Jockey elements also gets integrated into the AT-rich regions of the genome [146,147]. Substitution mutation in a similar DNA contacting loop in R1 and TRAS1 clades of LINE (Tyr-98 and N-180 for R1Bm EN and Asp-130 for TRAS1) affects the sequence specificity during integration

[148]. Swapping the APE domain of TRAS1 into SART1 changes the insertion specificity into TRAS1 [57]. This shows that the primary determinant of the DNA specificity of integration is the APE domain itself [145,148,149].

In some APE LINEs (TRAS, R1Bm, SART1, RT1Ag, TARTDm, and L1Hs), a putative myb like domain were identified in between APE and RT domain using secondary structure prediction program [87]. TRAS and SART which are telomeric repeat specific LINEs were found to insert at specific but different nucleotide positions in opposite orientation into the telomeric repeats, $(TTAGG)_n$ [111,150]. In TRAS1 and SART1 elements, the myb like domain has been postulated to be responsible for the general targeting into the telomeres, while their APE determines the insertion position [57].

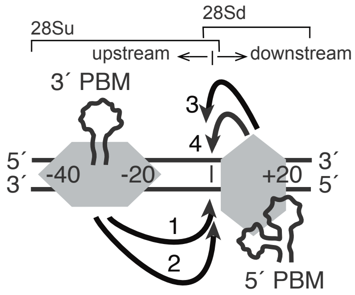


Figure 1-4: R2 integration mechanism. 1. DNA cleavage of the first/ antisense strand. 2. First strand DNA synthesis by TPRT. 3. DNA cleavage of the second/ sense strand. 4. Second strand DNA synthesis. Grey hexagons represent R2 encoded protein. Protein bound to 3' protein binding motif (PBM) RNA binds upstream of the insertion site (28Su) protecting from -20 to -40 base pairs upstream of the insertion site, and protein bound to 5' PBM RNA binds downstream of the insertion site (28Sd) protecting up to 20 base pair downstream of the insertion site in DNA footprinting assay [143]. Black straight line represents doubles stranded target DNA. Adapted from [72].

1.11 First strand cleavage and first strand synthesis by TPRT

In R2 RLE LINEs, the integration process is catalyzed by upstream and downstream protein subunits (Figure 1-4). The R2 protein bound to 3' PBM RNA binds 20 to 40 bp upstream from the cleavage site, thus forming the upstream subunit [30]. The regions of the protein that contact the upstream sequences are yet to be elucidated. The integration mechanism begins when the endonuclease encoded by the upstream subunit cleavages the first strand, i.e., the antisense strand with respect to the 28S rRNA gene promoter (Figure 1-4 step 1) [33]. At the site of cleavage, a 3'-OH is exposed which then acts as a primer for first strand synthesis by TPRT, where reverse transcriptase of the upstream subunit catalyzes reverse transcription of R2 mRNA into cDNA (Figure 1-4 step 2). Annealing of RNA template to the target DNA for initiation of TPRT is not required. R2 RNA template that has the 3' end corresponding to the precise boundary of R2 elements with the 28S gene was found to be most efficiently used for TPRT [50]. R2 RNA that were polyadenylated at the 3’ end (8 nt) or had a truncated 3’ end (3-6 nt) were not efficiently used for TPRT. The utilization of 3’ truncated RNA or the one ending at precise 3’ end of R2 element resulted in the addition of non-templated nucleotides on the target DNA by the reverse transcriptase [50,51]. The ability of reverse transcriptase to add extra non-templated nucleotides on target DNA before engaging the RNA template for TPRT could possibly explain the mechanism for generating poly(A) tail or simple repeat sequences at the 3’ end of many LINEs. The presence of 28S downstream sequences in the R2 transcript could yield accurate R2 target- 3’ junction in reactions *in vitro* as found in endogenous R2 of many insects [51]. While non-specific RNA could position the R2 protein at the insertion site, they were unable to initiate TPRT [3]. 3’ PBM corresponding to the 3’ UTR region was required and sufficient for these steps of the integration mechanism.

In APE LINEs, the ORF2p encoded EN interacts with DNA via the Bβ6- Bβ5 loop and makes a single-strand endonucleolytic nick in genomic DNA at a degenerate consensus sequence (e.g., 5′-TTTT/A-3′ or variants of that sequence) [85,145]. The resultant 3′ -OH group at the nick site is used by the ORF2 RT to prime L1 first strand synthesis, which generally begins within the L1 RNA 3′ poly(A) tail [151]. The L1 reverse transcriptase has shown to act in *cis,* and preferentially reverse transcribe their encoding transcript [93]. The 3′ poly(A) tract in L1 RNA is required for efficient L1- mediated retrotransposition *in vitro*. When the L1 RNA polyadenylation signal is replaced by sequences derived from a non-polyadenylated long noncoding RNA, the RNA could stably accumulate in cells and could be translated, however, they were unable to retrotranspose in *cis* [152]. Addition of 26 to 20 poly(A)tract downstream of this construct increased their retrotransposition. Also, the poly(A) is found to be essential to allow ORF2 to bind and mobilize RNA in *trans* [152]. However, poly(A) tail is shown to be not required for L1 retrotransposition in assays using cultured cells [56].

Retrotransposition assay using cell lines lacking one or more components of the cellular non-homologous end joining (NHEJ) mechanism, a principal form of DNA double-strand break (DSB) repair, has suggested endonuclease independent L1 retrotransposition [153–155]. Double strand break created in the DNA could initiate L1 reverse transcription.

1.12 Second strand cleavage and second strand synthesis

For R2 RLE LINEs, at the end of first strand synthesis, the 5’ PBM bound to the downstream subunit is removed by the reverse transcriptase of upstream subunit which triggers second strand cleavage 2 bp upstream relative to the first strand cleaved site and is catalyzed by the endonuclease of the downstream subunit (Figure 1-4 step 3) [31]. However, it is unclear how a 'site-specific' DNA endonuclease cleaves two different sites on the target DNA specifically. Also, it is unclear if second strand cleavage requires a protein dimer. After second strand cleavage occurs, the reverse transcriptase of downstream subunit is thought to synthesize the second strand thus completing element integration (Figure 1-4 step 4). In *Bombyx mori* and many other animals, a uniform R2 5' junctions were observed within species, and in some cases, a particular length of the 28S gene sequence was found to be duplicated at the 5' junction [156]. However, in species of Drosophila, there is a variable addition of non-templated nucleotides and variable deletions of the 28S target site at the R2 5' junctions [157]. This difference in uniformity at 5' junctions could be explained by whether the 5' end of the RNA processed form 28S co-transcript had the 28S gene sequence or not [7]. If the RNA template contained the 28S sequence, the first strand cDNA generated by reverse transcription could anneal to the top strand to allow precise priming for second strand synthesis, and hence uniform 5' junctions. If the RNA does not have the 28S sequences, then random 3-5 non-templated nucleotides are added to the cDNA and the varying microhomologies between the added nucleotides and the upstream target DNA sequences are used to prime second strand synthesis [7]. This gives rise to different length deletions of 28S sequences and/or non-templated nucleotide additions observed in some species.

Analysis of L1 elements that are flanked by target site duplication (TSD) did not show strict consensus second strand cleavage site, however weak preference was observed for the sequence 5′-TYTN/R in few inversion/deletion and inversion/duplication events in L1s [158]. The formation of inversion/duplication product could be explained by an alternative form of L1 retrotransposition called as "Twin priming" [159]. During this process, the second strand cleavage occurs before the completion of TPRT which generates an overhang that anneals primes the second strand synthesis. While the 3' poly(A) tail end of L1 RNA is getting templated for the first strand synthesis, internal region towards the 5' region of the same RNA anneals to the generated top strand overhang and gets templated for second strand synthesis. When the RNA is removed, the single-stranded cDNA pairs at the region of microhomology to complete remaining DNA synthesis [159]. This explains the L1 inversion flanked by TSD. Sequence analysis of L1 (L1Ta) insertions shows that only about 30-35% of them are full-length insertions [160]. About 40-45% of the insertions were truncated at the 5' ends and about 25% showed inversion/deletion events. The 5' truncations could be because of L1 reverse transcriptase falling off the RNA template before the complete strand synthesis [161]. The RNA template could also be degraded by cellular enzymes that could result in truncated copies. Details in the different steps of L1 integration requires further investigation.

1.13 Regulation of LINEs

Although many LINEs are parasitic and have evolved a mechanism to limit host genome damage, their mobility still poses a potential threat to the host. Host organisms have evolved mechanisms to combat LINE activity, however, the host must be able to discriminate LINE sequences from host genes. R2 elements are controlled by transcriptional repression rather than by post-transcriptional or post-translational degradation. Out of the hundreds of rDNA units present in the rDNA loci, host activates only ~40 rDNA units that contain the fewest R2 insertions [7,162,163]. The remaining rDNA units are inactivated by heterochromatin formation. If the area that is activated for transcription contains R2 insertions, then R2 transcripts are produced. However, if a region free from R2 insertions are unavailable to be transcriptionally activated, the cell is required to include the R2 inserted units in their transcription domain. New copies of R2 by retrotranspositions are formed by the transcribed rDNA unit with R2 insertions [7]. It is not known if the host can differentiate between the inserted and un-inserted rDNA units or not. However, it is speculated that the small RNA silencing pathway that induces heterochromatin formation of the R2 inserted units could possibly be a likely identifying mechanism [164,165]. Also, crossovers that are characteristic to the rDNA locus could change the boundary of transcriptionally active rDNA blocks to exclude or include the expression of R2 inserted units [166–168]. The long-term stability of R2 in rDNA loci is the ability of R2 to reside in the rDNA units outside of the transcriptionally active domain for many generations.

Cytosine methylation (5-methylcytosine) is an important DNA modification in eukaryotes that are correlated with the transcriptional repression of LINEs in somatic and germline cells. Methylation of CpG sites inactivates the LINE promoter and is one of the mechanisms involved in suppression of LINE-1 expression in a variety of cells [169]. Deletion of the cytosine-5-methyltransferse 3-like gene (Dnmt3L) in mice leads to loss of *de novo* cytosine methylation in LINEs, which has led to reactivation of spermatocytes and spermatogonia [170]. 5-hydroxymethylation of cytidine which is a mark in L1 5' UTR in pluripotent cells, could also be involved in regulation of L1 transcription [171]. Members of the human APOBEC3 gene family (APOBEC3A (A3A) and APOBEC3B (A3B)) catalyze the deamination of cytidine to uridine residues in single-strand DNA substrates several members of the APOBEC3 family and robustly inhibit LINE-1 and Alu retrotransposition in cultured cells [172–175]. PIWI proteins that belong to Argonaute family of proteins bind 26-31 nucleotide piwi-interacting RNAs (piRNA). The piRNA directs the PIWI protein to cleave the LINE transcript thus forming a defense system against LINEs. Several RNA based mechanism to control LINEs have also been implicated [176]. 5' UTR of L1 LINEs includes an antisense promoter that results in antisense RNA which binds to the L1 mRNA to form a substrate for siRNA biogenesis and the siRNA created cleaves target L1 transcripts [177]. 5' UTR of L1 also harbors binding sites for *cis*-acting transcription factors like Runx3, Sp1, and SRY-related (Sox) proteins, and host factors can recognize these sites to control L1 expression. Recent studies have revealed 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 L1 and SVA LINEs, which, in turn, inhibits their expression in embryonic stem cells [175,178].

1.14 Gap in the knowledge and dissertation work

Characterization of reverse transcriptase and endonuclease domain of LINEs have yielded important insights into the understanding of retrotransposition. However, a complete understanding of integration mechanism is yet to be achieved. Functional investigation of additional motifs/domains within the ORF structure of the protein encoded by both RLE and APE LINEs needs to be done to uncover the details of TPRT mechanism. R2Bm RLE LINE is extensively used as a LINE model system, and have facilitated a great part of the *in vitro* biochemical studies to advance the knowledge of RLE LINE integration, and by analogy, also of APE LINE integration. The first half of the integration mechanism that involves first strand cleavage and first strand synthesis by TPRT is well understood, but the second half of the integration remains obscure and uncharacterized. Second strand cleavage is very anemic and can only be achieved in a very narrow window of protein-RNA-DNA ratios in our *in vitro* assays. Second strand synthesis has not yet been biochemically demonstrated *in vitro* and we just speculate that the reverse transcriptase from the downstream subunit catalyzes second strand synthesis.

The functional role of the overall R2Bm ORF structure is not fully comprehended. While the ZF and Myb domains in N-terminal region of R2 protein have shown to bind downstream of R2 target DNA insertion site [42], the DNA binding motifs responsible for upstream binding is yet to be identified. It is also unclear which regions of the protein coordinate the activity of reverse transcriptase and endonuclease domain at each step of the integration mechanism which would possibly modulate protein's conformational change. The function of the conserved zinc knuckle CCHC motif and the upstream α-helices, located in the linker region between RT and RLE, is unknown. The zinc knuckle is universally conserved in all LINEs, and so is the upstream predicted α-helices. It is likely that these two motifs provide at least some of the missing nucleic acid binding and protein structure coordination functions in LINEs.

A highly basic region immediately before RT domain including the RT-1 and RT0 motifs have been recently identified to be involved in RNA binding [18], however, it is unknown how the protein differentiates between 3' and 5' PBM RNA. It is also not clear how the same residues could be involved in binding two distinct RNA regions. Given the different secondary structures at the 3' and 5' regions of RNA and the different conformational changes they induce when bound to the protein [31]; we speculate that there must be different residues in the protein that are involved in binding 3' and 5' PBM RNA with some extent of overlap. The specific 3' and 5' PBM binding residues of the R2 transcript remains to be identified. Based on the sequence similarity of reverse transcriptase domain, the telomerase and group II introns are speculated to share a close evolutionary relationship with LINEs [90,179]. RNA binding domain for telomerase and group II introns are also located upstream of the RT domain, and the 0 motif of R2 and group II intron bear sequence identity [90,180–182]. Mobile group II introns and Telomerase have their RNA binding residues limited not only to N-terminal regions but also involve residues in the finger, palm and thumb regions of Reverse transcriptase for complete element integration [180,182]. We believe that a complete picture of domains involved in RNA binding is not yet known for RLE LINEs.

Chapter 2 of my dissertation addresses the unknown aspects of second half of the integration reaction. A 4-way junction intermediate DNA structure has been identified that leads to second strand cleavage followed by successful second strand synthesis in our *in vitro* assays. The reverse transcriptase from the downstream protein subunit has been shown to primarily catalyze second strand synthesis. The findings from this study had led to an updated R2 integration model.

Chapter 3 of my dissertation work focuses on identifying the exact biochemical function of the universally conserved α-helices and zinc knuckle of the linker region. The functional relevance of the linker region in terms of integration mechanism of LINEs have been explained. Double point mutations were generated in the highly conserved amino acid residues across the linker region. The effects of these mutations were analyzed through a series of biochemical assays. Both the α-helices and zinc knuckle were identified to be essential for accurate nucleic acid positioning and protein conformational change required for effective reverse transcriptase and endonuclease activity at different steps of integration mechanism. In addition, the α-helix was found to be crucial in binding an intermediate

DNA structure required to proceed to the second half of the integration mechanism. The results could be implied not only to other RLE LINEs but also to APE LINEs given their universal presence.

Chapter 4 and Chapter 5 includes the results of mutational studies carried out to probe the nucleic acid binding function of conserved residues in the R-box region and the N-terminal RT-1 region, respectively. Two motifs are already identified in the N-terminal region to be involved in binding both 3' and 5' PBM RNA [18]. An additional residue in the RT-1 region is identified that distinguishes between the two structurally different PBM regions and preferentially bind to 3' PBM RNA. While R-box has been identified to be involved in DNA binding [72], I have assayed the function of two highly conserved R-box arginine residue which are speculated to participate in DNA binding function. Chapter 4 includes conclusions of my dissertation work along with limitations and future directions.

1.15 References

1. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
2. Arkhipova, I. R. & Morrison, H. G. Three retrotransposon families in the genome of Giardia lamblia: two telomeric, one dead. *Proc. Natl. Acad. Sci. U. S. A.* (2001). doi:10.1073/pnas.231494798
3. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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).
4. Singer, M. F. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell* (1982). doi:10.1016/0092-8674(82)90194-5
5. Kajikawa, M. & Okada, N. LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell* (2002). doi:10.1016/S0092-8674(02)01041-3
6. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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).
7. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol. Spectr.* (2015). doi:10.1128/microbiolspec.MDNA3-0011-2014
8. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol. Spectr.* (2014). doi:10.1128/microbiolspec
9. Eickbush, T. H. & Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* **134,** 221–34 (2008).
10. Gilbert, N. & Moran, J. V. in *Mobile DNA II* 836–869 (American Society of Microbiology, 2002). doi:10.1128/9781555817954.ch35
11. Zingler, N., Weichenrieder, O. & Schumann, G. G. APE-type non-LTR retrotransposons: Determinants involved in target site recognition. *Cytogenetic and Genome Research* (2005). doi:10.1159/000084959
12. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* (2002). doi:10.1016/S0092-8674(02)00828-0
13. Symer, D. E. *et al.* Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* (2002). doi:10.1016/S0092-8674(02)00839-5
14. Kapitonov, V. V., Tempel, S. & Jurka, J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* (2009). doi:10.1016/j.gene.2009.07.019
15. Kojima, K. K. & Fujiwara, H. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res.* **15,** 1106–1117 (2005).
16. Kapitonov, V.V. and Jurka, J. RandI-1, a family of RandI non-LTR retrotransposons from the Chlamydomonas reinhardtii genome. *Repbase Reports* **4,** 196 (2004).
17. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol. Biol.*

Evol. **22,** 2157–2165 (2005).

18. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res.* **42,** 8405–8415 (2014).
19. Malik, H. S. & Eickbush, T. H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. *Genetics* (2000).
20. Zingler, N., Weichenrieder, O. & Schumann, G. G. APE-type non-LTR retrotransposons: Determinants involved in target site recognition. *Cytogenet. Genome Res.* **110,** 250–268 (2005).
21. Mizrokhi, L. J., Georgieva, S. G. & Ilyin, Y. V. jockey, a mobile drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. *Cell* (1988). doi:10.1016/S0092-8674(88)80013-8
22. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* (1990). doi:10.1128/MCB.10.12.6718.Updated
23. Minakami, R. *et al.* Identification of an internal cis-element essential for the human li transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res.* (1992). doi:10.1093/nar/20.12.3139
24. George, J. A. & Eickbush, T. H. Conserved features at the 5' end of Drosophila R2 retrotransposable elements: Implications for transcription and translation. *Insect Mol. Biol.* (1999). doi:10.1046/j.1365-2583.1999.810003.x
25. Hohjoh, H. & Singer, M. F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.* **15,** 630–9 (1996).
26. Martin, S. L. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. *Mol. Cell. Biol.* **11,** 4804–7 (1991).
27. Dmitriev, S. E. *et al.* Efficient Translation Initiation Directed by the 900-Nucleotide-Long and GC-Rich 5' Untranslated Region of the Human Retrotransposon LINE-1 mRNA Is Strictly Cap Dependent Rather than Internal Ribosome Entry Site Mediated. *Mol. Cell. Biol.* **27,** 4685–4697 (2007).
28. Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* **21,** 1429–39 (2001).
29. Kubo, S. *et al.* L1 retrotransposition in nondividing and primary human somatic cells. *Proc. Natl. Acad. Sci.* (2006). doi:10.1073/pnas.0601954103
30. Christensen, S. M. & Eickbush, T. H. 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).
31. Christensen, S. M., Ye, J. & Eickbush, T. H. 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).
32. Piskareva, O. & Schmatchenko, V. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. *FEBS Lett.* (2006). doi:10.1016/j.febslet.2005.12.077
33. Yang, J. & Eickbush, T. H. RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol. Cell. Biol.* **18,** 3455–65 (1998).
34. Han, J. S. & Boeke, J. D. LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression? *BioEssays* (2005). doi:10.1002/bies.20257
35. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: Similarities to eukaryotic splicing factor Prp8. *Mob. DNA* **8,** 1–15 (2017).
36. Berg, J. M. & Shi, Y. The galvanization of biology: A growing appreciation for the roles of zinc. *Science (80-.).* (1996). doi:10.1126/science.271.5252.1081
37. Ogata, K. *et al.* Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* (1994). doi:10.1016/0092-8674(94)90549-5
38. Burke, W. D., Malik, H. S., Jones, J. P. & Eickbush, T. H. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* **16,** 502–11 (1999).
39. T.H. Eickbush. in *Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. Mobile DNA II* 813–35 (ASM Press, 2002).
40. Thompson, B. K. & Christensen, S. M. Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons. *Mob. Genet. Elements* **1,** 29–37 (2011).
41. Shivram, H., Cawley, D. & Christensen, S. M. 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 (2011).

42. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).
43. Li, S. J. & Hochstrasser, M. A new protease required for cell-cycle progression in yeast. *Nature* (1999). doi:10.1038/18457
44. Kierzek, E. *et al.* 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.* (2008). doi:10.1093/nar/gkm1085
45. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
46. Clements, A. P. & Singer, M. F. The human LINE-1 reverse transcriptase: Effect of deletions outside the common reverse transcriptase domain. *Nucleic Acids Res.* (1998). doi:10.1093/nar/26.15.3528
47. Bibillo, A. & Eickbush, T. H. High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. *J. Biol. Chem.* **277,** 34836–34845 (2002).
48. Bibillo, A. & Eickbush, T. H. End-to-End Template Jumping by the Reverse Transcriptase Encoded by the R2 Retrotransposon. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M310450200
49. Bibiłło, A. & Eickbush, T. H. The reverse transcriptase of the R2 non-LTR retrotransposon: Continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* (2002). doi:10.1006/jmbi.2001.5369
50. Luan, D. D. & Eickbush, T. H. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell. Biol.* **15,** 3882–91 (1995).
51. Luan, D. D. & Eickbush, T. H. 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.* (1996). doi:10.1128/MCB.16.9.4726
52. Wu, J., Liu, W., Gong, P. & Gong, P. A structural overview of RNA-dependent RNA polymerases from the Flaviviridae family. *International Journal of Molecular Sciences* (2015). doi:10.3390/ijms160612943
53. Lu, G. & Gong, P. A structural view of the RNA-dependent RNA polymerases from the Flavivirus genus. *Virus Research* (2017). doi:10.1016/j.virusres.2017.01.020
54. Thompson, A. A. & Peersen, O. B. Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *EMBO J.* (2004). doi:10.1038/sj.emboj.7600357
55. Kajikawa, M., Ohshima, K. & Okada, N. 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.* (1997). doi:10.1093/oxfordjournals.molbev.a025730
56. Moran, J., Holmes, S. & Naas, T. High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
57. Takahashi, H. & Fujiwara, H. Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* (2002). doi:10.1093/emboj/21.3.408
58. Doucet, A. J. *et al.* Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet.* **6,** 1–19 (2010).
59. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3,** 433–437 (2013).
60. Bertram, K. *et al.* Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation Article Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. 701–713 (2017). doi:10.1016/j.cell.2017.07.011
61. Wan, R. *et al.* The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science (80-.).* (2016). doi:10.1126/science.aad6466
62. Nguyen, T. H. D. *et al.* Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530,** 298–302 (2016).
63. Yan, C., Wan, R., Bai, R., Huang, G. & Shi, Y. Structure of a yeast step II catalytically activated spliceosome. *Science (80-.).* (2017). doi:10.1126/science.aak9979
64. Bai, R., Yan, C., Wan, R., Lei, J. & Shi, Y. Structure of the Post-catalytic Spliceosome from Saccharomyces cerevisiae. *Cell* **171,** 1589–1598.e8 (2017).
65. Shi, Y. The Spliceosome: A Protein-Directed Metalloribozyme. *J. Mol. Biol.* **429,** 2640–2653 (2017).
66. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. Type II restriction endonucleases: Structure and mechanism. *Cellular and Molecular Life Sciences* **62,** 685–707 (2005).
67. Pingoud, V. *et al.* 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.* (2005). doi:10.1074/jbc.M409020200
68. Pingoud, V. *et al.* Evolutionary relationship between different subgroups of restriction endonucleases. *J. Biol.*

Chem. **277,** 14306–14 (2002).

69. Nishino, T., Komori, K., Ishino, Y. & Morikawa, K. Dissection of the Regional Roles of the Archaeal Holliday Junction Resolvase Hjc by Structural and Mutational Analyses. *J. Biol. Chem.* (2001). doi:10.1074/jbc.M104460200

70. Kvaratskhelia, M., Wardleworth, B. N., Norman, D. G. & White, M. F. A conserved nuclease domain in the archaeal Holliday junction resolving enzyme Hjc. *J. Biol. Chem.* (2000). doi:10.1074/jbc.M003420200

71. Middleton, C. L., Parker, J. L., Richard, D. J., White, M. F. & Bond, C. S. Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh869

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

73. Yang, J., Malik, H. S. & Eickbush, T. H. 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,** 7847–52 (1999).

74. Pingoud, A., Wilson, G. G. & Wende, W. Type II restriction endonucleases - A historical perspective and more. *Nucleic Acids Research* (2014). doi:10.1093/nar/gku447

75. Kulpa, D. A. & Moran, J. V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum. Mol. Genet.* **14,** 3237–3248 (2005).

76. Moran, J. V, DeBerardinis, R. J. & Kazazian, H. H. Exon shuffling by L1 retrotransposition. *Science* **283,** 1530–1534 (1999).

77. Khazina, E. *et al.* Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **18,** 1006–1014 (2011).

78. Khazina, E. & Weichenrieder, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc. Natl. Acad. Sci.* (2009). doi:10.1073/pnas.0809964106

79. Januszyk, K. *et al.* 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.* (2007). doi:10.1074/jbc.M702023200

80. Matsumoto, T., Hamada, M., Osanai, M. & Fujiwara, H. Essential Domains for Ribonucleoprotein Complex Formation Required for Retrotransposition of Telomere-Specific Non-Long Terminal Repeat Retrotransposon SART1. *Mol. Cell. Biol.* **26,** 5168–5179 (2006).

81. Heras, S. R. *et al.* Nucleic-acid-binding properties of the C2-L1Tc nucleic acid chaperone encoded by L1Tc retrotransposon. *Biochem. J.* (2009). doi:10.1042/BJ20090766

82. Kajikawa, M., Sugano, T., Sakurai, R. & Okada, N. Low dependency of retrotransposition on the ORF1 protein of the zebrafish LINE, ZfL2-1. *Gene* (2012). doi:10.1016/j.gene.2012.02.048

83. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* (2003). doi:10.1038/ng1223

84. Olivares, M., Thomas, M. C., Alonso, C. & López, M. C. The L1Tc, long interspersed nucleotide element from Trypanosoma cruzi, encodes a protein with 3'-phosphatase and 3'-phosphodiesterase enzymatic activities. *J. Biol. Chem.* (1999). doi:10.1074/jbc.274.34.23883

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

86. Ergün, S. *et al.* Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M312985200

87. Kubo, Y., Okazaki, S., Anzai, T. & Fujiwara, H. Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. *Mol. Biol. Evol.* (2001). doi:10.1093/oxfordjournals.molbev.a003866

88. König, P., Fairall, L. & Rhodes, D. 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 Res.* (1998). doi:10.1093/nar/26.7.1731

89. Taylor, M. S. *et al.* Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* (2013). doi:10.1016/j.cell.2013.10.021

90. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* (1990). doi:10.1002/j.1460-2075.1990.tb07536.x

91. Evgen'ev, M. B. & Arkhipova, I. R. Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet. Genome Res.* (2005). doi:10.1159/000084984

92. Piskareva, O., Denmukhametova, S. & Schmatchenko, V. Functional reverse transcriptase encoded by the

human LINE-1 from baculovirus-infected insect cells. *Protein Expr. Purif.* (2003). doi:10.1016/S1046-5928(02)00655-1

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

94. Monot, C. *et al.* The Specificity and Flexibility of L1 Reverse Transcription Priming at Imperfect T-Tracts. *PLoS Genet.* (2013). doi:10.1371/journal.pgen.1003499

95. Fanning, T. & Singer, M. The line-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic Acids Res.* (1987). doi:10.1093/nar/15.5.2251

96. Kajikawa, M., Ohshima, K. & Okada, N. 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 (1997).

97. Dewannieux, M., Heidmann, T. & Yaniv, M. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J. Mol. Biol.* (2005). doi:10.1016/j.jmb.2005.03.068

98. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gkr863

99. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* (2000). doi:10.1038/74184

100. Jakubczak, J. L., Burke, W. D. & Eickbush, T. H. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 3295–3299 (1991).

101. Luchetti, A. & Mantovani, B. Non-LTR R2 Element Evolutionary Patterns: Phylogenetic Incongruences, Rapid Radiation and the Maintenance of Multiple Lineages. *PLoS One* (2013). doi:10.1371/journal.pone.0057076

102. Kapitonov VV, J. J. Non-LTR retrotransposons in the Acanthamoeba castellanii protist genome. *Repbase Reports* **9,** 1143–1143 (2009).

103. Xiong, Y. & Eickbush, T. H. Dong, a non-long terminal repeat (non-LTR) retrotransposable element from Bombyx mori. *Nucleic Acids Res.* (1993). doi:10.1093/nar/21.5.1318

104. Burke, W. D., Müller, F. & Eickbush, T. H. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res.* (1995). doi:5t0487 [pii]

105. Kapitonov VV, J. J. A family of HERO non-LTR retro- transposons from the Californian leech genome. *Repbase Reports 1* **4,** 311 (2014).

106. Aksoy, S., Williams, S., Chang, S. & Richards, F. F. SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. *Nucleic Acids Res.* (1990). doi:10.1093/nar/18.4.785

107. Gabriel, a *et al.* A rapidly rearranging retrotransposon within the miniexon gene locus of Crithidia fasciculata. *Mol. Cell. Biol.* (1990). doi:10.1128/MCB.10.2.615

108. Burke, W. D., Malik, H. S., Rich, S. M. & Eickbush, T. H. Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. *Mol. Biol. Evol.* **19,** 619–630 (2002).

109. Kojima, K. K. & Fujiwara, H. Cross-Genome Screening of Novel Sequence-Specific Non-LTR Retrotransposons: Various Multicopy RNA Genes and Microsatellites Are Selected as Targets. *Mol. Biol. Evol.* (2004). doi:10.1093/molbev/msg235

110. Garrett, J. E., Knutzon, D. S. & Carroll, D. Composite transposable elements in the Xenopus laevis genome. *Mol. Cell. Biol.* (1989). doi:10.1128/MCB.9.7.3018

111. Kojima, K. K. & Fujiwara, H. Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol. Biol. Evol.* (2003). doi:10.1093/molbev/msg031

112. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* (2015). doi:10.1186/s13100-015-0041-9

113. Eickbush, T. H. Transposing without ends: the non-LTR retrotransposable elements. *New Biol.* (1992).

114. Chen, J. H. *et al.* A 1.9 Å crystal structure of the HDV ribozyme precleavage suggests both lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry* (2010). doi:10.1021/bi100670p

115. Webb, C. H. T., Riccitelli, N. J., Ruminski, D. J. & Lupták, A. Widespread occurrence of self-cleaving ribozymes. *Science* (2009). doi:10.1126/science.1178084

116. Ruminski, D. J., Webb, C.-H. T., Riccitelli, N. J. & Lupták, A. Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes. *J. Biol. Chem.* **286,** 41286–95 (2011).

117. Eickbush, D. G., Burke, W. D. & Eickbush, T. H. Evolution of the R2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* (2013). doi:10.1371/journal.pone.0066441

118. Lavie, L., Maldener, E., Brouha, B., Meese, E. U. & Mayer, J. The human L1 promoter: Variable transcription

initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* (2004). doi:10.1101/gr.2745804

119. Babushok & Kazazian Jr. Progress in Understanding the Biology of the Human Mutagen LINE-1. *Hum. Mutat.* (2007). doi:10.1002/humu

120. Kozak, M. The scanning model for translation: An update. *Journal of Cell Biology* (1989). doi:10.1083/jcb.108.2.229

121. Athanikar, J. N., Badge, R. M. & Moran, J. V. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh698

122. Tchénio, T., Casella, J. F. & Heidmann, T. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res.* (2000). doi:10.1093/nar/28.2.411

123. Yang, N., Zhang, L., Zhang, Y. & Kazazian, H. H. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res.* (2003). doi:10.1093/nar/gkg663

124. Anzai, T., Osanai, M., Hamada, M. & Fujiwara, H. Functional roles of 3′-terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki347

125. Eickbush, D. G. & Eickbush, T. H. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA cotranscript. *Mol. Cell. Biol.* **30,** 3142–3150 (2010).

126. Kierzek, E. *et al.* 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 (2009).

127. Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev.* **20,** 210–224 (2006).

128. Li, P. W. L., Li, J., Timmerman, S. L., Krushel, L. A. & Martin, S. L. The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: Implications for retrotransposition. *Nucleic Acids Res.* (2006). doi:10.1093/nar/gkj490

129. Kojima, K. Eukaryotic translational coupling in UAAUG stop-start codons for the bicistronic RNA translation of the non-long terminal repeat retrotransposon SART1. *Mol. Cell. ...* (2005). doi:10.1128/MCB.25.17.7675

130. Ruschak, A. M. *et al.* Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* **10,** 978–987 (2004).

131. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3,** 1–16 (1997).

132. Hohjoh, H. & Singer, M. F. Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. *J. Mol. Biol.* (1997). doi:10.1006/jmbi.1997.1159

133. Goodier, J. L., Cheung, L. E. & Kazazian, H. H. Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt512

134. Mandal, P. K., Ewing, A. D., Hancks, D. C. & Kazazian, H. H. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. *Hum. Mol. Genet.* (2013). doi:10.1093/hmg/ddt225

135. Goodier, J. L., Mandal, P. K., Zhang, L. & Kazazian, H. H. Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. *Hum. Mol. Genet.* (2010). doi:10.1093/hmg/ddq048

136. Dutko, J. A., Kenny, A. E., Gamache, E. R. & Curcio, M. J. 5' to 3' mRNA Decay Factors Colocalize with Ty1 Gag and Human APOBEC3G and Promote Ty1 Retrotransposition. *J. Virol.* (2010). doi:10.1128/JVI.02477-09

137. Larsen, L. S. Z. *et al.* Ty3 Nucleocapsid Controls Localization of Particle Assembly. *J. Virol.* (2008). doi:10.1128/JVI.01814-07

138. Larsen, L. S. Z. *et al.* Ty3 capsid mutations reveal early and late functions of the amino-terminal domain. *J. Virol.* (2007). doi:10.1128/JVI.02207-06

139. Fuller, A. M., Cook, E. G., Kelley, K. J. & Pardue, M. Lou. Gag proteins of drosophila telomeric retrotransposons: Collaborative targeting to chromosome ends. *Genetics* (2010). doi:10.1534/genetics.109.109744

140. Rashkova, S., Athanasiadis, A. & Pardue, M.-L. Intracellular targeting of Gag proteins of the Drosophila telomeric retrotransposons. *J. Virol.* (2003). doi:10.1128/JVI.77.11.6376

141. Khazina, E. & Weichenrieder, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 731–6 (2009).

142. Callahan, K. E., Hickman, A. B., Jones, C. E., Ghirlando, R. & Furano, A. V. Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. *Nucleic Acids Res.* **40,** 813–827 (2012).

143. Christensen, S. & Eickbush, T. H. Footprint of the Retrotransposon R2Bm Protein on its Target Site before and after Cleavage. *J. Mol. Biol.* **336,** 1035–1045 (2004).

144. Cost, G. J. & Boeke, J. D. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* (1998). doi:10.1021/bi981858s

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

146. Priimägi, A. F., Mizrokhi, L. J. & Ilyin, Y. V. The Drosophila mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. *Gene* (1988). doi:10.1016/0378-1119(88)90197-7

147. Chaboissier, M. C., Finnegan, D. & Bucheton, A. Retrotransposition of the I factor, a non-long terminal repeat retrotransposon of Drosophila, generates tandem repeats at the 3' end. *Nucleic Acids Res.* (2000). doi:10.1093/nar/28.13.2467

148. Maita, N., Aoyagi, H., Osanai, M., Shirakawa, M. & Fujiwara, H. Characterization of the sequence specificity of the R1Bm endonuclease domain by structural and biochemical studies. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkm397

149. Maita, N., Anzai, T., Aoyagi, H., Mizuno, H. & Fujiwara, H. Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J. Biol. Chem.* **279,** 41067–41076 (2004).

150. Okazaki, S., Tsuchida, K., Maekawa, H., Ishikawa, H. & Fujiwara, H. Identification of a pentanucleotide telomeric sequence, (TTAGG)n, in the silkworm Bombyx mori and in other insects. *Mol. Cell. Biol.* (1993). doi:10.1128/MCB.13.3.1424

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

152. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol. Cell* **60,** 728–741 (2015).

153. Sen, S. K., Huang, C. T., Han, K. & Batzer, M. A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkm317

154. Morrish, T. A. *et al.* Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* (2007). doi:10.1038/nature05560

155. Teng, S. C., Kim, B. & Gabriel, A. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature* (1996). doi:10.1038/383641a0

156. Burke, W. D., Malik, H. S., Jones, J. P. & Eickbush, T. H. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* (1999). doi:10.1093/oxfordjournals.molbev.a026132

157. George, J. A., Burke, W. D. & Eickbush, T. H. Analysis of the 5′ junctions of R2 insertions with the 28S gene: Implications for non-LTR retrotransposition. *Genetics* (1996).

158. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Mol. Cell. Biol.* (2005). doi:10.1128/MCB.25.17.7780-7795.2005

159. Ostertag, E. M. & Kazazian H.H., J. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* (2001). doi:10.1101/gr.205701

160. Myers, J. S. *et al.* A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements. *Am. J. Hum. Genet.* (2002). doi:10.1086/341718

161. Ostertag, E. M. & Kazazian Jr, H. H. Biology of Mammalian L1 Retrotransposons. *Annu. Rev. Genet.* (2001). doi:10.1146/annurev.genet.35.102401.091032

162. Zhou, J., Eickbush, M. T. & Eickbush, T. H. A Population Genetic Model for the Maintenance of R2 Retrotransposons in rRNA Gene Loci. *PLoS Genet.* (2013). doi:10.1371/journal.pgen.1003179

163. Ye, J. & Eickbush, T. H. Chromatin Structure and Transcription of the R1- and R2-Inserted rRNA Genes of Drosophila melanogaster. *Mol. Cell. Biol.* (2006). doi:10.1128/MCB.01409-06

164. Girard, A. & Hannon, G. J. Conserved themes in small-RNA-mediated transposon control. *Trends in Cell Biology* (2008). doi:10.1016/j.tcb.2008.01.004

165. Senti, K. A. & Brennecke, J. The piRNA pathway: A fly's perspective on the guardian of the genome. *Trends in Genetics* (2010). doi:10.1016/j.tig.2010.08.007

166. Ohta, T. & Dover, G. a. Population genetics of multigene families that are dispersed into two or more chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* (1983). doi:10.1073/pnas.80.13.4079

167. Lyckegaard, E. M. & Clark, A. G. Evolution of ribosomal RNA gene copy number on the sex chromosomes of Drosophila melanogaster. *Mol. Biol. Evol.* (1991).

168. Zhang, X., Eickbush, M. T. & Eickbush, T. H. Role of recombination in the long-term retention of transposable elements in rRNA gene loci. *Genetics* **180,** 1617–26 (2008).

169. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics* **13,** 335–340 (1997).

170. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431,** 96–99 (2004).

171. Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat. Rev. Genet.* (2012). doi:10.1038/nrg3080

172. Chen, H. *et al.* APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Curr. Biol.* (2006). doi:10.1016/j.cub.2006.01.031

173. Hulme, A. E., Bogerd, H. P., Cullen, B. R. & Moran, J. V. Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene* (2007). doi:10.1016/j.gene.2006.08.032

174. Schumann, G. G. APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. *Biochem. Soc. Trans.* (2007). doi:10.1042/BST0350637

175. Wissing, S., Montano, M., Garcia-Perez, J. L., Moran, J. V. & Greene, W. C. Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. *J. Biol. Chem.* (2011). doi:10.1074/jbc.M111.251058

176. Malone, C. D. *et al.* Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. *Cell* (2009). doi:10.1016/j.cell.2009.03.040

177. Yang, N. & Kazazian, H. H. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.* **13,** 763–771 (2006).

178. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* (2014). doi:10.1101/gad.241661.114

179. Arkhipova, I. R., Pyatkov, K. I., Meselson, M. & Evgen'ev, M. B. Retroelements containing introns in diverse invertebrate taxa. *Nat. Genet.* **33,** 123–124 (2003).

180. Gu, S.-Q. *et al.* Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16,** 732–747 (2010).

181. Rouda, S. & Skordalakes, E. Structure of the RNA-Binding Domain of Telomerase: Implications for RNA Recognition and Binding. *Structure* **15,** 1403–1412 (2007).

182. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat. Struct. Mol. Biol.* **17,** 513–518 (2010).

Chapter 2

Completion of LINE Integration Involves a 4-way Branched DNA Intermediate[1]

Aruna Govindaraju, Brijesh Khadgi, Monika Pradhan, and Shawn M. Christensen

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

2.1 Abstract

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

2.2 Introduction

Long interspersed elements (LINEs) are an abundant and diverse group of autonomous transposable elements (TEs) that are found in eukaryotic genomes across the tree of life. LINEs also mobilize the non-autonomous short interspersed elements (SINEs). SINEs appropriate the protein machinery of LINEs to replicate. The movement of LINEs and SINEs have been implicated in progression to cancer and in genome evolution including modulating gene expression, genome rearrangements, DNA repair, and as a source of new genes. LINEs replicate by a process called target primed reverse transcription (TPRT) where the element RNA is reverse transcribed into DNA at the site of insertion using a nick in the target DNA to prime reverse transcription [1–3]. LINEs encode protein(s) that are used to perform the critical steps of the insertion reaction. LINE proteins bind their own mRNA, recognize target DNA, perform first-strand target-DNA cleavage, and perform TPRT. The proteins are also hypothesized to perform second-strand target-DNA cleavage and second-strand element-DNA synthesis, although the evidence for this is sparse [1–18]. The early branching clades of LINEs encode a restriction-like endonuclease (RLE) while the later branching LINEs encode an apurinic-apyrimidinic DNA endonuclease (APE) [19–22]. Both types of elements are thought to integrate through a functionally equivalent integration process [3,23–25].

Second-strand DNA cleavage has remained puzzling because the cleavage sites are generally not palindromic: The sequence around the second cleavage site is often unrelated to the sequence around the first strand site. In addition, the cleavages can produce blunt or staggered that lead to either a target site duplication or a target site deletion depending upon the stagger of the cleavage events for that element. The staggered cleavages can be a few bases away (e.g., 2 bp in R2Bm) or quite distant, e.g., 126 bp in R9 [26,27]. In APE LINEs, the cleavages are generally staggered such as to generate a modest 10-20 target site duplication upon insertion [28–30]. The endonuclease from APE bearing LINEs (APE LINEs) appears to have some specificity for the first DNA cleavage site but much less so for the second strand on a linear target DNA [21,28,29,31,32]. The endonuclease from the RLE bearing LINEs (RLE LINEs) is similarly involved in target site recognition [9]. In both cases, however, additional specifiers for cleavage have been invoked to account for the different specificity of the first and second strand cleavages including the endonuclease being tethered to the DNA by unidentified DNA binding domains in the protein. Another complicating factor is that the first cleavage event should occur in the presence of element RNA while the second cleavage event, according to *a priori reasoning*, should occur in the absence of element RNA, but this has been difficult to demonstrate *in vitro* [18].

Second-strand DNA synthesis has remained unresolved for over 20 years and it has never been directly observed *in vitro* [2,13,23,33,34]. Second-strand synthesis is hypothesized to be primed off of the free 3'-OH generated by the second-strand cleavage event and synthesized by the element encoded reverse transcriptase. It is unknown how the proposed primer-template association is generated as the target (ds)DNA ends drift away from each other post second strand DNA cleavage in *in vitro* reactions [4,18].

The R2 element from *Bombyx mori*, R2Bm, is one of a number of model systems that have been used to study the insertion reaction of LINEs [25]. R2 elements are site specific, targeting the "R2 site" in the 28S rRNA gene [25]. The R2 element encodes 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 *in vitro* reactions.

In vitro studies of the R2Bm protein and RNA have led to the current model of integration for R2Bm (Figure 1b) [18]. 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 thought to be involved in the integration reaction. The 5' and 3' PBM RNAs dictate the roles of the two subunits and coordinate a series of DNA cleavage and polymerization steps resulting in element integration by TPRT (Figure 1a). The protein subunit bound to the element's 3' PBM interacts with 28S rDNA sequences upstream of the R2 insertion site. The upstream subunit's RLE cleaves the first (bottom/antisense) DNA strand. After first-strand target-DNA cleavage, the subunit's RT performs TPRT using the 3'-OH generated by the cleavage event to prime first-strand cDNA synthesis. The protein subunit bound to the 5' PBM RNA interacts with 28S rDNA sequences downstream of the R2 insertion site by way of the ZF and Myb domains. The downstream subunit's RLE cleaves the second (top/sense) DNA strand. Second-strand DNA cleavage, however, is not thought to occur until after the 5' PBM RNA is pulled from the subunit, presumably by the process of TPRT, putting the protein in a "no RNA bound" conformation. Confusingly, second-strand DNA cleavage does not occur in the absence of RNA in our *in vitro* reactions. Second strand cleavage had, until this report, required a narrow range of R2 protein, 5' PBM RNA, and target DNA ratios to be observed [18]. Additionally, second-strand cleavage divorced the upstream target-DNA from the downstream target-DNA making initiation of second-strand DNA synthesis from the upstream target-DNA to the TPRT product attached to the downstream target-DNA problematic [4,18].

The DNA endonuclease plays a central role in the integration reaction of LINEs. The RLE found in the early branching LINEs is a variant of the PD-(D/E)XK superfamily of endonucleases [9,20]. In a previous paper, we had reported the similarity of the LINE RLE as having sequence and structural homology to archaeal Holliday junction resolvases [9]. 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 perform integration functions on branched DNAs is explored.

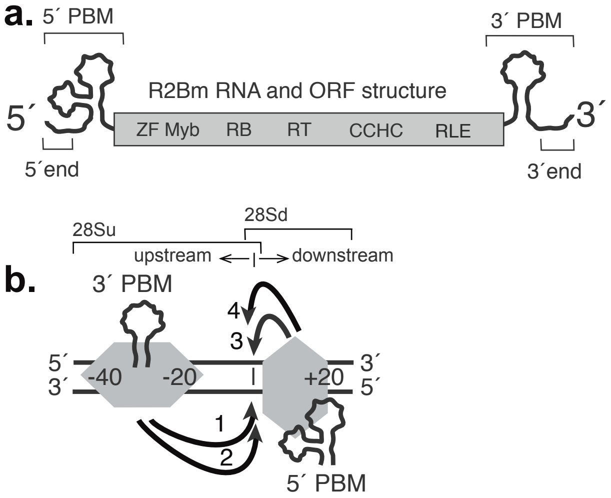


Figure 2- 1: R2Bm structure and integration reaction.

(a) R2Bm RNA (wavy line) and open reading frame (ORF) structure (gray box). The ORF encodes conserved domains of known and unknown functions: zinc finger (ZF), Myb (Myb), reverse transcriptase domain (RT), a cysteine-histidine rich motif (CCHC), and a PD-(D/E)XK type restriction-like endonuclease (RLE). RNA structures 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 used in this paper: 5′ PBM RNA (320 nt), 3′ PBM RNA (249 nt), RNA at the 5' end of the element (25 or 40 nt) and RNA 3' end (25 or 40 nt).

(b) The four-step integration model is depicted on a segment of 28S rDNA (black parallel lines). An R2 protein subunit (gray hexagon) is bound upstream of the insertion site (vertical bar) and an R2 protein subunit is bound downstream of the insertion site. The upstream subunit is associated with the 3' PBM RNA while the downstream subunit is associated with the 5' PBM RNA. The footprint of the protein subunits on the target DNA are indicated. The upstream footprints from -40 bp to -20 bp, but grows to just over the insertion site (verticle line) after first-strand DNA cleavage. The downstream subunit footprints from just prior to the insertion site to +20 bp [8,18]. The four steps of integration are: (1) DNA cleavage of the bottom/first-strand of the target DNA, (2) TPRT, (3) DNA cleavage of the top/second-strand of the target DNA, and (4) second strand DNA synthesis. The fourth step has not been directly observed *in vitro*. The overlapping portions of the target site used in this paper are indicated with brackets.

2.3 Results

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

Holliday junction resolvases bind to and symmetrically cleave 4-way DNA junctions (Holliday junctions), resolving the junctions into linear DNAs. Holliday junction resolvases recognize DNA structure rather than DNA sequence. The R2 RLE, which shares structural and amino acid sequence homology to Archael Holliday junction resolvases, may exhibit similar DNA binding and cleavage activities. 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). The linear and junction DNAs were formed by annealing complementary oligos. The linear and the junction DNA shared a common DNA oligo 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 DNA sequences to be probed. DNA binding was analyzed by electrophoretic mobility shift assay (EMSA). 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. 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 in the competition reaction as the R2 protein did not bind to the linear DNA until most of the junction DNA had been bound: A difference greater than two-fold.

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 [4,27].

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 (Supplemental File 1). Interestingly, in the presence of 5' PBM RNA, the behavior was different (see next section).

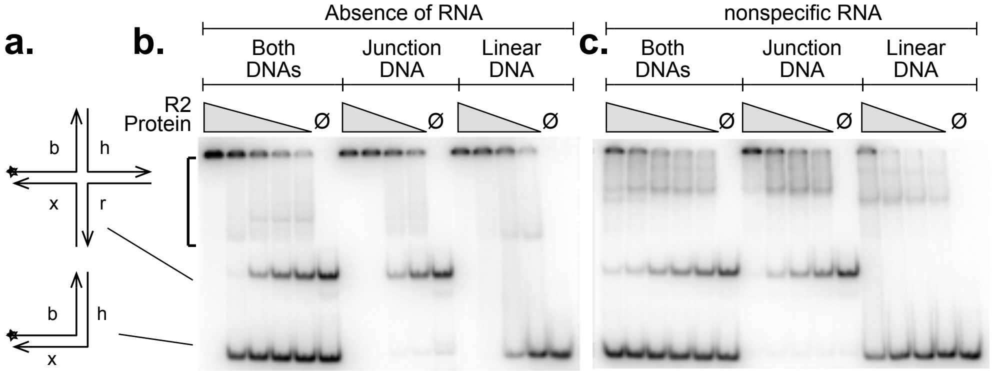


Figure 2- 2: R2 protein binds preferentially to a nonspecific (ns) 4-way DNA junction over a linear nsDNA.
(a) Diagrams of the nonspecific 4-way junction and linear DNA DNA constructs. The design and sequence of the 4-way junction was from and formed by annealing the b, x, h, and r DNA oligos. Each arm of the resulting junction was 25 bp. The linear DNA was generated by annealing oligo b to an oligo that was a combination of the x and h oligos. Thus junction and linear DNAs shared a common DNA oligo (oligo b). The shared DNA oligo was 5' end-labeled (star) with ^{32}P prior to formation and purification of the linear and junction DNAs.
(b) R2BM protein bound to 4-way nsDNA junction and to linear nsDNA, separately and in competition, in the absence of RNA. Gray triangles represent an R2Bm protein titration series. The bracket indicates the region of the gel where DNA, junction and linear, bound by protein migrate to. The DNA only (no R2Bm protein) lanes are marked with a Ø. The migration position of the unbound junction and linear DNAs are indicated.
(c) Same as in panel b, except that the binding reactions were in the presence of nonspecific RNA (nsRNA).

2.3.2 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 for nonspecific RNA, 3' PBM RNA, and 5' PBM RNA is reported in Figure 3. 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. The R2 RNAs were more efficient at pulling material out of the well and into the gel. A similar phenomenon is observed when R2 protein is bound to its normal (linear) 28S target DNA in the presence of R2 RNA [4,18,27]. Unlike binding to linear 28S target DNA, the presence of 5’ PBM RNA greatly inhibited the binding of R2 protein to the 4-way junction DNA. Only the presence of 5’ PBM RNA greatly affected the binding of R2 protein to junction DNA and inhibition scaled with 5’ PBM RNA concentration. Binding to nonspecific linear DNA and 3-way junction was less affected by the presence of 5’ RNA but still reduced in its presence (Supplemental Figure 2). This inhibition is not observed if downstream 28S rDNA sequences are present in any of the DNA constructs [8,18].

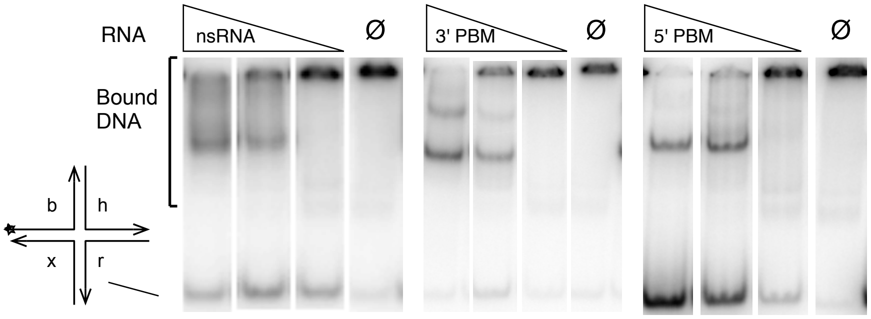


Figure 2- 3: The R2 5’ PBM RNA is inhibitory to R2 protein binding to a nonspecific 4-way DNA junction. EMSA gels showing a decrease in the ability of the R2 protein to bind to 4-way nsDNA junction in the presence of R2 5’ PBM RNA is presented. The migration positions of junction bound by protein as well as unbound junction in the EMSA are indicated. The DNA binding reactions where such that nearly 100% of the junction DNA was bound in the absence (Ø) of RNA. A titration (white triangles; 18 pmole, 675 fmole, and 75 fmole) of nsRNA, 3' PBM RNA, and 5' PBM RNA were added to the DNA binding reactions.

2.3.3 The R2 protein does not resolve nonspecific 4-way junction DNA

DNA from reactions of R2 protein bound to nonspecific linear and non-specific 4-way junctions across a range of protein concentrations in the absence of RNA were analyzed for DNA cleavage events by denaturing polyacrylamide gel electrophoresis (Supplemental Figure 3). Each strand of the junction and linear DNAs was tracked independently for DNA cleavage events by sequentially radiolabeling the 5’ ends of the different DNA strands. A complicated pattern of random low intensity background cleavages occurred, particularly in protein excess. A similar phenomenon of background cleavages occurs for R2 protein bound to its normal 28S target DNA in the absence of RNA when R2 protein is in excess. The background cleavages on the non-specific junction were not structure driven as the cleavages occurred in identical positions in the linear DNA of the same sequence. The presence of any of the three RNAs (5’ PBM RNA > 3’ PBM RNA > nonspecific RNA) abolished the random background DNA cleavage.

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

R2Bm inserts into a specific site in the 28S rDNA. In previous studies, it was determined that the protein subunit bound to target sequences downstream of the insertion site provides the endonuclease involved in second-strand (i.e., top-strand) DNA cleavage. Second-strand cleavage, however, has always been tricky to achieve and study. Second-strand cleavage has, until this paper, required a narrow range of 5’ PBM RNA, R2 protein, and DNA ratios. The prior data indicated that first-strand DNA cleavage is probably required before the second-strand can be cleaved and that the downstream subunit must be bound to the DNA (which required 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 be expected to pull the 5’ PBM RNA from the downstream subunit putting the downstream subunit into the "no RNA bound" state and thus initiating second-strand DNA cleavage.

Given the R2 protein is able to bind branched DNAs (Figures 2 and 3) in the absence of RNA, we investigated the role of DNA structure on the downstream subunit's ability to cleave DNA in the absence of RNA. The DNA constructs contained the binding site for the downstream R2 protein subunit but not binding site for the upstream-binding R2 protein subunit in order to isolate activities associated with the downstream subunit. The upstream DNA sequence was replaced by non-specific DNA derived from the 4-way junction used in the previous figures. Interestingly, linear DNAs containing downstream 28S DNA were not substrates for second strand cleavage regardless of the presence or absence of a first strand DNA cleavage event (Figure 4a, constructs iii, and iv). Neither was a post-TPRT analog (construct v) able to be cleaved by the R2 protein. The TPRT analog was a 3-way junction containing downstream 28S DNA that was precleaved at the first (bottom) strand cleavage site and covalently linked to cDNA sequences corresponding to the 3' end of the R2 element, as would be expected from a TPRT reaction. 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 unable to cleave the top-strand of these 3-way junctions. It did not matter if the R2 3' sequence containing arm was in the form of an RNA-DNA duplex or a DNA duplex.

2.3.5 Specific 4-way junction(s) are cleaved by R2 protein

Unlike the linear and TPRT-junction (Figure 4a, constructs iii-v) DNAs, a 4-way junction that included target sequence and R2 sequences was found to be cleavable by R2 protein (Figure 4a and 4b, construct viii). Construct viii was similar to the TPRT-junction (construct v) but with an additional arm: the 5' R2 arm. Both the R2 5' arm and the R2 3' arm were 25 bp in length and consisted of a RNA-DNA duplex. Construct viii mimics a hypothetical association between the cDNA and the target DNA. The 5' end of the R2Bm mRNA is believed to contain rRNA sequence corresponding to the upstream target DNA [33,35–37]. The reverse transcribed cDNA could then hybridize to the top strand of the target to form the 4-way junction. A completely covalently closed all DNA version of the same junction was also able to be cleaved, albeit to a lesser degree (see construct vi, Figure 4a and 4b) as was a construct lacking the R2 3' arm (construct vii).

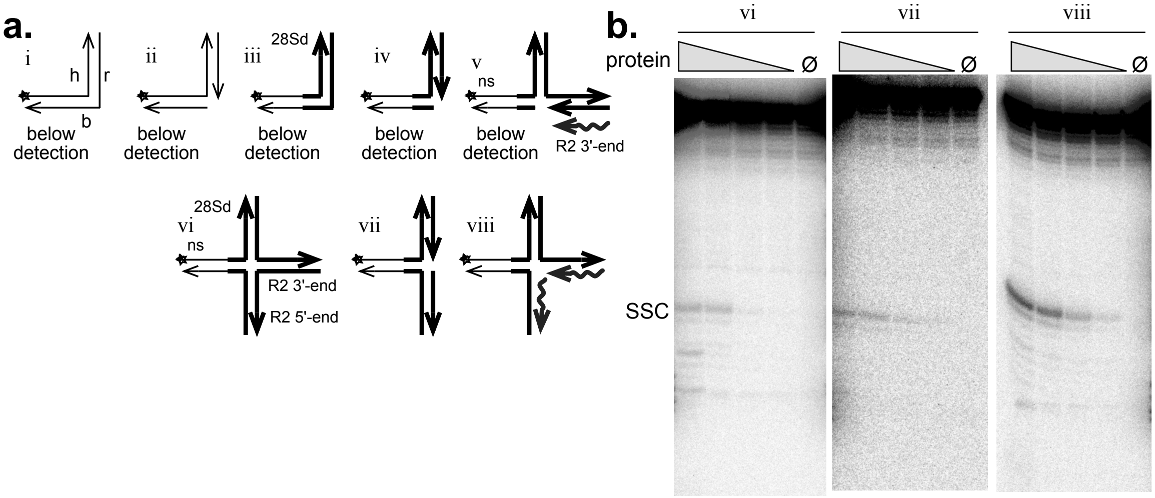


Figure 2- 4: A 4-way DNA junction involving 28S rDNA and R2 sequences is a substrate for second-strand DNA cleavage.
(a) Several linear, 3-way, and 4-way branched DNA constructs are diagramed. Straight lines represent DNA and wavy lines represent RNA. Thin lines represent non-specific DNA derived from the constructs depicted in Figure 2. Thick lines represent 28S rDNA as well as R2 element derived sequences. The R2 sequences are from the 5' and 3' ends of the element. The 28S sequence is the downstream DNA (28Sd) plus 7 bp of upstream DNA. The "arms" in each construct are 25 bp in length. Each construct is numbered for discussion purposes. The star indicates that the strand was end labeled as in previous figures. Two variations of construct v were tested, one having a DNA duplex in the R2 3' arm and the other having the RNA/DNA hybrid that would have been the result of TPRT. No detectable second-strand DNA cleavage was found on constructs i-v. Second-strand DNA cleavage was detectable on constructs vi-viii. **(b)** Denaturing gel analysis of DNA cleavage reactions on constructs vi, vii, and viii. The band resulting from second-strand cleavage (SSC) at the R2 site is indicated.

2.3.6 Further exploration of second-strand DNA cleavage

To further explore the structure requirements for second-strand cleavage, a number of structural-variants (i.e., partial-junctions) of Figure 4 construct viii were tested for cleavability (Figure 5a, constructs i-viii). Figure 4 construct viii is identical to Figure 5a construct i except that the 28S downstream arm was increased to 47 bp in length instead of the original 25 bp used in Figure 4 construct viii. This adjustment was to set the downstream DNA in the Figure 5a constructs equal to the amount of downstream DNA included in our historical linear DNA constructs used in previous publications [9]. The reason for testing the cleavability of partial junctions (Figure 5a, junctions ii-viii) was to determine to what extent, if any, the DNA cleavage signal observed in Figure 4 may have been coming from the minor, but present, contaminating partial junctions in the binding and cleavage reactions. It was also to determine if constructs mimicking cellular removal of the RNA component (e.g., by cellular RNases; construct vi-viii) faired better or worse at being cleaved by the R2 protein than constructs with intact RNA-DNA duplexes. It appears that several of the partial junctions (complexes ii and iii) can be cleaved and thus likely partially contribute the overall cleavage in reactions containing the full junction (complex i). The 4-way junction that lacked both RNA components (complex vi) was nearly uncleavable indicating the need for double stranded R2 arms. The 4-way junction that lacked the 5' end RNA but contained the 3' end RNA; construct vii) also failed to appreciably cleave indicating the importance of the presence a RNA-DNA duplex in R2 5' arm. The 4-way junction that lacks the 3' end RNA but contained the 5' end RNA (construct viii) cleaved well. Indeed, it was more efficiently cleaved than construct i indicating that the presence of duplex in the R2 3' arm is partially inhibitory but that the presence of duplex in the 5' arm is stimulatory.

In order to investigate the relative importance of upstream target sequences on second-strand DNA cleavage, 73 bp of upstream 28S DNA was incorporated into the 4-way junction Figure 5b; constructs ii-iv). In construct ii the 47 bp of downstream 28S DNA was replaced with nonspecific DNA and construct iii contained the full target DNA sequence (73 bp of upstream 28S DNA and 47 bp of downstream 28S DNA). Unexpectedly, construct ii was able to be cleaved, albeit much less efficiently that construct i which contained the downstream target DNA but not upstream as in previous figures. The fact that construct ii is able to be cleaved indicates that perhaps the 12 bp (7 bp of upstream and 5 bp of downstream DNA) common to both constructs i and ii might be involved in helping to direct DNA cleavage. Paradoxically, construct iii, which contains the full target sequence, was less efficient at being cleaved than even construct ii. Adding the flap, or displaced strand (construct iv), expected to occur during template jumping noticeably increased cleavability of the junction.

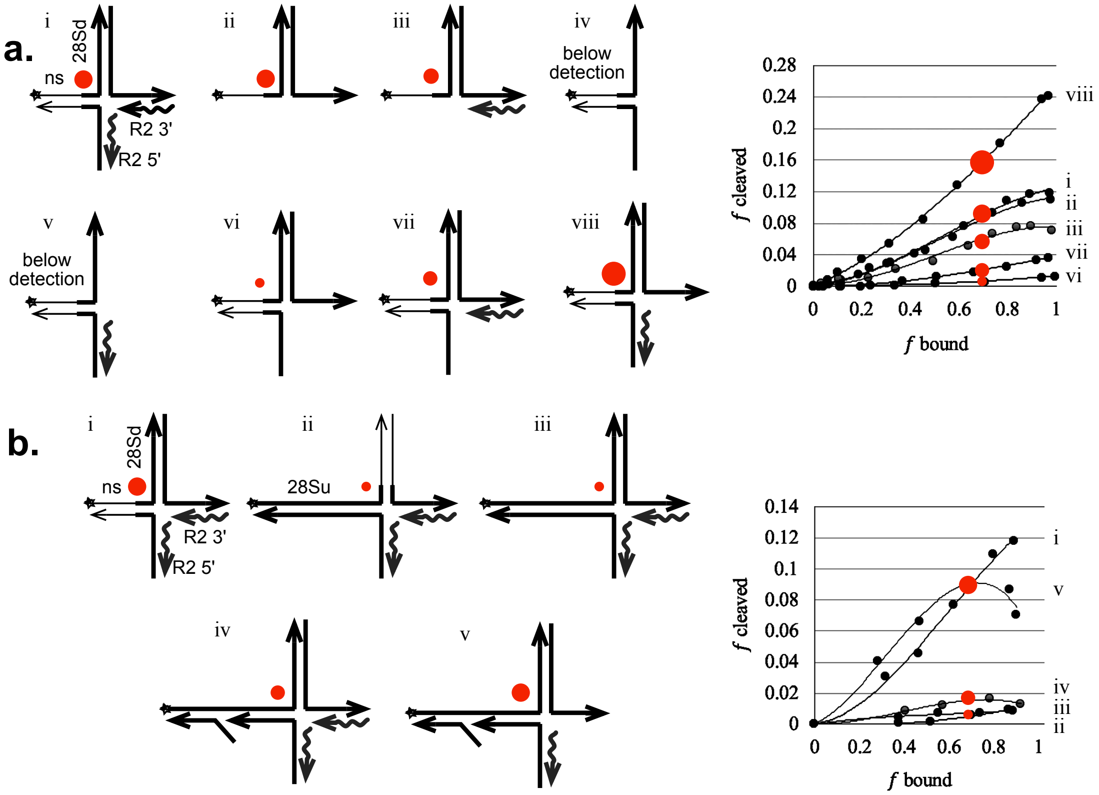


Figure 2- 5: Further exploration of second-strand DNA cleavage.
Various R2Bm/28S derived 4-way junctions were tested for DNA cleavage across a range of protein concentrations and analyzed by EMSA and denaturing gel electrophoresis. A diagram of each construct is given as a graph of the fraction cleaved (*f* cleaved) as a function of the fraction bound (*f* bound) for each set of constructs. All other abbreviations and symbols are as in other figures.
(a) Testing derivatives of the 4-way junction from Figure 4 to test for cleavage on partial junctions. The constructs have been numbered. The 28S downstream (28Sd) DNA arm was increased 47 bp so as to equal to the amount of downstream DNA historically used in our linear 28S target DNA [8,18]. Diameter of the red dot depicts relative cleavability of the construct by R2Bm. Abbreviations and symbols are as in previous figures.
(b) Constructs designed to test DNA cleavage on 4-way junctions that include upstream 28S DNA. The 28S upstream (28Su) DNA arm is 73 bp and corresponds to the amount of upstream DNA normally used in our linear target DNA [4,27]. Black lines are DNA with thin lines being non-specific DNA and thick line being either 28S or R2 derived DNA.

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

To test if second-strand cleavage could progress to second-strand synthesis dNTPs were added to the DNA cleavage reaction. The construct used to test for second-strand synthesis was construct i of Figure 5a as it cleaved relatively well and we had more experience with it compared to the other junctions presented in Figure 5, some of which had yet to be tested for DNA cleavage at the time. A range of R2 protein concentrations was used and the reactions were analyzed by denaturing (Figure 6a) and native (Figure 6b) polyacrylamide gel electrophoresis. The labeled strand of the 4-way junction was 72 nt uncleaved and 24 nt in length upon second-strand DNA cleavage (marked as SSC on the denaturing gel). Second-strand synthesis (SSS), i.e., extension of the labeled strand post DNA cleavage, would generate a 50 nt product when analyzed on a denaturing gel. Second-strand DNA synthesis was observed only at the higher end of the protein titration series in the denaturing gels (Figure 6a). The reason for this becomes obvious in the native (EMSA) gels (Figure 6b). Upon cleavage, the 4-way junction is resolved into two linear DNAs: one DNA containing the downstream and R2 3' arms and one DNA containing the "upstream" and R2 5' arms. The R2 protein appeared to remain bound to the DNA that contained the downstream 28S DNA after DNA cleavage while DNA with the DNA containing the non-specific "upstream" DNA was released. The release DNA primer-template is extended by the R2 RT only when protein is in excess. The migration positions of product of second-strand cleavage and second-strand synthesis are indicated next to the EMSA gels.

The signal above full length oligo on the denaturing gels 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 [38,39].

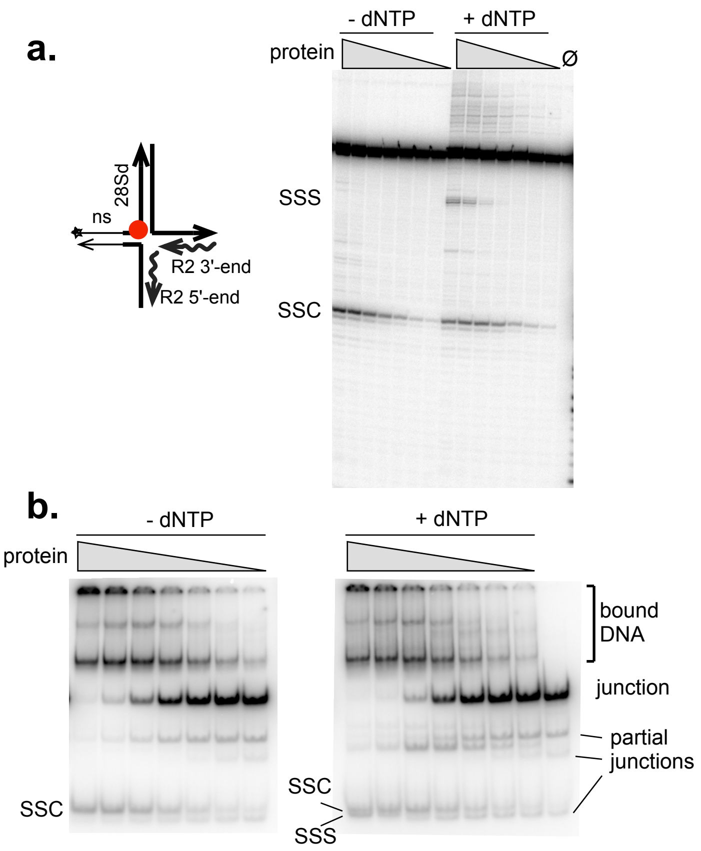


Figure 2- 6: Second-strand DNA cleavage followed by second-strand DNA synthesis.
(a) Diagram of the 4-way junction and denaturing gel analysis of DNA cleavage (-dNTP) and cleavage plus second-strand synthesis (+dNTP) reactions across a range of protein concentrations (gray triangle). The bands resulting from second-strand cleavage (SSC) and second-strand synthesis are marked. The junction used in the analysis is the same junction as Figure 5, construct i.
(b) The companion EMSA gels to the denaturing gel presented in panel a. Bands resulting from SSC and SSS are indicated.

2.3.8 Second-strand synthesis on precleaved DNA constructs

Although the primer-template is released from the protein-DNA complex when the upstream DNA is not present in the 4-way junction, one might expect that this would not occur *in vivo* with in a junction that contained the full target sequence. In part, this expectation is because it is hypothesized that the downstream subunit performs second-strand synthesis [4]. Unfortunately, junctions with full target sequence do not cleave well (Figure 5b) and second-strand synthesis is below the detection level when tested *in vitro*. For this reason, a post-second strand cleavage analog was generated. In order to keep the second-strand cleavage products tethered together, the R2 3' and 5' end "RNAs" were covalently linked, although instead of RNA DNA was used for convenience. The upstream 28S DNA containing second-strand cleavage product was able to undergo primer extension (i.e., second-strand synthesis) in the tethered configuration. The 5' end cDNA strand was used as the template (Figure 7a and 7b).

In order to determine which R2 protein subunit is used for second-strand synthesis, linear (Figure 7c, complexes iv and v) and tethered (Figure 7c, complexes i and iii) post-second strand cleavage products were tested for their relative ability to undergo second-strand synthesis (Figure 7d). The results are consistent with the downstream subunit being responsible for second-strand cleavage as complex iii was the most efficient substrate for second-strand synthesis and complex v is the least efficient substrate.

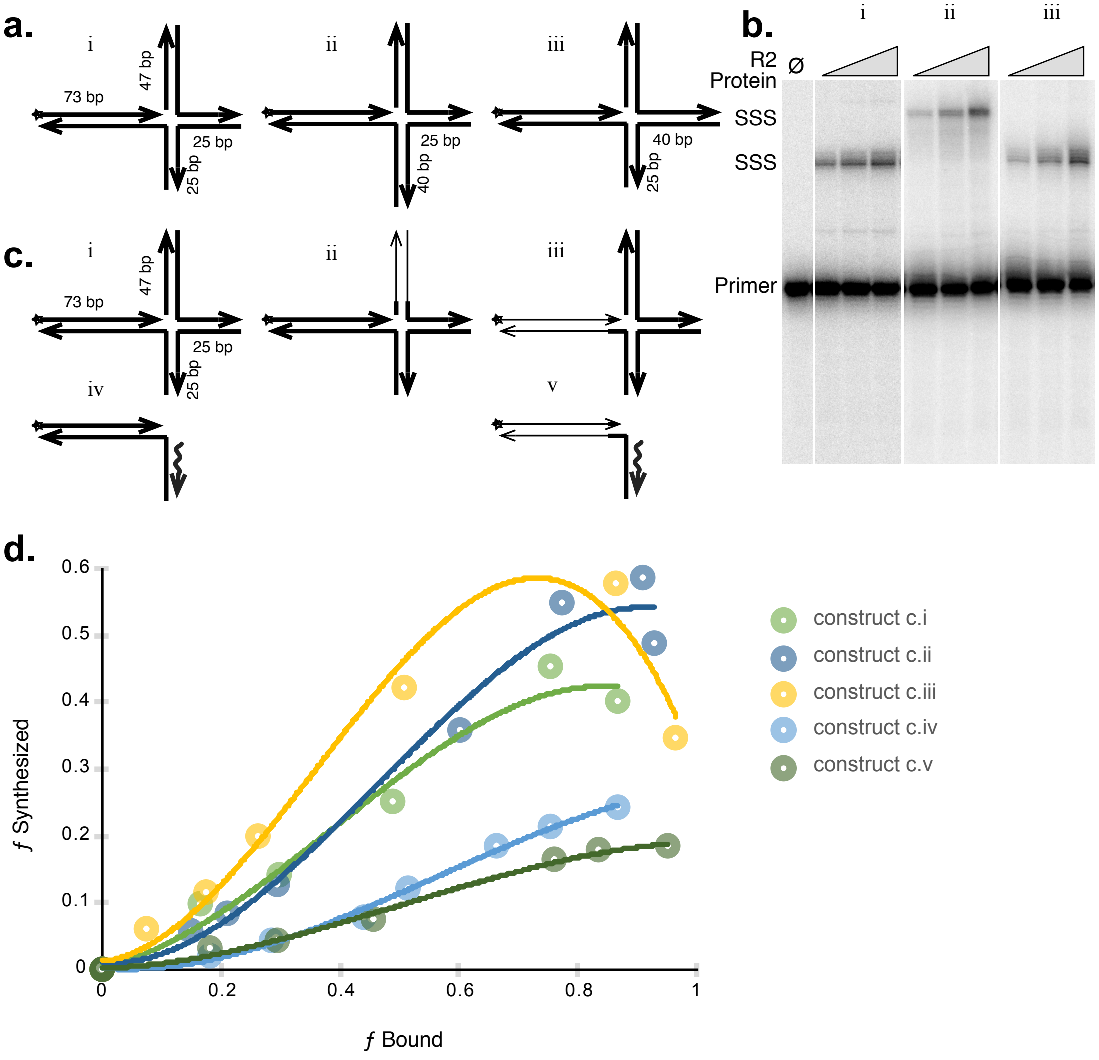


Figure 2- 7: Second-strand synthesis on precleaved DNA constructs.
(a) Constructs designed to hold the pre-cleaved products close proximity and to test which arm is used as a template. The length of 5' and 3' arms were varied (40 bp vs 25 bp). The 28S downstream arm was 47 bp and the 28S upstream

arm was 73 bp.
(b) Denaturing gel analysis of DNA synthesis reactions on the constructs presented in panel a.
(c) Constructs designed to test whether the upstream or the downstream protein subunit is likely responsible for second strand synthesis.
(d) A graph of the fraction synthesized (*f* synthesized) as a function of the fraction bound (*f* bound).

2.4 Discussion

2.4.1 R2 protein is not a general Holliday junction resolvase but does cleave its own integration intermediate in a resolvase-like reaction.

R2 protein was found to bind nonspecific 4-way DNA junctions, Holliday junctions, in preference to nonspecific linear DNA. The R2 protein appears to have a large surface for binding junction DNA when in the minus RNA conformation. This makes mechanistic sense in the context of R2 integration as it would be the minus RNA conformation of the R2 protein that would be expected to carry out second strand DNA cleavage. The presence of 5' RNA abolished binding to the nonspecific junction DNA (and nonspecific DNA in general). It is not known what part of the R2 protein binds the 4-way DNA junction, it may not be the endonuclease. It is also unknown whether the 5' PBM binding site overlaps the junction binding surface or if the lack of RNA promotes protein conformational changes that then reveal the junction binding surface. The binding surfaces for the 5' and 3' PBM RNAs are likely to be distributed across a large portion of the R2 protein, although currently the only identified RNA binding area is domain -1 and domain 0 [40]. The CCHC zinc-knuckle has also been hypothesized to bind to element RNA, but its true function remains unknown. It could be that the 5' PBM RNA forms a 4-way junction like mimic. The DNA binding surfaces of Holliday junction resolvases are large and highly positively charged, so it would make sense that R2 protein might make some use of this positive surface to bind help bind R2 RNA [41].

Although R2 binds to nonspecific DNA junctions in the absence of RNA, it was not able to subsequently resolve those junctions; DNA cleavages, particularly symmetrical DNA cleavages, did not occur. Therefore, R2 protein is not a Holliday junction resolvase in the strictest sense. However, with a more specific 4-way junction containing 28S rDNA and R2 sequences, the second/top-strand 28S rDNA cleavage event was nearly symmetrical with the bottom/first-strand cleavage that had been engineered into the 4-way junction. This DNA cleavage activity is very Holliday junction resolvase-like.

The presence of the template jump and the 5' arm being double stranded appeared to be the most important junction determinants, beyond the presence of target sequence in the downstream 28S rDNA arm, for cleavability. Interestingly, unless the R2 protein exists as a dimer in solution (of which there is no convincing evidence of), the bound versus DNA activity graph is linear and thus consistent with the endonuclease being monomeric [4,18]. If the second protein subunit was required for cleavage and the protein was not already dimeric, large amounts of protein would be expected to be needed to drive the second protein subunit, whose binding site is missing, into the protein-DNA complex for DNA cleavage to occur. Given this line of thought, however, it is difficult to envision how the 4-way junction lacking the much of the binding site for the downstream R2 subunit is cleaved, unless it is largely structure driven. The DNA sequence at the center of the junction also might be important, but the constructs tested do not address this prospect as all of the R2 specific junctions contained 5-7 bases of 28S sequence to either side of the insertion site. In addition, each junction contained at least 25 bp of R2 5' end sequence and 25 bp of R2 3' end sequence. The R2 3' arm appeared to be less important. Having the R2 3' arm duplexed was even inhibitory. Removal of the R2 3' arm, in an all DNA version was still cleavable, although only just. The presence of the first strand cleavage event appeared to also play a role in cleavability as a covalently closed all DNA version of the 4-way junction also had a difficult time being cleaved by R2 protein, although the lack of a RNA-DNA hybrids, especially in the 5' arm, may have contributed to the reduced cleavability.

The presence of a full target site in the 4-way junction was inhibitory towards DNA cleavage. There are several possible reasons for the reduced cleavability: (1) the upstream arm is competing for R2 protein subunits, (2) the presence of the upstream DNA causes protein subunits into suboptimal position or conformation to cleave, and (3) steric clash between subunits bound to both arms. When a displaced strand is included on the upstream 28S arm, cleavability is increased. The increase may be due to increase flexibility of the upstream arm because of the gap or the need for a specific stability/size of the invading strand segment. The presence of just a flap on the upstream (linear) DNA is not sufficient for driving DNA cleavage (unpublished data).

2.4.2 A new model for R2Bm Integration

The deeper understanding of the second half of the insertion reaction for R2Bm has allowed for an improved R2Bm integration model to be put forth (Figure 8A). 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 or recombination event from the 5’ end of the R2 RNA to the top-strand of the 28S rDNA upstream of the R2 insertion site forming a 4-way junction (step 3). It is this step that, to date, does not occur *in vitro* and may require host factors to form, if it exists at all. An association of the cDNA to the upstream target DNA is, however, consistent with a lot of previous data and a 4-way junction presents a simple unified mechanism for 5' junction formation, second strand DNA cleavage, and second strand DNA synthesis leading to full length element insertions.

The model makes sense of earlier *in vivo* experiments in which 'upstream' ribosomal RNA sequence attached to 5' end of the R2Bm element RNA had been noted as a requirement for full length element insertion [36,37]. More recently, bioinformatics and *in vitro* studies of the R2 RNA transcript have determined that R2 RNA is co-transcribed with ribosomal RNAs as part of the same large transcript [33,42]. The R2 RNA is then processed from bulk of the ribosomal RNA by an HDV-like ribozyme found near the 5' end of the R2 RNA [33,42]. For a number of R2 elements, however, the final processed R2 RNA retains some ribosomal RNA on the 5' end, 27 nt of ribosomal RNA in the case of R2Bm [33]. For elements that retain this much ribosomal RNA, the template jump may be more of a strand invasion or recombination event rather than a template jump [36,37]. For other R2 elements, however, the ribozyme leaves no ribosomal sequence on the processed R2 RNA (e.g., *Drosophila simulans* R2) and a template jump, as diagramed in Figure 8A, is envisioned to occur [14,33,35,39]. The RT of both APE LINEs and RLE LINEs has been shown to have the ability to jump from the end one template to the beginning of another without any homology [39]. Template jumps have long been hypothesized to be involved in 5' junction formation for both types of elements [14,33,35,39]. In addition to template jumping, LINE reverse transcriptases are able to use both DNA and RNA as a template during DNA synthesis and to displace a duplexed strand while polymerizing [14].

Recently the R2 RLE's reported similarity to Archaeal Holliday junction resolvases begged the question as to whether or not R2 can bind and cleave branched DNAs [9,43]. It turns out that the R2 protein can indeed bind to and cleave 4-way junctions in the absence of RNA. Second-strand DNA cleavage is step 4 in Figure 8A. Second-strand cleavage occurs across from first-strand cleavage on R2 specific 4-way junctions, a reaction reminiscent of Holliday junction resolvase. Second-strand cleavage is dependent on both structure and sequence as sequences from the immediate insertion site area and downstream of the insertion site helped to drive cleavage.

The south arm, i.e., the R2 5' arm, was a critical cleavage determinant. The presence of 5' PBM RNA prevents binding to non-specific 4-way junctions and prevents DNA cleavage of specific junctions. The R2 protein only cleaves in the absence of RNA. The three-way TPRT junction was not a good substrate for DNA cleavage.

For elements with rRNA sequences at the 5' end, like R2Bm, it is not clear what happens to the displaced

RNA strand from the heteroduplex or the displaced 'bottom strand' target DNA flap while the cDNA strand is forming the junction depicted in Figure 8A step 3, and what role, if any, the displaced strands plays in DNA cleavage. The displaced RNA was not included in the R2Bm integration 4-way junction constructs and the flap was non-specific DNA. In addition, it remains to be investigated as to whether or not the jump/recombination dislodges the upstream protein subunit as the 27 nt of ribosomal sequence encroaches on the minimal DNAse footprint observed of the upstream subunit when the subunit is bound to linear 28S rDNA [4,27]. The construct in Figure 5 that contained the full target sequence along with a displaced target DNA strand behaved much more like the junctions lacking upstream target sequences than did junctions with full target sequence and no displaced target DNA. The recombined cDNA/target DNA duplex was 27 bp in these constructs matching that expected for R2Bm [33].

The fifth and final line of evidence in support of the model is that cleavage of the 4-way junction generates natural primer-template for second-strand DNA synthesis. The 'downstream bound' subunit appears prime second-strand DNA synthesis (Figure 8A, step 5). *In vivo* host factors may help keep junction halves held together long enough to prime second-strand synthesis. *In vitro,* the primer-template is released, at least when the upstream target DNA arm consists of nonspecific DNA.

2.4.3 Extrapolating the R2 model to LINEs with different cleavage staggers.

The position of the second-strand DNA cleavage site relative to the first-strand cleavage site is quite variable across species even more so across the R2 clade. The stagger of the first and second DNA cleavage events in R2Bm is a small 5' overhang of 2 bp that leads to 2 bp target site deletion upon insertion of the element. In Drosophila, the R2 endonuclease produces blunt cleavages [35]. Other R2 elements produce small 3' overhangs. The model presented in Figure 8A works equally well for elements with any of these small staggers. The model can be adapted for elements with moderate 3' overhang staggers by hypothesizing a local melting or displacement of the TSD region followed by template switch to generate the 4-way junction. APE LINEs tend to produce a moderate 3' overhanging stagger in the range of 10-20. It remains to be determined if APE LINEs use 4-way junction structure to drive second-strand DNA cleavage and synthesis. Bioinformatic analysis of 5' junctions of full length L1 and Alu elements is suggestive of template jumping to the upstream target sequence and that DNA repair process might be an alternative path to 5' junction formation for abortive insertion events [13,15,44–46]. Twin priming in L1 might be a related, albeit aberrant, phenomenon to second-strand synthesis [47]. An association between the cDNA and the upstream target DNA has been hypothesized for some R1 elements [35]. Ribosomal sequences are also important for element-RNA/target-DNA interactions during first strand synthesis for R1Bm as well as several other site-specific LINEs but do not appear to be as important for R2Bm [24,48,49]. A few LINEs have very larger staggers. The R9 Av element, an R2 clade member, produces a 126 bp stagger [50]. For large staggers, a D-loop opening allows for the template jump and formation of the 4-way junction.

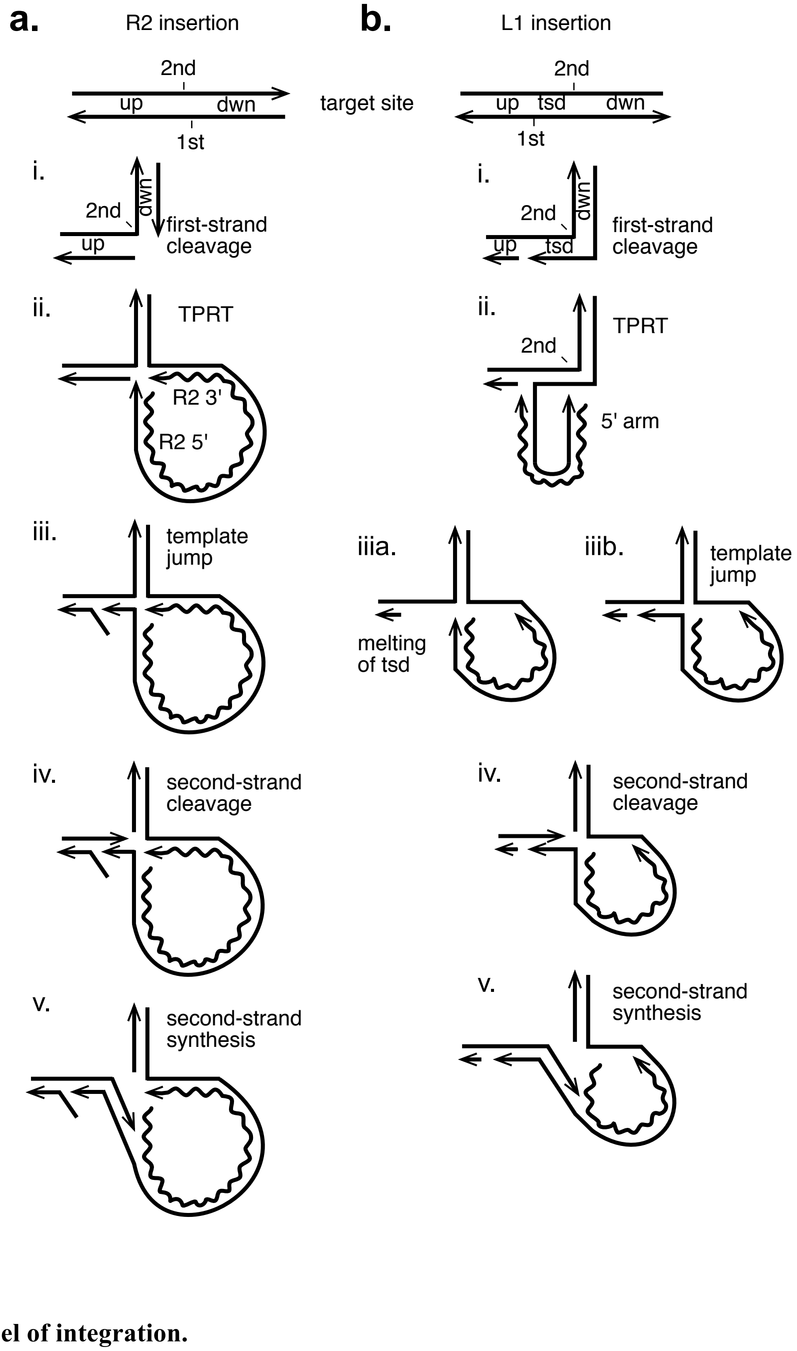


Figure 2- 8: A new model of integration.
(a) R2 integration. The R2 28S target site is diagramed with the positions of the first and second-strand cleavages that

will lead to insertion of a R2 new element. The initial steps of the integration reaction (i, ii) are as in Figure 1 except that the target site is bent 90° near the second strand insertion site for diagrammatic purposes. Step iii depicts a template jump/recombination event near the second-strand cleavage site that generates the 4-way junction. Step iv depicts second-strand cleavage. Finally, step v depicts second-strand DNA synthesis. Abbreviations: up (target sequences upstream of the insertion site), dwn (target sequences downstream of the insertion site).
(b) L1 integration. A target site is diagramed with the first and second-strand cleavages staggered such that a target site duplication (tsd) would occur upon element insertion. The steps are as in R2 except that the template jump displaces/melts the tsd region of the target to generate the 4-way junction.

2.5 Materials and Methods

2.5.1 Protein purification

R2Bm protein expression and purification were carried out as previously published [9]. 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 [52].

2.5.2 Nucleic acid preparation

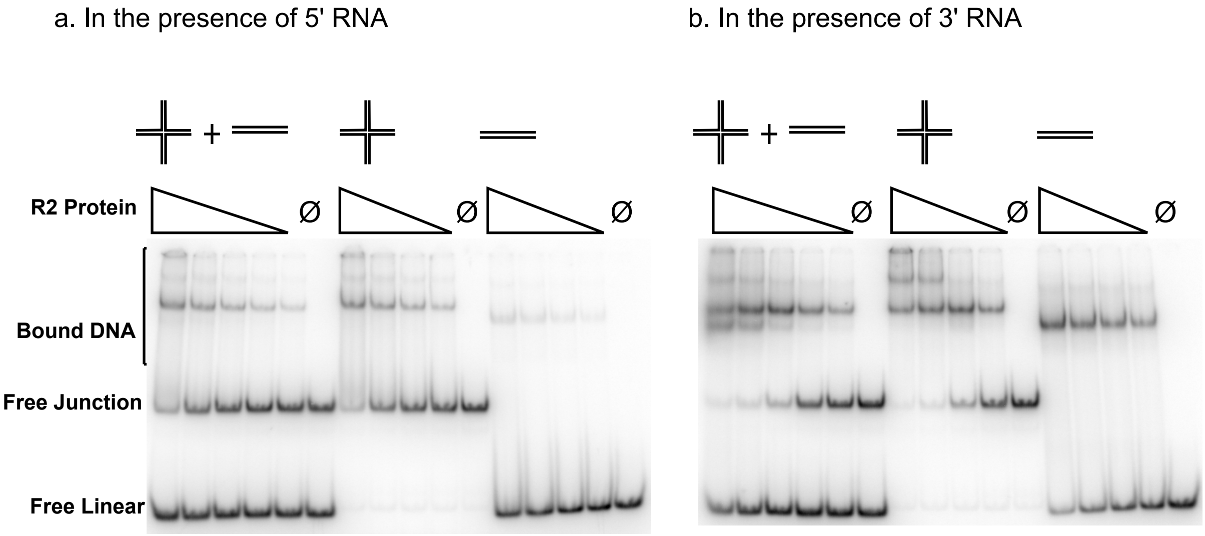
Oligos containing 28S R2 target DNA, non-target (non-specific) DNA, and R2 sequences were ordered from Sigma-Aldrich. The upstream (28Su) and downstream (28Sd) target DNA designations are relative to the R2 insertion dyad within the 28S rRNA gene. The oligo sequences are reported in Supplemental Table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

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. [51]. 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 95º C, followed by 10 minutes at 65º C, 10 minutes at 37º C and finally 10 minutes at room temperature. One of component oligos had been 5' 32P 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).

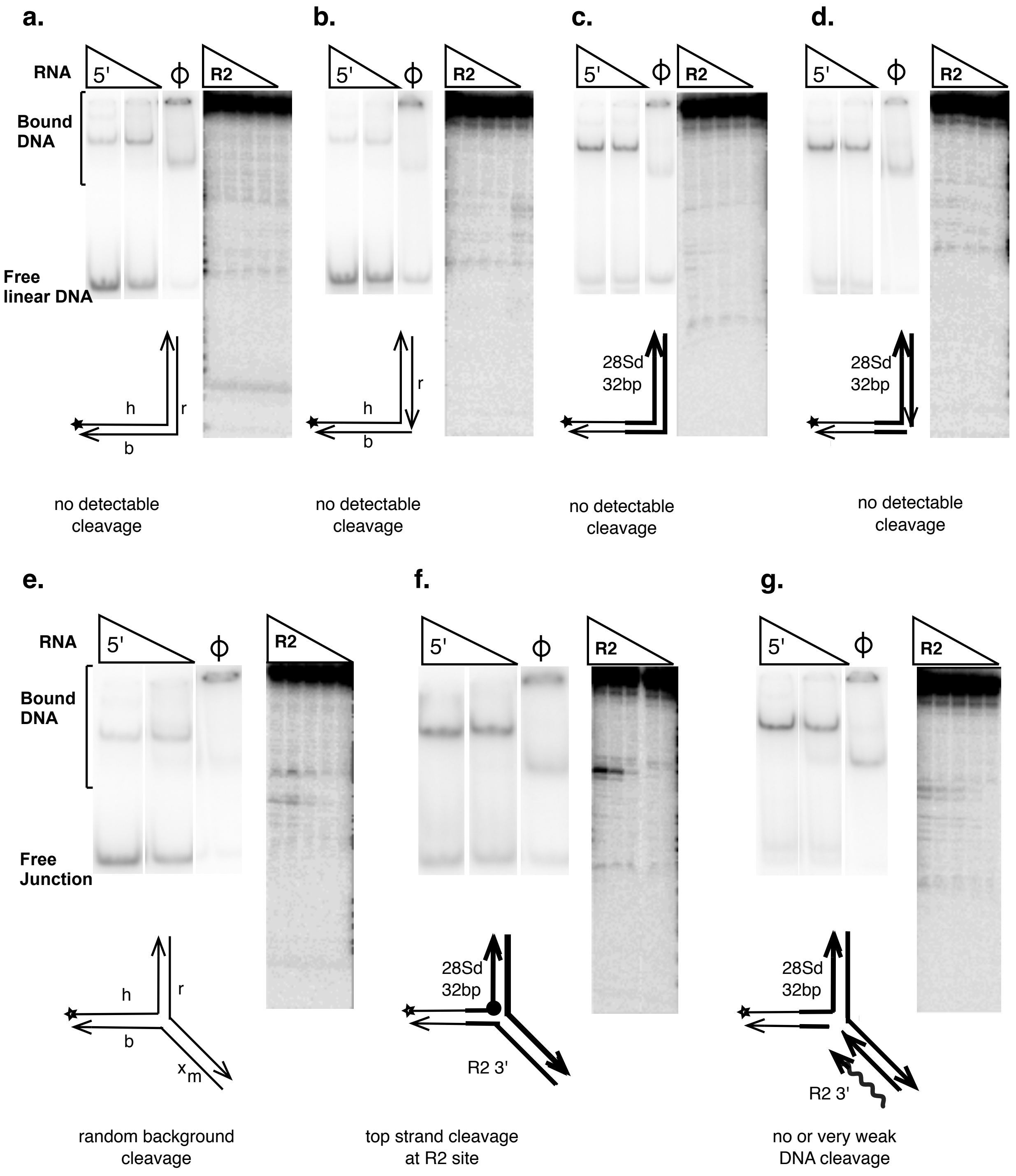
2.5.3 R2Bm reactions and analysis

R2 protein and target DNA binding and cleavage reactions were performed largely as previously reported [9]. 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 [52].

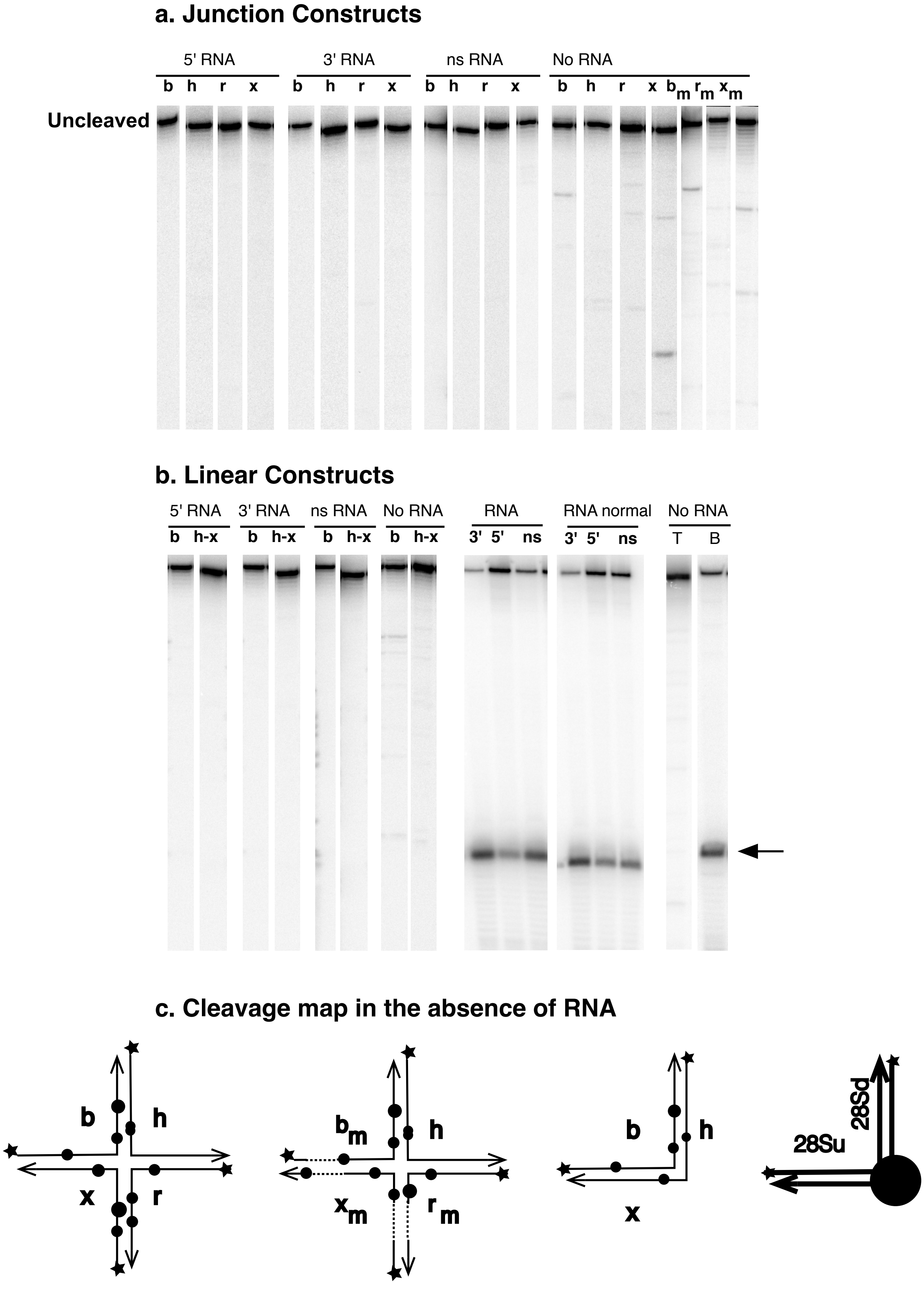
2.6 Supplemental file



Supplemental Figure 2-1: 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.



Supplemental Figure 2-2: Effect of 5' PBM RNA on binding to linear and 3-way junctions (EMSA gels). DNA Cleavage in the absence of RNA on linear and 3-way junctions (Denaturing gels). Panels a through g shows either linear or 3-way junction DNA constructs. Non-specific DNA is a straight thin line. R2 DNA and 28S rDNA are thick lines. R2 RNA strand is a squiggly line. Cleavages are shown as a black circle.



a. Junction Constructs

b. Linear Constructs

c. Cleavage map in the absence of RNA

Supplemental Figure 2-3: Non-specific four-way junction exhibiting structure independent cleavages. a) Lanes show each of the labeled strand of non-specific four-way junction in the presence of R2 protein and excess RNA in a denaturing condition. Uncleaved strand is the slowest running band and anything below that is considered as cleaved. new junctions were built with local sequence modification being made on b, r and x strands and they are represented with a suffix m. b) Lanes show each of the labeled strand of non-specific linear DNA, sharing the sequence with that of the non-specific junction and R2 target DNA in the presence of R2 protein and excess RNA in a denaturing condition. Arrow represents cleavage at R2 insertion site. c) Junction and linear DNA constructs are represented in lines. Thin lines are non-specific sequences and thick lines are R2-specific sequences. Cleavages are placed as solid black circle in the corresponding locations and circle diameter roughly correlates to cleavage intensity. Broken lines represent local sequence changes being made in that region. (See table x for sequence information). Solid black star represents 5' p32 labeling and arrows indicate 5' - 3' directionality. Cleavages are preserved whether they are in a four-way junction (a) or in a duplex form (b), but changes when the local sequences are altered.

Supplemental Table 1. Table presenting the DNA and RNA oligonucleotides used to build the linear and junction DNAs. 'Comp' stands for complementary strand.

Oligo Name	Sequence
b-strand	CCTCGAGGGATCCGTCCTAGCAAGCCGCTGCTACCGGAAGC TTCTGGACC
h-strand	GGTCCAGAAGCTTCCGGTAGCAGCGAGAGCGGTGGTTGAAT TCCTCGACG
r-b strand	CGTCGAGGAATTCAACCACCGCTCTCGCTGCTACCGGAAGCT TCTGGACC
Pre-cleaved r-b	1) CGCTGCTACCGGAAGCTTCTGGACC 2) CGTCGAGGAATTCAACCACCGCTCT
r-strand	CGTCGAGGAATTCAACCACCGCTCTTCTCAACTGCAGTCTAG ACTCGAGC
x-strand	GCTCGAGTCTAGACTGCAGTTGAGAGCTTGCTAGGACGGAT CCCTCGAGG
h-x strand	GGTCCAGAAGCTTCCGGTAGCAGCGGCTTGCTAGGACGGAT CCCTCGAGG
b_m-strand	CCTGCAGTGATCCGTCCTAGCAAGCCGCTGCTACCGGAAGCT TCTGGACC
r_m-strand	CGTCGAGGAATTCAACCACCGCTCTTCTCACCGATAAGTACG ACTCGAGC
x_m-strand	GCTCGAGTCGTACTTATCGGTGAGAGCTTGCTAGGACGGATC ACTGCAGG
Ns/28Sd 25 bp	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTC ATCTAATT
Comp ns/28Sd 25 bp	AATTAGATGACGAGGCATTTGGCTACCTTAAGCTACCGGAA GCTTCTGGA
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	TCCAGAAGCTTCCGGTAGCTTAAGGTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTA
Comp 28Sd 47 bp/comp R2 3’ RNA	TAATCCATTCATGCGCGTCACTAATTAGATGACGAGGCATTTGGCTATTTTCATCGCCGGATCATCATGCCA
28Su 73 bp/ns	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGGGTCCAGAAGCTTCCGGTAGCAGCGAGAGCGG
Comp ns/ comp R2 3’ RNA	CCGCTCTCGCTGCTACCGGAAGCTTCTGGACCCTATTTTCATCGCCGGATCATCATGCCA
Comp R2 5’ RNA/ Comp 28Su 73 bp	AAATTAAAATTATGCGTATCGCCCCCCTTAAGAGAGTCATAGTTACTCCCGCCGTTTACCCGCGCTTGCTTGAATTTCTTCACGTTGACATTCAGAGC
28Su 73bp/28Sd 47bp	GCTCTGAATGTCAACGTGAAGAAATTCAAGCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAGGTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTA

2.7 References

1. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910 (2002).
3. Moran, J. V. & Gilbert, N. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 836-869 (ASM Press, Washington, DC, 2002).
4. Christensen, S. M. & Eickbush, T. H. 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).
5. Kulpa, D. A. & Moran, J. V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**, 655-660 (2006).
6. Dewannieux, M. & Heidmann, T. LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. *Cytogenet Genome Res* **110**, 35-48 (2005).
7. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A 3’ Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell* **60**, 728-741 (2015).
8. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).

9. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287 (2016).
10. Martin, S. L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. *RNA Biol* **7**, 67-72 (2010).
11. Martin, S. L. The ORF1 Protein Encoded by LINE-1: Structure and Function During L1 Retrotransposition. *J Biomed Biotechnol* **2006**, 45621 (2006).
12. Matsumoto, T., Hamada, M., Osanai, M. & Fujiwara, H. Essential domains for ribonucleoprotein complex formation required for retrotransposition of telomere-specific non-long terminal repeat retrotransposon SART1. *Mol Cell Biol* **26**, 5168-5179 (2006).
13. Zingler, N. et al. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15**, 780-789 (2005).
14. Kurzynska-Kokorniak, A., Jamburuthugoda, V. K., Bibillo, A. & Eickbush, T. H. DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *J Mol Biol* **374**, 322-333 (2007).
15. Ichiyanagi, K., Nakajima, R., Kajikawa, M. & Okada, N. Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. *Genome Res* **17**, 33-41 (2007).
16. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* **357**, 1383-1393 (2006).
17. Suzuki, J. et al. Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. *PLoS Genet* **5**, e1000461 (2009).
18. Christensen, S. M., Ye, J. & Eickbush, T. H. 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).
19. Eickbush, T. H. & Malik, H. S. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 1111-1146 (ASM Press, Washington, DC, 2002).
20. Yang, J., Malik, H. S. & Eickbush, T. H. 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 (1999).
21. Feng, Q., Moran, J. V., Kazazian, H. H. J. & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916 (1996).
22. Weichenrieder, O., Repanas, K. & Perrakis, A. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986 (2004).
23. Han, J. S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA* **1**, 15 (2010).
24. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3**, MDNA3-0001 (2015).
25. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).
26. Gladyshev, E. A. & Arkhipova, I. R. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* **448**, 145-150 (2009).
27. Christensen, S. & Eickbush, T. H. Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* **336**, 1035-1045 (2004).
28. Zingler, N., Weichenrieder, O. & Schumann, G. G. APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res* **110**, 250-268 (2005).
29. Christensen, S., Pont-Kingdon, G. & Carroll, D. Comparative studies of the endonucleases from two related Xenopus laevis retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. *Genetica* **110**, 245-256 (2001).
30. Ostertag, E. M. & Kazazian, H. H. J. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35**, 501-538 (2001).
31. Feng, Q., Schumann, G. & Boeke, J. D. Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proc Natl Acad Sci U S A* **95**, 2083-2088 (1998).
32. Maita, N., Aoyagi, H., Osanai, M., Shirakawa, M. & Fujiwara, H. Characterization of the sequence specificity

of the R1Bm endonuclease domain by structural and biochemical studies. *Nucleic Acids Res* **35**, 3918-3927 (2007).

33. Eickbush, D. G., Burke, W. D. & Eickbush, T. H. Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441 (2013).
34. Kajikawa, M., Yamaguchi, K. & Okada, N. A new mechanism to ensure integration during LINE retrotransposition: A suggestion from analyses of the 5' extra nucleotides. *Gene* **505**, 345-351 (2012).
35. Stage, D. E. & Eickbush, T. H. Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. *Genome Biol* **10**, R49 (2009).
36. Fujimoto, H. et al. Integration of the 5' end of the retrotransposon, R2Bm, can be complemented by homologous recombination. *Nucleic Acids Res* **32**, 1555-1565 (2004).
37. Eickbush, D. G., Luan, D. D. & Eickbush, T. H. Integration of Bombyx mori R2 sequences into the 28S ribosomal RNA genes of Drosophila melanogaster. *Mol Cell Biol* **20**, 213-223 (2000).
38. Bibillo, A. & Eickbush, T. H. End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* **279**, 14945-14953 (2004).
39. Bibillo, A. & Eickbush, T. H. 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).
40. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res* **42**, 8405-8415 (2014).
41. Wyatt, H. D. & West, S. C. Holliday junction resolvases. *Cold Spring Harb Perspect Biol* **6**, a023192 (2014).
42. Eickbush, D. G. & Eickbush, T. H. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol* (2010).
43. Mukha, D. V., Pasyukova, E. G., Kapelinskaya, T. V. & Kagramanova, A. S. Endonuclease domain of the Drosophila melanogaster R2 non-LTR retrotransposon and related retroelements: a new model for transposition. *Front Genet* **4**, 63 (2013).
44. Gasior, S. L., Roy-Engel, A. M. & Deininger, P. L. ERCC1/XPF limits L1 retrotransposition. *DNA Repair (Amst)* **7**, 983-989 (2008).
45. Coufal, N. G. et al. Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc Natl Acad Sci U S A* **108**, 20382-20387 (2011).
46. Richardson, S. R. et al. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061 (2015).
47. Ostertag, E. M. & Kazazian, H. H. J. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**, 2059-2065 (2001).
48. Anzai, T., Osanai, M., Hamada, M. & Fujiwara, H. Functional roles of 3'-terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. *Nucleic Acids Res* **33**, 1993-2002 (2005).
49. Luan, D. D. & Eickbush, T. H. 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 (1996).
50. Arkhipova, I. R. et al. Genomic impact of eukaryotic transposable elements. *Mob DNA* **3**, 19 (2012).
51. Middleton, C. L., Parker, J. L., Richard, D. J., White, M. F. & Bond, C. S. Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res* **32**, 5442-5451 (2004).
52. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).

Chapter 3

The Linker Region of LINEs Binds a Key Integration Intermediate and Modulates DNA Cleavage and Polymerization throughout Integration [1]

Monika Pradhan and Shawn M. Christensen

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

3.1 Abstract

Non-LTR retrotransposons, or Long Interspersed Elements (LINEs), replicate by Target Primed Reverse Transcription (TPRT). Insertion involves two half reactions. Each half reaction involves DNA cleavage followed by DNA synthesis. The linker region, located just beyond the reverse transcriptase in the LINE open reading frame, contains a conserved set of predicted helices followed by a gag-like zinc knuckle. Point mutations in the helices reduce binding to a Holiday junction-like integration intermediate known to be the entry point into the second half of the integration reaction. These mutations also severely impair the DNA endonuclease and reverse transcriptase activities of the integration reaction during both half reactions. Mutations in the gag-like zinc knuckle also impair DNA cleavage and DNA synthesis in some instances. The linker appears to function as a branched DNA binding platform and as a protein-DNA conformational switch area used to properly position nucleic acid substrates into the active sites of the reverse transcriptase and the DNA endonuclease.

3.2 Introduction

Long Interspersed Elements (LINEs), also known as non-long terminal repeat (non-LTR) retrotransposons, are an ancient and abundant group of transposable elements (TEs) that are found across the eukaryotic tree of life [1–3]. LINEs integrate into new sites by a process called Target Primed Reverse Transcription (TPRT). The element encoded DNA endonuclease creates a nick in the host chromatin to expose a free 3'-OH group. The 3'-OH group is used by the element encoded reverse transcriptase to prime reverse transcription of the element RNA at the site of insertion. LINEs encode an invariant gag-like zinc-knuckle cysteine/histidine rich motif ($CX_{2\text{-}3}CX_{7\text{-}8}HX_{4}C$) downstream of the reverse transcriptase [4,5]. The spacing of the cysteines and histidine in the knuckle is unique to the knuckle found in LINEs. Immediately upstream of the zinc knuckle is a set of predicted helices [6].

The R2 LINE from *Bombyx mori* (R2Bm) is a site specific LINE that has served as a model system in which to dissect the integration reaction of LINEs at the biochemical level as the protein can be purified in active form and used in *in vitro* assays [4,7,8]. The R2 ORF encodes a multifunctional protein with N-terminal zinc-finger(s) (ZF) and myb domains that are involved in DNA binding; an RNA binding (RB) domain; a central reverse transcriptase (RT); a linker region containing several conserved predicted helices (HINALP motif), and a gag-like zinc knuckle (CCHC motif), and a PD-(D/E)XK type II restriction-like endonuclease (RLE) domain (Figure 3-1A) [4,6,9–13] The R2 RNA sequence corresponding to the 5' and 3' untranslated region (UTR) folds into distinct structures that are known to bind R2 protein, and hence are termed as 5' PBM and 3' PBM, respectively (Figure 3-1A) [14–16]. Binding to the 5' PBM and 3' PBM RNAs control protein conformation and role in the integration reaction (Figure 3-1B) [13]. Selective addition of the RNA, DNA, and protein components allow for distinct stages of the integration reaction to be assayed.

R2 protein bound to 3' PBM adopts a conformation that allows the protein to bind the upstream 28S DNA sequences (28Su) relative to the insertion site. The domain(s) of the R2 protein that contacts the 28Su to form upstream protein subunit remain largely unidentified [17–19] R2 protein bound to the 5' PBM adopts a conformation that allows the protein to bind the downstream 28S DNA sequences (28Sd). The ZF and Myb motifs of R2 protein include major residues that are known to interact with the 28Sd forming downstream protein subunit [11]. The upstream and downstream protein subunits catalyze the integration of R2 elements in two half reactions each consisting of DNA cleavage followed by DNA synthesis [13]. The five steps of integration are: (1) The endonuclease from upstream subunit nicks the target DNA exposing a 3'-OH at the insertion site; (2) The exposed 3'-OH is used as a primer by the upstream subunit's reverse transcriptase for TPRT; (3) A template jump or recombination event occurs where the cDNA from the 5' end of the reverse transcribed becomes associated with the upstream target DNA sequences to form a four-way junction; (4) The downstream subunit cleaves the four-way DNA junction; (5) the 3'-OH generated by the cleavage event is used as the primer for second strand DNA synthesis of the element.

The role of the linker region, located after the RT in all LINEs, remains unknown [6]. Point mutations were introduced into the linker's gag-like zinc knuckle and presumptive α-finger (Figure 3-1C). The spacing of the CCHC motif is unique to LINEs [20,21]. In a previous *in vivo* study using APE bearing human LINE-1 elements, mutating the first two cysteines in the linker region's CCHC motif significantly reduced LINE-1 retrotransposition [22]. In another *in vivo* study with human LINE-1, reduced levels of RNP complex was observed when first two cysteines were mutated which indicated its possible role in nucleic acid binding [23]. When the zinc knuckle structure was altered by substituting first three cysteines into serine, no reduction in RNA binding activity was reported for human LINE-1 elements *in vitro* [24]. However, in the same study, sequences C-terminal to the RT was found to be involved in RNA binding. Mutating residues upstream of the presumptive α-finger in LINE-1 elements reduced retrotransposition activity *in vivo* [22]. The helices upstream of the zinc knuckle, along with the zinc knuckle itself, reportedly align with the α-finger and the non-zinc knuckle of the eukaryotic splicing factor, Prp8 [6,25,26]. In this study, a series of double mutations are generated throughout the presumptive α-finger and zinc knuckle of R2Bm and tested *in vitro* for loss of function under conditions that test for DNA binding, first-strand DNA cleavage, first-strand DNA synthesis, second-strand DNA cleavage, and second-strand DNA synthesis.

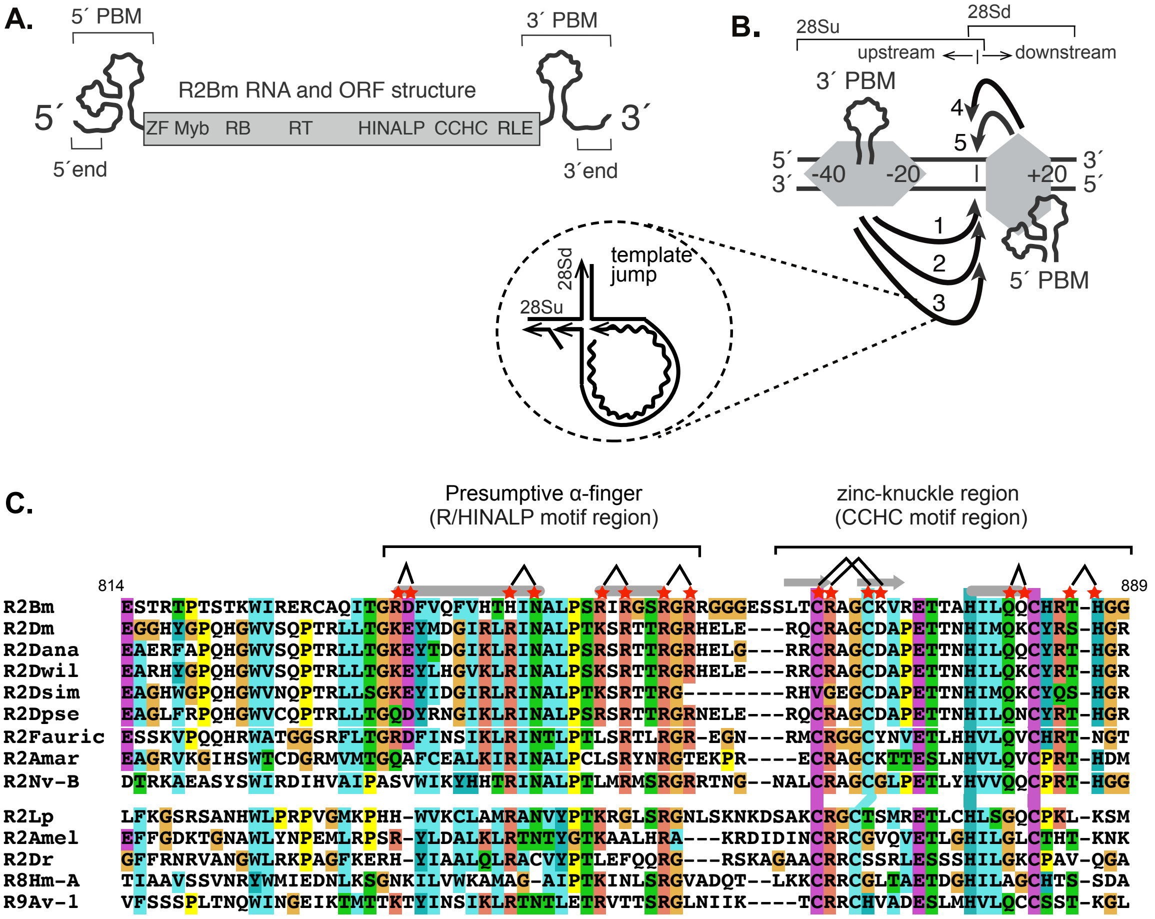

A.
5´ PBM
3´ PBM
5´
R2Bm RNA and ORF structure
ZF Myb RB RT HINALP CCHC RLE
3´
5´end
3´end
B.
28Su
28Sd
upstream
downstream
3´ PBM
4
5
5´
3´
-40
-20
+20
3´
5´
1
2
3
5´ PBM
28Sd
28Su
template jump
C.
Presumptive α-finger
(R/HINALP motif region)
zinc-knuckle region
(CCHC motif region)
814
889
R2Bm ESTRTPTSTKWIRERCAQITGRDFVQFVHTHINALPSRIRGSRGRRGGGESSLTCRAGCKVRETTAHILQQCHRT-HGG
R2Dm EGGHYGPQHGWVSQPTRLLTGKEYMDGIRLRINALPTKSRTTRGRHELE---RQCRAGCDAPETTNHIMQKCYRS-HGR
R2Dana EAERFAPQHGWVSQPTRLLTGKEYTDGIKLRINALPSRSRTTRGRHELG---RRCRAGCDAPETTNHILQQCYRT-HGR
R2Dwil EARHYGPQHGWVSQPTRLLTGKEYLHGVKLRINALPSKSRTTRGRHELE---RRCRAGCDAPETTNHILQKCYRT-HGR
R2Dsim EAGHWGPQHGWVNQPTRLLSGKEYIDGIRLRINALPTKSRTTRG---------RHVGEGCDAPETTNHIMQKCYQS-HGR
R2Dpse EAGLFRPQHGWVCQPTRLLTGQDYRNGIKLRINALPSRSRTTRGRNELE---RQCRAGCDAPETTNHILQNCYRT-HGR
R2Fauric ESSKVPQQHRWATGGSRFLTGRDFINSIKLRINTLPTLSRTLRGR-EGN---RMCRGGCYNVETLHHVLQVCHRT-NGT
R2Amar EAGRVKGIHSWTCDGRMVMTGQAFCEALKIRINALPCLSRYNRGTEKPR----ECRAGCKTTESLNHVLQVCPRT-HDM
R2Nv-B DTRKAEASYSWIRDIHVAIPASVWIKYHHTRINALPTLMRMSRGRRTNG--NALCRAGCGLPETLYHVVQQCPRT-HGG
R2Lp LFKGSRSANHWLPRPVGMKPHH-WVKCLAMRANVYPTKRGLSRGNLSKNKDSAKCRGCTSMRETLCHLSGQCPKL-KSM
R2Amel EFFGDKTGNAWLYNPEMLRPSR-YLDALKLRTNTYGTKAALHRA---KRDIDINCRRCGVQVETLGHILGLCTHT-KNK
R2Dr GFFRNRVANGWLRKPAGFKERH-YIAALQLRACVYPTLEFQQRG---RSKAGAACRRCSSRLESSSHILGKCPAV-QGA
R8Hm-A TIAAVSSVNRYWMIEDNLKSGNKILVWKAMAG-AIPTKINLSRGVADQT--LKKCRRCGLTAETDGHILAGCHTS-SDA
R9Av-1 VFSSSPLTNQWINGEIKTMTTKTYINSIKLRTNTLETRVTTSRGLNIIK----TCRRCHVADESLMHVLQCCSST-KGL

Figure 3-1: R2Bm structure, integration model, and multiple sequence alignment. A. R2Bm RNA structure. Box in the middle represents open reading frame (ORF) with conserved motifs. Abbreviations: zinc finger (ZF), Myb (Myb), RNA binding (RB), reverse transcriptase domain (RT), a conserved predicted α-helices with HINALP residues (HINALP), a gag-like zinc knuckle with Cysteine/Histidine rich motif (CCHC) and a PD-(D/E)XK type II restriction-like endonuclease (RLE). R2Bm RNA segments corresponding to the 5' and 3' untranslated regions known to adopt distinct structures and bind the R2Bm protein are labeled within the brackets as 5' and 3' protein binding motifs (PBMs) respectively. 5' PBM and 3' PBM are used to generate data in the paper. Sequences of R2Bm RNA at the 5' and 3' end are used to construct four-way junction target DNA, and are marked within the brackets as 3' end and 5' end. B. R2 target site, 28S rDNA, and insertion model. R2 protein associated with the 3' PBM RNA binds 20 to 40 bases upstream (28Su) of the insertion site (vertical line) and protein associated with the 5' PBM RNA binds to 20 bases downstream of the insertion site [11,27]. Insertion occurs in five steps: (1) First strand cleavage by upstream protein

subunit endonuclease. (2) First strand synthesis (TPRT) by the upstream protein subunit reverse transcriptase. (3) Template jump/ recombination to upstream target DNA (28Su) resulting in a four-way junction branched structure (zoomed in diagram). (4) Second strand cleavage by endonuclease of the downstream protein subunit. (5) Second strand synthesis by reverse transcriptase of downstream protein subunit. C. Multiple sequence and secondary structure alignment of the linker region of RLE LINEs. Red star represents the residues that were mutated and half triangle represents double point mutants generated in the presumptive α-finger and the zinc knuckle regions. Double point mutants generated for this study were: GR/AD/A, H/AIN/AALP, SR/AIR/A, SR/AGR/A, C/SC/SHC, CR/AAGCK/A, HILQ/AQ/A and RT/AH/A. The first four mutants are in the presumptive α-finger region and the last four mutants are in the zinc knuckle region as indicated by the brackets on the top. Secondary structures are predicted by Ali2D and grey bars represent α-helices and arrow represents β-strands. Abbreviations: R2Bm = Bombyx mori, R2Dm = Drosophila melanogaster, R2Dana = Drosophila ananassae, R2Dwil = Drosophila willistoni, R2Dsim = Drosophila simulans, R2Dpse = Drosophila pseudoobscura, R2Fauric = Forficula auricularia, R2Amar = Anurida maritima, R2Nv-B = Nasonia vitripennis, R2Lp = Limulus polyphemus, R2Amel = Apis mellifera, R2Dr = Danio rerio, R8Hm-A = Hydra magnipapillata, R9Av-1 = Adineta vaga.

3.3 Results

3.3.1 Double point mutants were generated in the presumptive α-finger and zinc knuckle regions of the linker

To investigate the role of the linker region's presumptive α-finger (HINALP motif region), and zinc knuckle (CCHC motif region), a number of double point mutants were generated (Figure 3-1C). The mutations in the presumptive α-finger region included GR/AD/A, VH/ATH/A, H/AIN/ALP, SR/AIR/A and SR/AGR/A. The H/AIN/AALP and SR/AIR/A mutations resulted in a reduction of soluble protein being recovered compared to wild type (WT) protein. The VH/ATH/A mutation did not produce soluble protein and was dropped from the study. The mutations in the zinc knuckle region were C/SC/CHC, CR/AAGCK/A, E/AT/AT, HILQ/AQ/A and RT/AH/A (Figure 3-1C). The C/SC/CHC mutation resulted in greatly reduced soluble protein being recovered compared to wild type (WT) protein. The E/AT/AT mutation did not yield usable quantities of protein and was dropped from the study.

3.3.2 Mutations in the core residues of the HINALP and CCHC motifs affect target DNA binding and leads to loss of DNA cleavage specificity

There were four double point mutants created in the HINALP region and four in the zinc knuckle region. The H/AIN/AALP and the C/SC/SHC mutants appear to have nearly identical phenotypes. Both sets of mutations severely impair DNA binding to the linear DNA as well as the ability to form the correct DNA-RNA-Protein complexes in

EMSA gels on linear DNA (Figure 3-2A). Only the well complex and a diffuse smear leading down from the well to the free DNA are observed (Figure 3-2A). This observation is true for both upstream binding conditions (i.e., presence of 3' PBM RNA) and downstream binding conditions (i.e., presence of 5' PBM RNA). The Cysteine and Histidine residues of the zinc knuckle motif are the presumptive zinc coordinating residues. The C/SC/SHC mutation may promote local misfolding of the linker. The H/AIN/AALP mutation may have also affected the folding of the linker.

In the presence of 3' PBM RNA, the H/AIN/AALP and C/SC/SHC mutants showed little to no first strand cleavage at the insertion site (Figure 3-2B). Second-strand DNA cleavage was similarly abolished in the presence of 5' PBM RNA. Instead of site specific DNA cleavage abundant promiscuous cleavages were observed at aberrant sites on both strands of the target DNA.

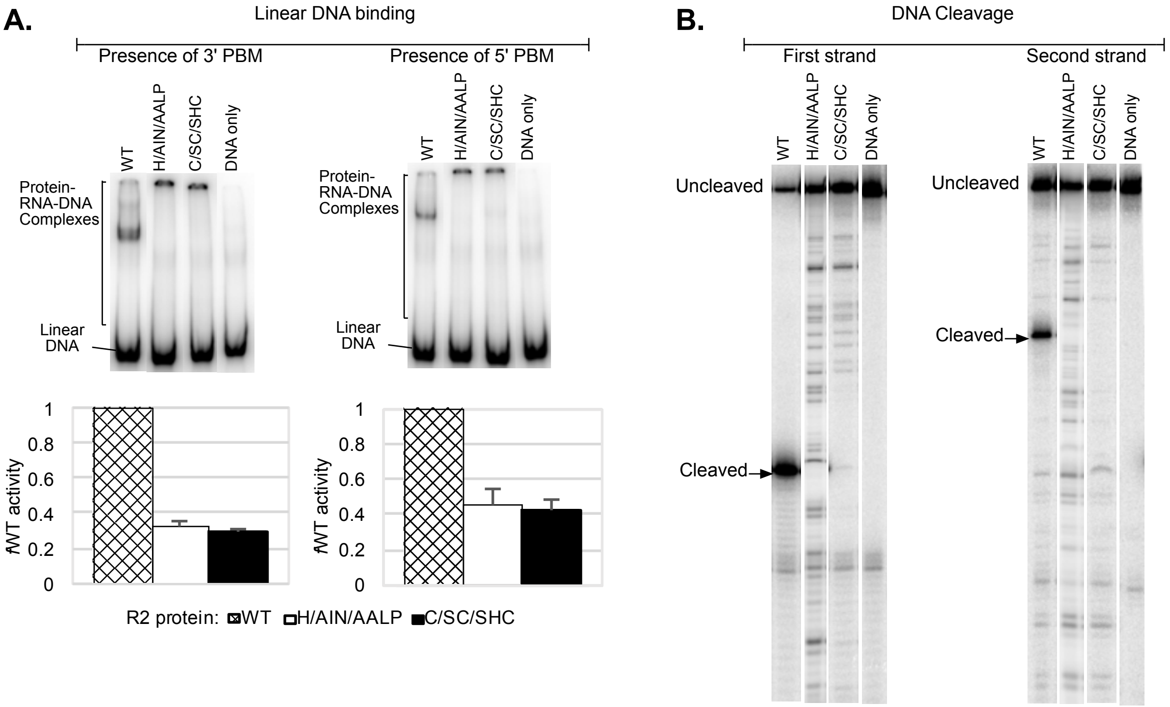


Figure 3-2: H/AIN/AALP and C/SC/SHC mutants affect linear DNA binding and cleavage activity. A. EMSA gels and bar graphs reporting mutant's ability to bind to target DNA in the presence of 3' and 5' PBM RNAs. Wild type (WT) protein activity is set to 1 and the mutant protein activity is then given as a fraction of WT activity (*f*WTactivity). **B.** Denaturing gels showing aberrant cleavages on the first and second cleavage sites and strands.

3.3.3 Mutations in the presumptive α-finger affect DNA binding, especially to a specific branched integration-intermediate analog

To better determine if the presumptive α-finger is involved in securing protein to upstream and/or downstream target DNA sequences, mutations surrounding the core HINALP motif were tested. The GR/AD/A, SR/AIR/A and SR/AGR/A mutants were tested for their ability to bind linear target in the presence of 3' PBM RNA and in the presence of 5' PBM RNA. Two positive controls were used, WT R2 protein and R2 protein with a catalytic residue of the RLE mutated to alanine (KPD/A) so as to knockout DNA cleavage but not DNA binding so that the α-finger mutations that either do or do not affect DNA cleavage (see the next section) are appropriately controlled for. The DNA binding ability of the mutant relative to the control R2 proteins were assayed using Electrophoretic Mobility Shift Assays (EMSAs) (Figure 3-3A). Duplicate lanes were loaded and duplicate binding reactions were run. Vector control extract and no protein lanes served as negative control lanes.

Upstream target DNA binding was moderately reduced (24%) by the GR/AD/A mutation and very mildly reduced (13%) by the SR/AIR/A mutation. But upstream target DNA binding activity was significantly increased by up to 32% by SR/AGR/A mutant (Figure 3-3A). Downstream target DNA binding activity for GR/AD/A and SR/AGR/A mutants was similar to WT activity, with only a mild decrease of ~13%. The SR/AIR/A mutation decreased binding in the range of 19-28%. All the three mutants did not seem to affect the migration pattern of protein-RNA-DNA complexes much if at all, although, more of the well complex formation was observed for SR/AIR/A mutant (Figure 3-3A). The ability of the mutants to bind to linear target DNA in the absence of RNA is presented in supplementary figure 1.

The ability of the mutants to bind a four-way junction integration intermediate was also tested. The four-way junction mimics the branched structure adopted by 28S rDNA after the template jump step, and contains 28Sd rDNA sequence (north arm), a non-specific sequence (west arm), a R2 5'-end RNA-DNA duplex (south arm), and a R2 3'-end RNA-DNA duplex (east arm) (Figure 3-3B) (Chapter 2). The four-way junction DNA was radiolabeled at the top strand of the 5' end of the west arm. The junction DNA was incubated with R2 protein in the absence of RNA and aliquots were run in EMSA gel (Figure 3-3B). After quantitation as described above, we could see that two mutants significantly reduced the ability of R2 protein to bind to the four-way junction, SR/AIR/A by 63% and SR/AGR/A by

48% while GR/AD/A mutant's binding activity was comparable to that of WT activity showing only a mild reduction of 12%.

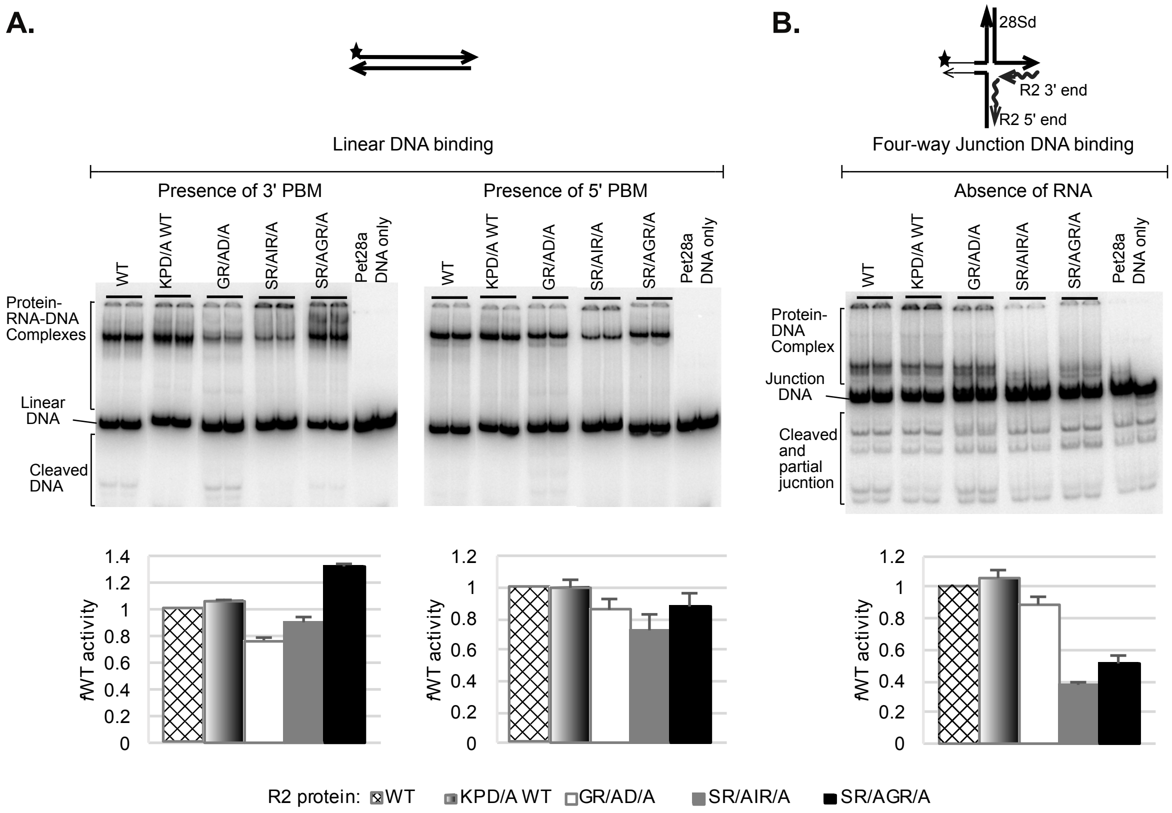


Figure 3-3: DNA binding by α-finger mutant proteins. A. EMSA gels and bar graphs reporting the relative ability of the mutants to bind to linear target DNA. WT and KPD/A WT served as positive controls while Pet28a and DNA only lanes served as negative controls. Standard deviation is presented on top of the bars. B. EMSA gels and bar graphs reporting binding to an analog of the branched insertion intermediate (Chapter 2). The black star in the substrate diagrams indicates the strand that was 5' end labeled.

3.3.4 Mutations in the presumptive α-finger reduce first-strand DNA cleavage

The ability of the GR/AD/A, SR/AIR/A and SR/AGR/A mutants to perform first-strand DNA cleavage was assayed. The R2 proteins were pre-bound to 3' PBM followed by incubation with target DNA. A protein titration series was used (seven 1:3 protein dilutions). An aliquot of each reaction was run on an EMSA gel (Figure 3-4A) and on a denaturing (8M urea) polyacrylamide gel (Figure 3-3B, full length denaturing gel is available in supplementary figure 3-2). The target DNA was ^{32}P labeled at the 5' end of the bottom strand (i.e., 28S antisense strand) so that the cleavage of this strand could be tracked in the denaturing gel.

At higher protein concentration lanes (first two) in EMSA gel (Figure 3-4A), Protein-DNA complexes corresponding to the one seen in the absence of RNA were observed for WT, GR/AD/A and SR/AGR/A mutants as the RNA concentration had been held constant and as protein neared parity with the RNA concentration, DNA-complexes appeared along with protein-RNA-DNA complexes before everything becomes stuck in the wells. The mutations did not appear to greatly affect the migration pattern of protein-RNA-DNA complexes as compared to WT. The cleavage activity of each of the mutant is reported as a scatter plot of the fraction of cleaved DNA (*f*cleaved), calculated from the urea denaturing gels, as a function of the fraction of bound (*f*bound) DNA, calculated from the EMSA gels. GR/AD/A mutant did not affect the first strand cleavage activity of R2 protein, however, the SR/AIR/A and SR/AGR/A mutants significantly reduced the ability of the bound protein to undergo first strand DNA cleavage (Figure 3-4B and 4C). No cleavages beyond the R2 cleavage site were observed for either WT or mutants (Supplementary figure 2).

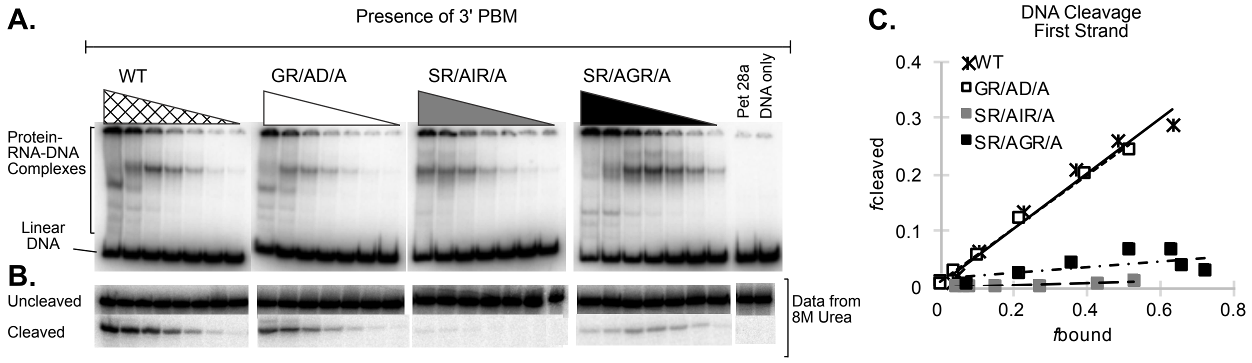


Figure 3-4: First strand DNA cleavage activity by α-finger mutant proteins. **A.** EMSA gel of protein bound to linear target DNA. The triangles above the EMSAs represent a protein titration series. **B.** Denaturing gel of the same reactions in A. showing the signals for uncleaved and cleaved (first strand) product of target DNA. Fraction of target DNA that undergoes first strand cleavage (*f*cleaved) is quantitated from denaturing gel. **C.** Scatter plot of fraction of cleaved target DNA (*f*cleaved) plotted as a function of fraction of target DNA bound by protein (*f*bound) at each protein concentrations. Data points for WT, GR/AD/A, SR/AIR/A and SR/AGR/A are represented by asterisk, white box, grey box, and black box respectively.

3.3.5 Mutations in the presumptive α-finger reduce first strand cDNA synthesis

To investigate if HINALP region affects TPRT (first-strand DNA synthesis), pre-cleaved target DNA with nick at the insertion site on first/bottom strand was incubated with R2 protein in the presence of 3’ PBM RNA and dNTPs (Figure 3-5A). The target DNA was radiolabeled at the 5’ end of the bottom strand to track the formation of the TPRT product. Aliquots of reactions across a protein titration series were assayed on EMSA and denaturing polyacrylamide gels. A graph of the fraction of target DNA that underwent TPRT (*f*synthesis) as a function of fraction of target DNA bound by R2 protein (*f*bound) is reported in Figure 3-5B. The gels are presented in supplemental figure 3A and 3B. GR/AD/A and SR/AIR/A mutants completely abolished the TPRT activity while SR/AGR/A mutant reduced first strand synthesis activity by approximately 50% (Figure 3-5B).

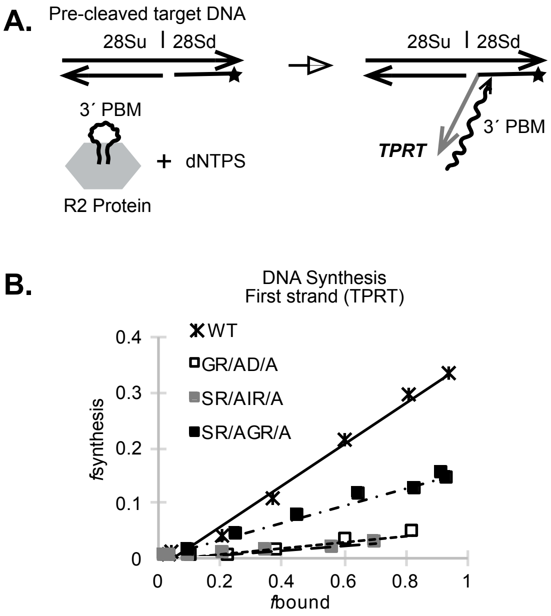


Figure 3-5: First strand synthesis (TPRT) by α-finger mutant proteins. A. Experimental setup for first strand synthesis assay in which pre-cleaved target DNA was incubated with R2 protein in the presence of 3’ PBM RNA and dNTPs. **B.** Scatter plot showing fraction of the DNA that underwent synthesis (*f*synthesis) as a function of fraction of the DNA that was bound by R2 protein (*f*bound) across a protein titration series. The symbols and abbreviations are as in the previous figures.

3.3.6 Mutations in the presumptive α-finger affect second-strand DNA cleavage

In order to determine the role, if any, the GR/AD/A, SR/AIR/A and SR/AGR/A mutants have on second-strand cleavage, two different cleavage assays were undertaken: (1) on linear target DNA in the presence of 5' PBM RNA, and (2) cleavage on 4-way junction DNA in the absence of RNA. On linear DNA, R2 protein binds downstream of the insertion site in the presence of 5' PBM RNA but only cleaves once the RNA dissociates from the complex. The dissociation occurs as the RNA to protein ratio drops across the protein titration series (RNA is held constant) [16]**.** In EMSA gel, the migration pattern of protein-RNA- DNA complexes of mutants were similar to that of WT, however, a band corresponding to a second strand cleaved product located immediately below the major protein-RNA- DNA complex was absent for SR/AIR/A and SR/AGR/A mutants (Figure 3-6A). In denaturing gel, the signal for second strand cleaved product was not visible for SR/AIR/A and SR/AGR/A mutants (Figure 3-6B). Non-specific cleavages were not observed for any of the mutants (Supplementary figure 4). While GR/AD/A showed WT activity, SR/AIR/A and SR/AGR/A mutants knocked out the endonuclease activity of R2 protein to make second strand cleavage on linear target DNA (Figure 3-6C).

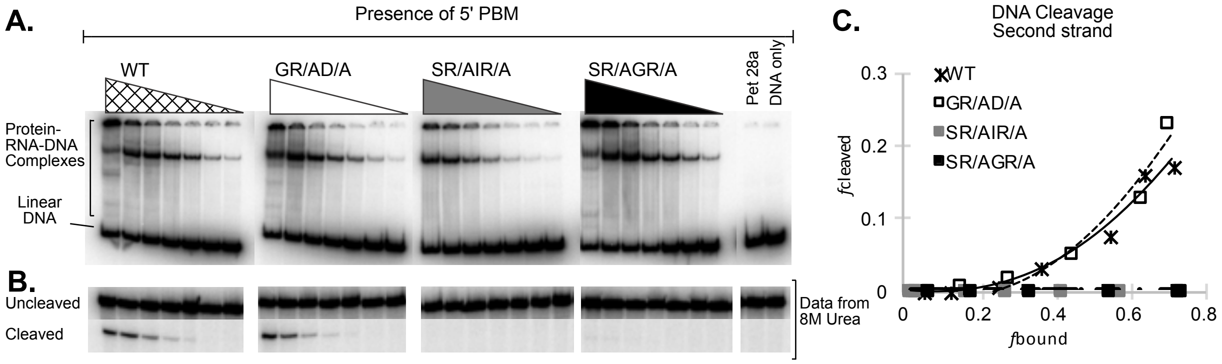


Figure 3-6: Second strand DNA cleavage by α-finger mutant proteins on linear target DNA. **A.** EMSA gel used to calculate the fraction of target DNA bound by R2 protein. **B.** Denaturing gel used to calculate the fraction of target DNA cleaved by the R2 protein. **C.** Scatter plot of second strand cleavage activity. Symbols and abbreviations are as in previous figures.

As noted above, second strand cleavage activity was also tested using a 4-way junction integration intermediate (Figure 3-7). Second strand DNA cleavage is believed to occur when the protein is in the "no RNA" bound state [16] and that the proper substrate for DNA cleavage is a 4-way junction intermediate formed by template jump (Christensen S.M., unpublished data). A diagram of the junction DNA used is shown in Figure 3-3B. The junction DNA was radiolabeled at the 5' end of the west arm to track cleavages on the top strand of the 28S DNA. The cleavage activity for mutants was tested against WT as indicated in the previous target DNA cleavage assays but in the absence of RNA. EMSA gel shows slightly smeary migration pattern of Protein-DNA complexes in the mutants (Figure 3-7A) as compared to that of WT. The signal for cleaved product of four-way junction DNA was almost invisible in denaturing gel (Figure 3-7B), and non-specific cleavages on junction DNA were not reported for any of the mutants (full length denaturing gel in supplementary figure 5). Endonuclease activity to cleave the second strand on a four-way junction DNA was completely knocked out by SR/AIR/A and SR/AGR/A mutants while GR/AD/A mutant showed WT cleavage activity as shown in the scatterplot (Figure 3-7C).

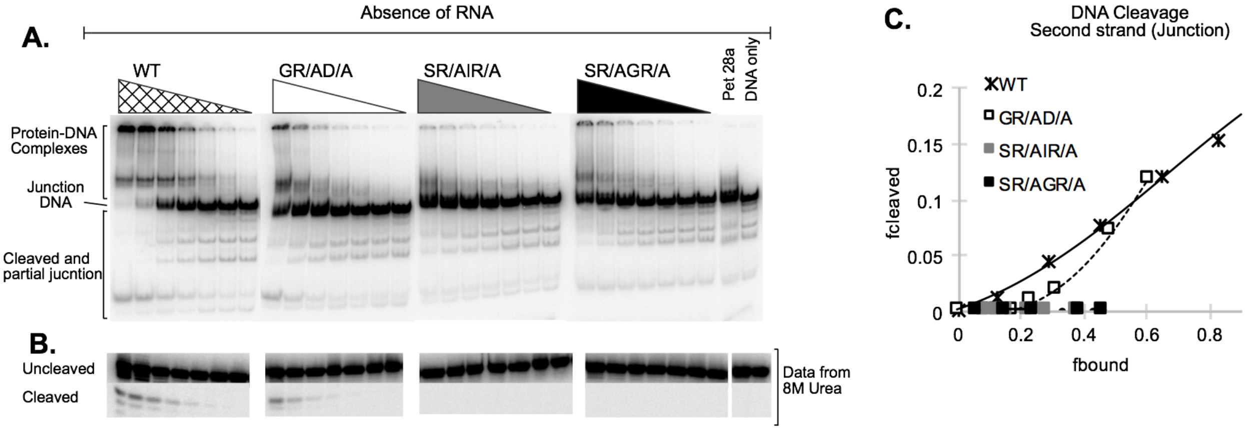


Figure 3-7: Second strand DNA cleavage activity by α-finger mutant proteins on four-way junction DNA. **A.** EMSA gel used to calculate the fraction of target DNA bound by the R2 protein. **B.** Denaturing gel used to calculate the fraction of target DNA cleaved by the R2 protein. **C.** Scatter plot of second strand cleavage activity. Symbols and abbreviations are as in previous figures.

3.3.7 Mutations in the presumptive α-finger affect second strand synthesis

In addition to testing second strand cleavage activity of HINALP mutants, the same mutants were subjected to experiments designed to test second stand DNA synthesis activity. As DNA cleavage is not very efficient, pre-cleaved DNA was used, and as the upstream and the downstream ends separate *in vitro* post DNA cleavage, the two ends were held together by a covalent linkage between the east and south arms (i.e., between R2 5' end sequence and

R2 3' end sequence) (see diagram in Figure 3-8A) (Chapter 2). This post second-strand cleavage analog was developed and reported in a previous study. The HINALP mutants were tested for second-strand DNA synthesis activity using this construct (Figure 3-8B). The 5' end of the west arm was radiolabeled to visualize the newly synthesized second-strand in denaturing gel (represented by black star in Figure 3-8A). The graph shown in Figure 3-8B was obtained from EMSA and denaturing gels (Supplementary figure 6A and 6B) as described previously for first strand synthesis assay. GR/AD/A mutant seems to act more like WT except that at the highest protein concentration, the amount of second strand synthesis goes down. SR/AIR/A mutant looks more like WT until about 40% of the target DNA is protein-bound but with increasing protein concentrations, the second strand synthesis decreases significantly. SR/AGR/A mutant drastically diminishes the ability of R2 protein to synthesize second strand as shown in the Figure 3-8B graph.

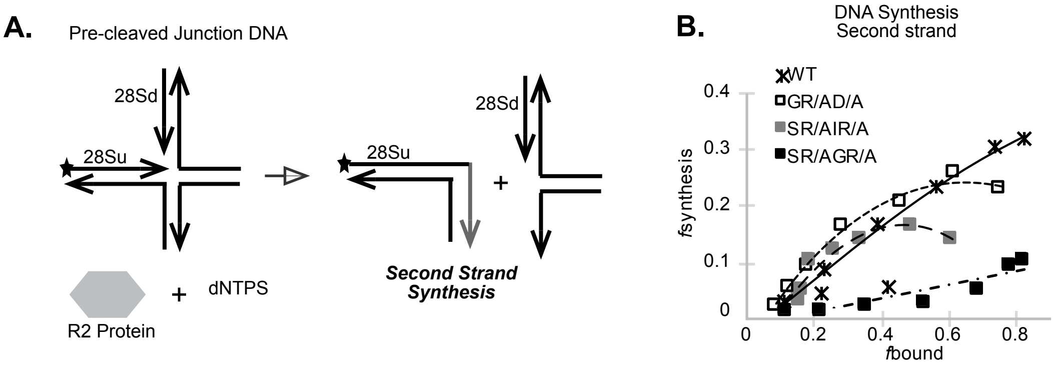


Figure 3-8: Second strand DNA synthesis activity by α-finger mutant proteins. **A.** Experimental setup for second strand synthesis assay in which pre-cleaved four-way junction DNA was incubated with R2 protein in the presence of dNTPs. **B.** Scatter plot of second strand synthesis activity. Symbols and abbreviations are as in previous figures.

3.3.8 Mutating residues in the zinc knuckle region affect target DNA cleavage and second stand synthesis

While C/SC/SHC mutant showed to affect target DNA binding and cleavage, the role of CCHC region was further investigated with the help of three additional double point mutants in this region: CR/AAGCK/A, HILQ/AQ/A, and RT/AH/A (Figure 3-1C). The mutants were assayed for DNA cleavage and new strand synthesis activities as described previously.

All the three mutants only slightly reduced the ability of the R2 protein to cleave the first strand at the insertion site (supplementary figure 7A), and they did not seem to have any effect on the first strand synthesis activity by TPRT (supplementary figure 7B). Although CR/AAGCK/A, HILQ/AQ/A and RT/AH/A mutants were nearly WT for first strand cleavage and synthesis, at least two of the mutants, HILQ/AQ/A and RT/AH/A significantly abolished second strand cleavage activity on a linear DNA (Figure 3-9A). In addition to the decrease in second strand cleavage activity at the insertion site, the endonuclease of RT/AH/A mutant was also found to be cleaving at a nearby site on top strand of linear target DNA as shown in Figure 3-9B. The second strand cleavage activity of the mutants was also tested using the four-way junction target DNA, however, all the three mutants showed WT activity (Supplementary figure 7C). Yet again, the endonuclease of RT/AH/A mutant showed an additional cleavage at a non-R2 specific site (Supplementary figure 7D).

Second strand synthesis assay with a pre-nicked four-way junction DNA, as shown in Figure 3-8A, was conducted for the three CCHC region mutants as described before for HINALP region mutants. The second strand synthesis product formation per bound unit of target DNA for CR/AAGCK/A looked very similar to that of WT, but for HILQ/AQ/A and RT/AH/A there was huge reduction in second strand synthesized product formation as shown in Figure 3-10.

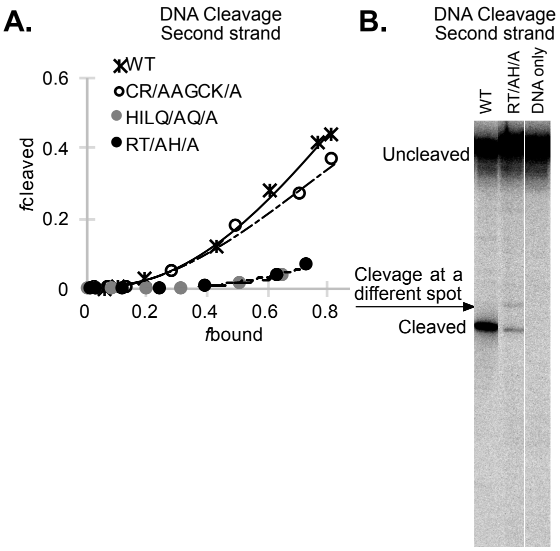


Figure 3-9: Second strand cleavage by zinc knuckle mutants. A. Scatterplot of second strand cleavage activity on linear target DNA as a function bound DNA. **B.** Denaturing gel showing RT/AH/A mutant's aberrant cleavage on

linear DNA. **C.** Scatterplot of second-strand cleavage activity on four-way junction DNA. Symbols and abbreviations are as in previous figures.

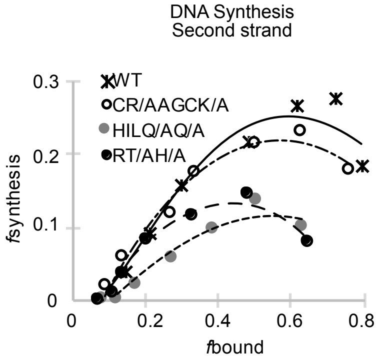


Figure 3- 10: Second strand synthesis activity of zinc knuckle mutants. Experimental setup was as in figure 3-8.

3.4 Materials and Methods

3.4.1 Protein and nucleic acid preparations

Protein was expressed and purified as previously published [17]. A QuikChange site-directed mutagenesis kit (Stratagene #200523–5) was used to generate the GR/AD/A, SR/AIR/A, SR/AGR/A, H/AIN/ALP, C/SC/SHC, CR/AAGCK/A, HILQ/AQ/A and RT/AH/A mutants. 5' PBM (320nt), 3' PBM (249 nt), linear target DNA, and 4-way junction were prepared as previously published [17].

3.4.2 R2Bm reactions and analysis

DNA binding, first and second strand cleavage, and first and second strand synthesis reactions were performed as previously reported [17].

For DNA binding assays, a mastermix containing all the components except for the protein was made and aliquoted. The binding reaction was initiated by adding 3ul of protein at the known and equalized concentrations across all proteins being tested in a data set. Duplicate reactions were prepared for each data set and two different data sets were generated, each at different protein concentrations. WT and WT KPD/A proteins acted as binding activity references and positive controls for endonuclease active and endonuclease deficient mutations, respectively.

For DNA cleavage assays, a master mix containing all the components except protein and DNA was made and aliquoted. Protein from protein dilution series was allowed to bind to RNA for 5 minutes at 37°C prior to adding the target DNA to start the cleavage reaction. The reaction was incubated for 30 minutes at 37°C. The reactions were kept on ice before running on 5% native (1X Tris-borate-EDTA) polyacrylamide gels and on denaturing (8M urea) 7% polyacrylamide gels.

First and second strand synthesis reactions contained labeled target DNA in the master mix along with all other components except for protein. Pre-cleaved linear DNA was used so that mutants deficient in DNA cleavage could be tested along with mutants with normal cleavage ability. Target DNA substrate for second strand synthesis assay was a four-way junction DNA pre-cleaved at the second strand and is described in Chapter 2. Similar to the cleavage assay the reactions were analyzed by both native and denaturing polyacrylamide gels.

All gels were dried and quantitated using a phosphorimager (Molecular dynamics STORM 840) and FIJI [28].

3.5 Discussion

3.5.1 The primary role of the linker does not appear to be binding element RNA.

The CCHC mutations reduced the accumulation of ORF2 protein into ribonucleoprotein (RNP) complex, implying a possible role in binding element RNA [23]. Likewise, sequences upstream of the presumptive α-finger were found to reduce retrotransposition activity *in vivo* [22]. Domain swapping experiments between the human and mouse L1 elements also indicate that sequence just upstream of the zinc knuckle are important for retrotransposition *in vivo* [29]. The upstream sequences are functionally linked to the zinc knuckle and other parts of the protein in a complicated yet modular way that is not well understood. A number of these domain swaps were in the middle of the presumptive α-finger. In addition, a polypeptide containing 180 amino acids of the C-terminal end of ORF2 of L1Hs containing much of the α-finger and the zinc knuckle was found to bind non-specifically to RNA *in vitro,* but mutating the cysteines did not affect nucleic acid binding [24].

Our *in vitro* study has found that mutations in the zinc knuckle and α-finger in R2Bm do not overtly reduce binding to the element 5' PBM RNA or to 3' PBM RNA. It should be noted, however, that RNA binding is inferred by the formation of distinct DNA-RNA-protein complexes in our EMSA gels [12,16]. Protein-DNA and Protein-DNA-

RNA complexes with either the 5' PBM RNA or 3' PBM RNA have unique well defined migration patterns in EMSA gels [13]. Amino acids that affect incorporation of the RNA into the protein-nucleic acid complexes can thus be detected as a change in the ratio of Protein-DNA to Protein-DNA-RNA complexes in our generic protein titration series. The RT -1 and RT 0 domains were determined to be RNA binding domains using an identical assay system [12]. RNA titrations instead of protein titrations were also carried out on several of the mutants with no indication of changes to RNA binding (data not shown). That said, an RNA binding role cannot be ruled out. The RNA binding surface might be too large and widely distributed across the surface of the R2 protein for point mutants to make an observable difference in our assays. This is one reason why double point mutants were used, instead of single point mutants [12].

Mutations to the core CCHC motif of the zinc knuckle (C/SC/SHC) and to the HINALP motif of the presumptive α-finger (H/AIN/AALP) are consistent with local disruption of protein structure leading an inability to form stable gel migrating protein-nucleic acid complexes in EMSA gels. We are unable to discern from the EMSA with these two mutants if RNA was bound or not as no distinct Protein-DNA or Protein-DNA-RNA bands were observed. All other mutations in the zinc knuckle and α-finger regions retained the ability to efficiently form the proper protein-RNA-DNA complexes in patterns similar to WT protein.

3.5.2 The linker presents nucleic acids to the RLE and RT during the first half of the integration reaction.

A comparative summary of the DNA binding, cleavage, and synthesis results for each of the mutants tested in this study is presented in Table 1. Mutations to the core of the CCHC motif (C/SC/SHC) and to the core of the HINALP motif (H/AIN/AALP) lead to an unrestrained DNA endonuclease and an inability to form stable upstream bound protein-nucleic acid complexes. All other mutants are able to form normal upstream protein-RNA-DNA complexes. Two of the α-finger mutations (SR/AIR/A and SR/AGR/A) led to the endonuclease being overly restrained and not cleaving. The inability to perform first strand cleavage was not related to the mutant's ability to bind to upstream DNA sequences as one of the mutants was unimpaired in DNA binding in the presence of 3' PBM RNA and the other mutation actually increased the protein's ability to bind to target DNA in the presence of 3' PBM RNA. Rather, residues R849, R851, R854, and R856 are used to position the target DNA and/or the DNA endonuclease for first-strand DNA cleavage.

Once cleaved, α-finger GR/AD/A and SR/AIR/A mutants were unable to perform first strand cDNA synthesis (TPRT) on pre-nicked target DNA indicating a role of the mutated residues in positioning the RT and/or nucleic acid components relative to each other. Indeed, the GR/AD/A mutant lacked any other major phenotype beyond the inability to perform TPRT and a modest reduction in binding to upstream DNA sequences. The zinc knuckle mutants CR/AAGCK/A, HILQ/AQ/A, and RT/AH/A modestly reduced first strand DNA cleavage and retained near wild type first-strand DNA synthesis activity. Upstream DNA binding was not carefully examined but appeared to be normal.

3.5.3 The linker region is key to the second half of the integration reaction.

The second half of the integration reaction begins with R2 protein being associated with the 5' PBM RNA and thus becoming bound to DNA sequences downstream of the insertion site on linear target DNA. Mutations to the core of the CCHC motif (C/SC/SHC) and to the core of the HINALP motif (H/AIN/AALP) lead to an unrestrained DNA endonuclease and an inability to form stable downstream bound protein-nucleic acid complexes. All other mutants were able to form normal downstream protein-RNA-DNA complexes on linear target DNA and appeared to have minimal effect on binding to linear DNA. That said, the SR/AIR/A mutation did show a modest decrease in binding to the downstream sequence on linear DNA and the zinc knuckle mutants were not quantitatively tested.

The second half of the integration only proceeds when the downstream subunit is in the "no-RNA-bound" state [16]. Although second-strand DNA cleavage can occur on linear DNA, it requires a complicated set of 5' RNA, DNA, and protein ratios to do so and is non-productive in the sense that second-strand synthesis does not occur [13,16]. For this reason, it is now thought that the second half of the integration reaction, specifically second-strand DNA cleavage and second-strand synthesis, mechanistically requires the formation of the 4-way junction (see chapter 2). The 4-way junction appropriately cleaves the junction in the absence of RNA and the cleaved product is a substrate for second strand synthesis (see chapter 2).

All of the zinc knuckle and α-finger mutants tested, except for the CR/AAGCK/A mutant, were unable to perform second-strand cleavage on linear DNA (Table 1), yet, importantly, the zinc knuckle mutants did not impair second-strand cleavage on the more important 4-way junction. The α-finger mutations that lie closest to the zinc knuckle, SR/AIR/A and SR/AGR/A, greatly reduce binding to the 4-way junction and abolish second-strand DNA cleavage. Second-strand synthesis was similarly affected by the two sets of mutations. The results indicate that the α-finger is critical for 4-way junction recognition as well as presenting the bound DNA to the endonuclease and to the reverse transcriptase. The zinc knuckle mutants HILQ/AQ/A and RT/AH/A severely reduced second-strand synthesis indicating that the zinc knuckle residues are involved in positioning the cleaved junction and/or the reverse transcriptase for primer extension.

3.5.4 Structural and functional connections to APE LINEs and to Prp8

The protein encoded by R2Bm has been determined to consist of two globular domains. The larger of the two domains (colored in Figure 3-11) contains the RT, the RLE, and a region between the two called the linker [6]. The end of the linker region contains an invariant zinc knuckle and several conserved helices upstream of the zinc knuckle. In this paper, the upstream helices have been referred to as the "presumptive α-finger" of which the HINALP motif is central to the α-finger in R2Bm. APE LINEs also contain a "linker" with a presumptive α-finger and a zinc knuckle located beyond the RT (Figure 3-11).

The large globular domain of R2Bm, an RLE LINE, shares structural as well as sequence similarities to the large fragment of eukaryotic splicing factor Prp8 (see Figure 3-11). Prp8 has an RT, an RLE, and a linker region between the RT and RLE. Towards the end of the linker region in Prp8 is a non-zinc knuckle structure. Upstream of the non-zinc knuckle is a set of helices that align with the helices found upstream of the zinc knuckle in LINEs. The helices upstream of the non-zinc knuckle in Prp8 form a very prominent and important α-finger. The α-finger protrudes out over the reverse transcriptase (see Figure 3-11C) [26]. It is by analogy to the α-finger in Prp8 that the corresponding region of the RLE LINEs is called the "presumptive α-finger" [6]. In Prp8 the non-zinc knuckle, the α-finger, and the RT thumb work together to bind the splice sites and spliceosomal RNAs. The non-zinc knuckle and the α-finger are dynamic in Prp8 undergoing/promoting protein and protein-RNA confrontational changes across all aspects of the splicing reaction. Of particular interest is the fact that in the U4/U6.U5 tri-snRNP and in the B complex the α-finger and non-zinc knuckle bind to critical branched RNA structures.

The data reported here indicate that whatever the actual structure of the R2Bm linker is, the linker is central to the recognition of the 4-way junction integration intermediate. It also acts as a protein-DNA conformational switch or hub for correctly positioning the EN, the RT, and the substrate DNA relative to each other.

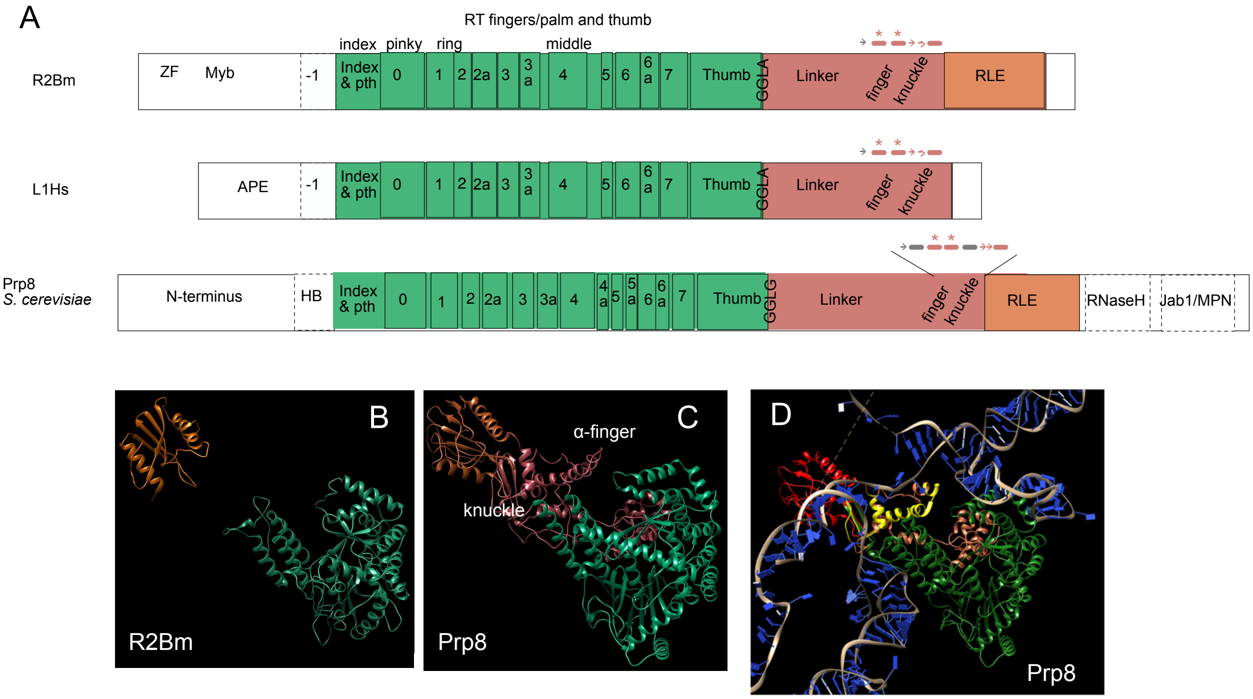


Figure 3-11: Similarities between R2Bm and Prp8. A. The ORF structure of R2Bm, human L1 (L1Hs), and *Saccharomyces cerevisiae* Prp8 are presented as color block diagrams [6,25,26,30–33]. The RT is green, the linker is maroon, and the RLE is orange. In the linker region, the sequences of the orange colored α-helices (rounded bars) with an asterisk align well. Remaining of the colored α-helix and β-strands (arrows) (may) form a structurally similar knuckle. **B.** Model of R2Bm's RT and RLE [6]. The ribbons have been colored as in the corresponding color block diagram. **C.** Cryo-Em structure of the large fragment of Prp8 [25]. Ribbon color is matching the corresponding color block diagram. **D.** Cryo-EM structure of the Prp8 and RNA from the B spliceosome complex [26]. Reverse transcriptase is colored in green, RLE in red, and the linker region in orange except for the α-finger and non-zinc knuckle shown in yellow. A branched structure formed by the RNA components of spliceosome is also shown.

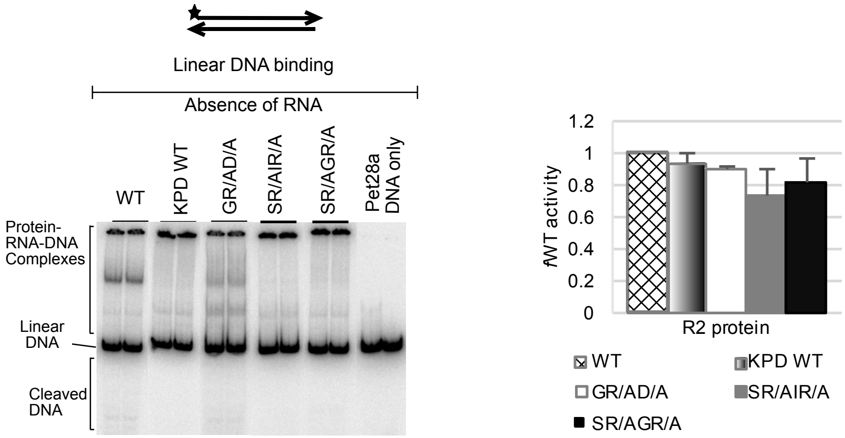
Table 3- 1: Summary of DNA binding, cleavage, and synthesis results.

	Linear					Junction			either
	DNA binding (3' PBM)	First strand cleavage	First strand synthesis	DNA binding (5' PBM)	Second strand cleavage	DNA binding	Second strand cleavage	Second strand synthesis	Non-R2 site cleavage
GR/AD/A	-	WT	∅	WT	WT	WT	WT	WT	None
SR/AIR/A	WT	∅	∅	-	∅	- - -	∅	- -	None
SR/AGR/A	++	∅	- -	WT	∅	- -	∅	∅	None
H/AIN/AALP	- - -	∅	N.A.	- - -	∅	N.T.	N.T.	N.A.	Yes
C/SC/SHC	- - -	∅	N.A.	- - -	∅	N.T.	N.T.	N.A.	Yes
CR/AAGCK/A	N.T.	-	WT	N.T.	-	N.T.	WT	WT	None
HILQ/AQ/A	N.T.	-	WT	N.T.	∅	N.T.	WT	- - -	None
RT/AH/A	N.T.	-	WT	N.T.	∅	N.T.	WT	- - -	Yes

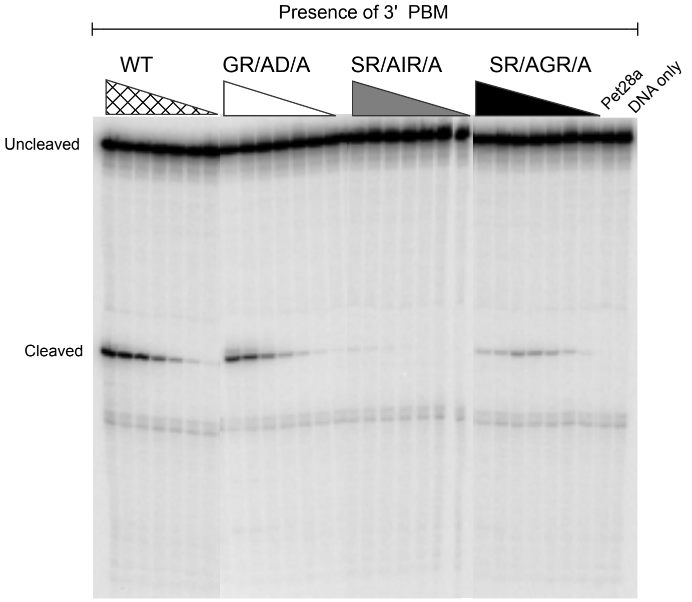
Not Applicable (N.A.), Not tested (N.T.)

"++"	: + 30% and above	
"+"	: +15% to 30%	
"WT"	: 15% to -15% of WT activity	: functionally WT
"-"	: -15% to -30%	: modest reduction
"- -"	: -30% to -50%	: major reduction
"- - -"	: -50% to 75%	: severe reduced
"∅"	: 75% and above	: functionally dead

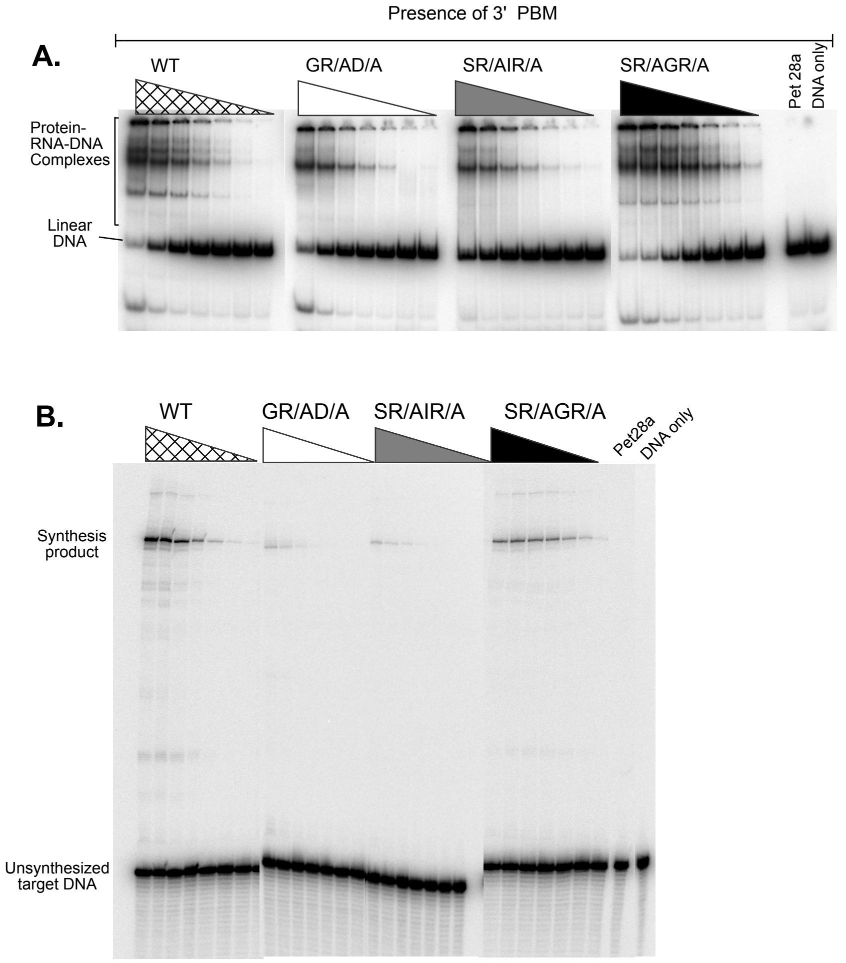
3.6 Supplementary files:



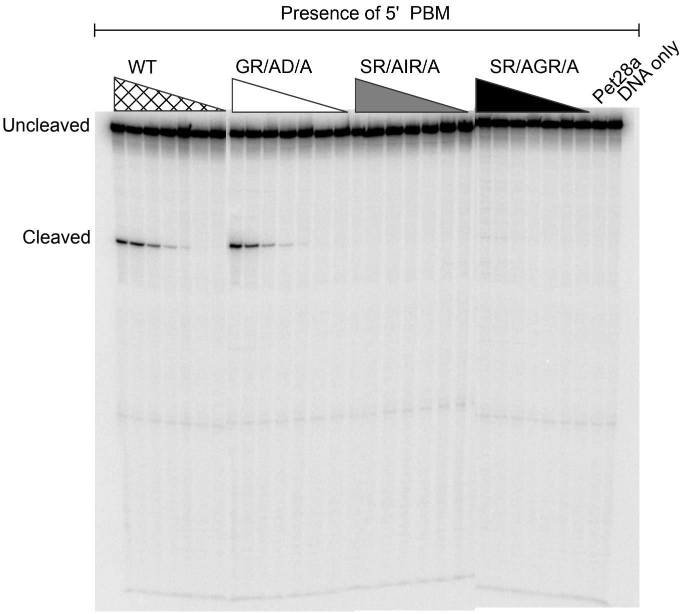
Supplementary figure 3-1: Linear Target DNA binding activity of α-finger mutant proteins in the absence of RNA. The symbol and abbreviations are as in Figure 3-3.



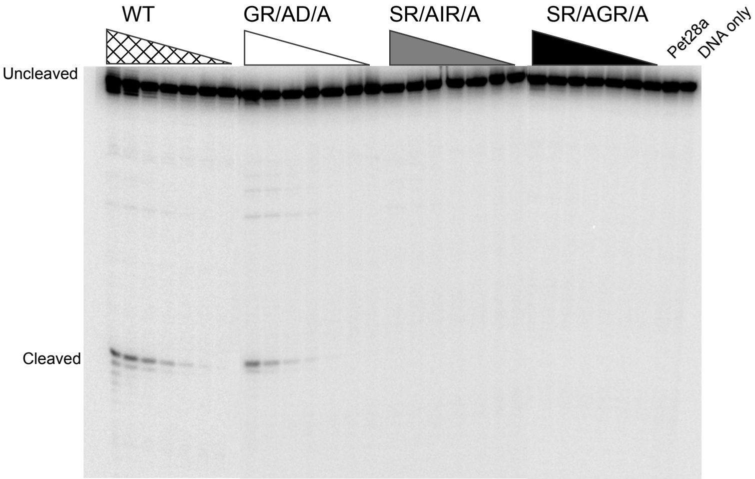
Supplementary figure 3-2: Denaturing gel showing first strand cleavage activity of α-finger mutant proteins. Gel shows uncleaved and cleaved linear DNA signals for seven protein titrations indicated by triangles in the presence of 3’ PBM RNA.



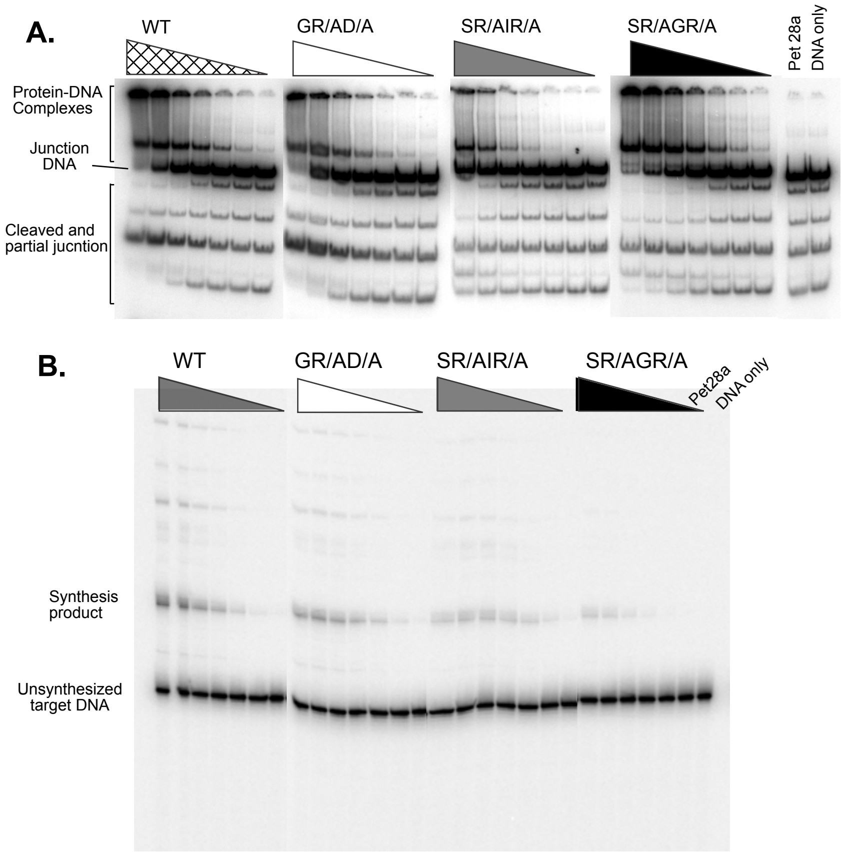
Supplementary figure 3-3: First strand synthesis (TPRT) activity of α-finger mutant proteins. A. EMSA gel used to calculate the fraction of target DNA bound by the R2 protein. **B.** Denaturing gel used to calculate the fraction of target DNA that undergoes first strand synthesis by TPRT.



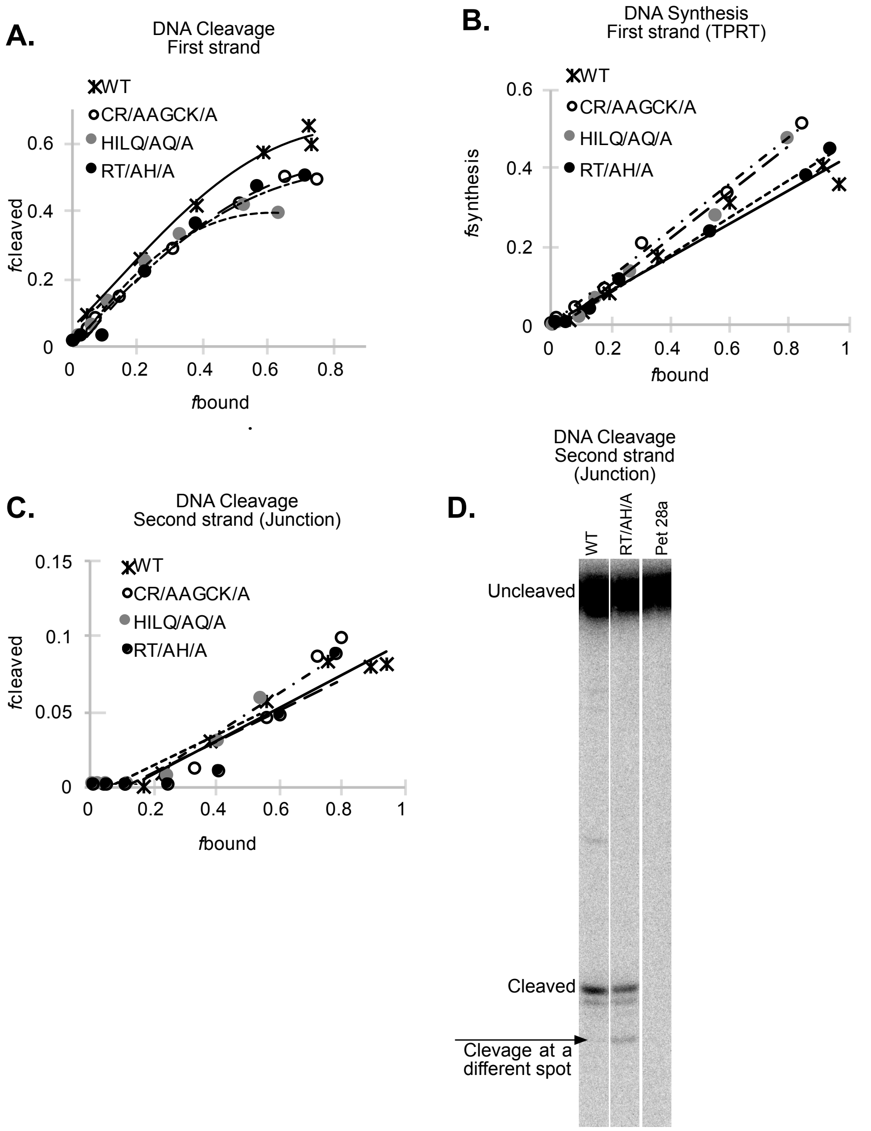
Supplementary figure 3-4: Denaturing gel showing second strand cleavage activity of α-finger mutant proteins. Gel shows uncleaved and cleaved linear DNA signals for seven protein titrations indicated by triangles in the presence of 5' PBM RNA.



Supplementary figure 3-5: Denaturing gel showing second strand cleavage activity of α-finger mutant proteins on a 4-way junction target DNA.



Supplementary figure 3-6: Second strand synthesis activity of α-finger mutant proteins on a pre-cleaved 4-way junction target DNA. A. EMSA gel used to calculate the fraction of target DNA bound by the R2 protein. **B.** Denaturing gel used to calculate the fraction of pre-cleaved 4-way junction target DNA that undergoes second strand synthesis.



Supplementary figure 3-7: Cleavage and synthesis activity of zinc knuckle mutant protein. A. First strand cleavage activity of zinc knuckle mutant proteins. Scatter plot of fraction of cleaved target DNA (*f*cleaved) plotted as a function of fraction of target DNA bound by protein (*f*bound) at each protein concentrations. **B.** First strand synthesis activity of zinc knuckle mutant proteins. The graph plots fraction of target DNA that undergoes first strand synthesis by TPRT (*f*synthesis) as a function of fraction of pre-cleaved linear target DNA bound by the protein (*f*bound). **C.** Second strand cleavage activity of zinc knuckle mutants on a 4-way junction target DNA. Scatter plot of target DNA cleaved at the second strand (*f*cleaved) as a function of fraction of 4-way junction DNA bound by the protein (*f*bound). **D.** Denaturing gel for second strand cleavage activity on a 4-way junction target DNA**.** The gel shows uncleaved and cleaved product formed by second strand cleavage on the junction DNA along with a non-specific cleavage at a non-R2 insertion site.

3.7 References:

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Ivancevic, A. M., Kortschak, R. D., Bertozzi, T. & Adelson, D. L. LINEs between Species : Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. **8,** 3301–3322 (2016).
3. Burton, F. H. *et al.* Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. *J. Mol. Biol.* (1986). doi:10.1016/0022-2836(86)90235-4
4. Jakubczak, J. L., Xiong, Y. & Eickbush, T. H. 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.* (1990). doi:10.1016/0022-2836(90)90303-4
5. Matsumoto, T., Hamada, M., Osanai, M. & Fujiwara, H. Essential Domains for Ribonucleoprotein Complex Formation Required for Retrotransposition of Telomere-Specific Non-Long Terminal Repeat Retrotransposon SART1. *Mol. Cell. Biol.* **26,** 5168–5179 (2006).
6. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: Similarities to eukaryotic splicing factor Prp8. *Mob. DNA* **8,** 1–15 (2017).
7. Kojima, K. K., Kuma, K. I., Toh, H. & Fujiwara, H. Identification of rDNA-specific non-LTR retrotransposons in cnidaria. *Mol. Biol. Evol.* (2006). doi:10.1093/molbev/msl067
8. Gladyshev, E. A. & Arkhipova, I. R. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* (2009). doi:10.1016/j.gene.2009.08.016
9. Burke, W. D., Calalang, C. C. & Eickbush, T. H. 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.* (1987). doi:10.1128/MCB.7.6.2221.Updated
10. Yang, J., Malik, H. S. & Eickbush, T. H. 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,** 7847–52 (1999).
11. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).
12. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res.* **42,** 8405–8415 (2014).
13. Christensen, S. M. & Eickbush, T. H. 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).
14. Kierzek, E. *et al.* 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.* (2008). doi:10.1093/nar/gkm1085
15. Kierzek, E. *et al.* 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 (2009).
16. Christensen, S. M., Ye, J. & Eickbush, T. H. 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).
17. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: Loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res.* **44,** 3276–3287 (2016).
18. Thompson, B. K. & Christensen, S. M. Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons. *Mob. Genet. Elements* **1,** 29–37 (2011).
19. Shivram, H., Cawley, D. & Christensen, S. M. 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 (2011).
20. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
21. Fanning, T. & Singer, M. The line-1 DNA sequences in four mammalian orders predict proteins that conserve

homologies to retrovirus proteins. *Nucleic Acids Res.* (1987). doi:10.1093/nar/15.5.2251

22. Moran, J., Holmes, S. & Naas, T. High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
23. Doucet, A. J. *et al.* Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet.* **6,** 1–19 (2010).
24. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3,** 433–437 (2013).
25. Wan, R. *et al.* The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science (80-.).* (2016). doi:10.1126/science.aad6466
26. Bertram, K. *et al.* Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. *Cell* (2017). doi:10.1016/j.cell.2017.07.011
27. Christensen, S. & Eickbush, T. H. Footprint of the Retrotransposon R2Bm Protein on its Target Site before and after Cleavage. *J. Mol. Biol.* **336,** 1035–1045 (2004).
28. Schindelin, J. *et al.* Fiji: An open source platform for biological image analysis. *Nat. Methods* (2012). doi:10.1038/nmeth.2019.Fiji
29. Wagstaff, B. J., Barnerßoi, M. & Roy-Engel, A. M. Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. *PLoS One* **6,** (2011).
30. Qu, G. *et al.* Structure of a group II intron in complex with its reverse transcriptase. *Nat. Struct. Mol. Biol.* (2016). doi:10.1038/nsmb.3220
31. Nguyen, T. H. D. *et al.* Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530,** 298–302 (2016).
32. Galej, W. P., Nguyen, T. H. D., Newman, A. J. & Nagai, K. Structural studies of the spliceosome: Zooming into the heart of the machine. *Current Opinion in Structural Biology* (2014). doi:10.1016/j.sbi.2013.12.002
33. Blocker, F. J. H. *et al.* Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* **11,** 14–28 (2005).

Chapter 4

Investigation on R-box Region of Restriction Like Endonuclease Encoded by R2 LINE

Monika Pradhan and Shawn M. Christensen

4.1 Abstract

Long Interspersed Elements (LINEs) are a major group of eukaryotic transposable elements that replicate in the genome by Target Primed Reverse Transcription (TPRT). A model system to study LINE integration mechanism by TPRT is R2 LINE that bear restriction like endonuclease (RLE) and inserts into specific target site of 28S rRNA gene. R2 elements encode a multifunctional protein with reverse transcriptase, endonuclease, and nucleic acid binding activities. R-box is a conserved arginine rich region located at the start of the RLE domain of R2 open reading frame and in *Bombyx mori,* the R-box region consists of RILRH residues. The RH has been identified as DNA binding residues in this region, while the single H residue is important for target DNA cleavage required for TPRT. In this study, we investigate the functional significance of each arginine residue in the R-box region by generating single point mutations and subjecting them to biochemical assays *in vitro*. The single mutations did not alter the DNA binding, first strand cleavage and TPRT product forming activities of R2 protein. Both the arginine residues appear to be important for second strand cleavage activity. The R-box arginine residues could possibly be involved in correctly positioning the target DNA towards the active site of R2 endonuclease required for DNA cleavage. Additional functional assays are required to accurately dissect the role of arginine residues in the R-box region.

4.2 Introduction

R2 is a site specific Long Interspersed Element (LINE) that inserts into the 28S ribosomal RNA (rRNA) genes of the host [1,2]. They are widely distributed in most lineages of arthropods and are also found in other taxa of animals like nematodes and tunicates[3–5]. Like all other LINEs, R2 elements integrate into the host genome by a process called target primed reverse transcription (TPRT) [6–8]. Endonuclease encoded by R2 element cleaves the target 28S rDNA sequence to expose a 3'-OH group at the cleavage site. The R2 element encoded reverse transcriptase uses the free 3'-OH group to prime reverse transcription of the element RNA into DNA at the insertion site [6–9]. R2 encodes a single open reading frame (ORF) with N-terminal DNA binding zinc finger (ZF) motif and myb motif, RNA binding (RB) motif; central reverse transcriptase (RT); and C-terminal linker region with predicted α-helices (HINALP motif) and gag-like zinc knuckle (CCHC motif), and type II restriction like endonuclease (RLE) (Figure 4-1A)[10–14]. The single ORF can be easily expressed in bacteria and the protein can be readily purified to be subjected to important biochemical assays to study LINE integration mechanism by TPRT. R2 integration mechanism has been most studied in the R2 element from *Bombyx mori* (R2Bm).

The R2 protein bound to 3' protein binding motif (PBM) of R2 RNA binds to the upstream sequences of the insertion site by using a DNA binding motif that is yet to be identified. The R2 protein bound to 5' PBM of R2 RNA binds downstream of the insertion site by using R2 protein's ZF and myb motifs [12]. R2 integration into the 28S rDNA site is catalyzed by the upstream and the downstream protein subunits. (Figure 4-1B). Endonuclease from the upstream subunit nicks the first strand exposing 3'-OH on the target DNA that acts as a primer for cDNA synthesis (Figure 4-1B step1)[6]. The 3'-OH is then used by the reverse transcriptase as a primer for first strand synthesis with R2 RNA as the template (Figure 4-1B step2). The reverse transcriptase removes the 5' PBM from this subunit and template jumps from RNA to the upstream target DNA sequences forming a 4-way junction target DNA structure (Figure 4-1B step3). The endonuclease from downstream subunit cleaves the second strand (Figure 4-1B step 4) [15]. Finally, the reverse transcriptase from the downstream subunit catalyzes the second strand to complete element integration (Figure 4-1B step 5) (chapter 2).

The endonuclease encoded by R2 elements is a variant of the PD-(D/E)XK superfamily of endonucleases and the PD-(D/E)XK serves as the catalytic core motif of many restriction-like endonuclease-fold (RLE) family of nucleases [11]. The PD-(D/E)XK superfamily comprises a large group of proteins that share structurally conserved consensus core with a four-stranded mixed β-sheet flanked by α-helix on each side (αβββαβ) [16–19]. There is an arginine rich region, also called as the "R-box" that consists of RXXR or NXRXXR found within or in front of the first α-helix in a subset of type II restriction enzymes (e.g., EcoRII, DpnII, MboI, PspGI, and Sso II) [20–22]. Mutational studies have identified the R-box region to be critical for DNA binding and cleavage [21,22]. Amino acids in the R-box region in some type II restriction enzymes are thought to be involved in making contacts to the region at the 3′-end of the recognition sequence of DNA [21,22]. Restriction enzymes that lack R-box residues uses the same region to interact with target DNA in order to correctly position the DNA at the cleavage site. A R-box (RXXR) was recently identified in R2 element and is located at the beginning of the first α-helix of the αβββαβ catalytic core [19]. This region is located immediately after the gag-like zinc knuckle (CCHC motif) that is identified to make contacts with the target DNA especially the 4-way junction intermediate structure required for second strand cleavage and synthesis (Chapter 3). In R2Bm element, the R-box region is formed by "RILRH" residues (Figure 4-1C). When the second arginine and the following histidine were mutated (RILR/AH/A), binding to the upstream and downstream target DNA sequences were reduced by 40% and 30%, respectively [19]. Mutating only the histidine residue (RILRH/A) slightly reduced the ability of the protein to cleave the first strand [19]. While the DNA binding activity of the second arginine residue was indirectly tested as the RH double mutant (RILR/AH/A), we expected that mutating the arginine residue alone could alter the target DNA interaction. The first arginine residue in the R-box was not tested and could also play a role in target DNA binding. This study investigates the role of the two arginine residues in the R-box by subjecting the single mutants (R/AILRH and RILR/AH) to different functional assays. Both R/AILRH and RILR/AH did not seem to alter the ability of R2 protein to bind to the target DNA, however, they seem to drastically reduce the ability of the protein to conduct second strand cleavage, indicating their possible role in precisely positioning the target DNA at the catalytic site of endonuclease.

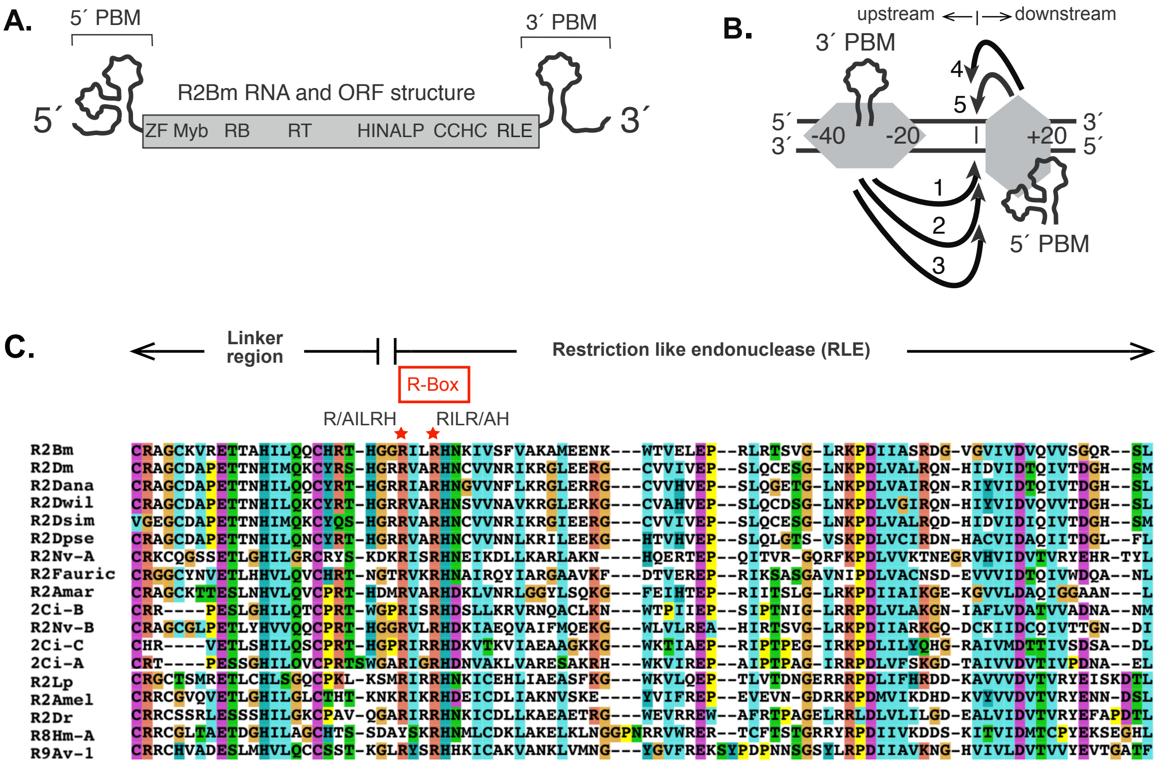

A.
5´ PBM
3´ PBM
R2Bm RNA and ORF structure
5´
3´
ZF Myb RB RT HINALP CCHC RLE
B.
upstream
downstream
3´ PBM
4
5
1
2
3
5´
3´
-40
-20
+20
3´
5´
5´ PBM
C.
Linker region
Restriction like endonuclease (RLE)
R-Box
R/AILRH
RILR/AH
R2Bm CRAGCKVRETTAHILQQCHRT-HGGRILRHNKIVSFVAKAMEENK---WTVELEP--RLRTSVG-LRKPDIIASRDG-VGVIVDVQVVSGQR--SL
R2Dm CRAGCDAPETTNHIMQKCYRS-HGRRVARHNCVVNRIKRGLEERG---CVVIVEP--SLQCESG-LNKPDLVALRQN-HIDVIDTQIVTDGH--SM
R2Dana CRAGCDAPETTNHILQQCYRT-HGRRIARHNGVVNFLKRGLERRG---CVVHVEP--SLQGETG-LNKPDLVAIRQN-RIYVIDTQIVTDGH--SL
R2Dwil CRAGCDAPETTNHILQKCYRT-HGRRVARHNSVVNAVKRGLERKG---CVAHVEP--SLQCDSG-LNKPDLVGIRQN-HIYVIDVQVVTDGH--SL
R2Dsim VGEGCDAPETTNHIMQKCYQS-HGRRVARHNCVVNRIKRGIEERG---CVVIVEP--SLQCESG-LNKPDLVALRQD-HIDVIDIQIVTDGH--SM
R2Dpse CRAGCDAPETTNHILQNCYRT-HGRRVARHNCVVNNLKRILEEKG---HTVHVEP--SLQLGTS-VSKPDLVCIRDN-HACVIDAQIITDGL--FL
R2Nv-A CRKCQGSSETLGHILGRCRYS-KDKRISRHNEIKDLLKARLAKN----HQERTEP--QITVR-GQRFKPDLVVKTNEGRVHVIDVTVRYEHR-TYL
R2Fauric CRGGCYNVETLHHVLQVCHRT-NGTRVKRHNAIRQYIARGAAVKF---DTVEREP--RIKSASGAVNIPDLVACNSD-EVVVIDTQIVWDQA--NL
R2Amar CRAGCKTTESLNHVLQVCPRT-HDMRVARHDKLVNRLGGYLSQKG---FEIHTEP--RIITSLG-LRKPDIIAIKGE-KGVVLDAQIGGAAN---L
2Ci-B CRR----PESLGHILQTCPRT-WGPRISRHDSLLKRVRNQACLKN---WTPIIEP--SIPTNIG-LRRPDLVLAKGN-IAFLVDATVVADNA--NM
R2Nv-B CRAGCGLPETLYHVVQQCPRT-HGGRVLRHDKIAEQVAIFMQEKG---WLVLREA--HIRTSVG-LRKPDIIARKGQ-DCKIIDCQIVTTGN--DI
2Ci-C CHR----VETLSHILQSCPRT-HGPRIRRHDKVTKVIAEAAGKKG---WKTIAEP--RIPTPEG-IRKPDLILYQHG-RAIVMDTTIVSDSA--DL
2Ci-A CRT----PESSGHILQVCPRTSWGARIGRHDNVAKLVARESAKRH---WKVIREP--AIPTPAG-IRRPDLVFSKGD-TAIVVDVTIVPDNA--EL
R2Lp CRGCTSMRETLCHLSGQCPKL-KSMRIRRHNKICEHLIAEASFKG---WKVLQEP--TLVTDNGERRRRPDLIFHRDD-KAVVVDVTVRYEISKDTL
R2Amel CRRCGVQVETLGHILGLCTHT-KNKRIKRHDEICDLIAKNVSKE----YVIFREP--EVEVN-GDRRKPDMVIKDHD-KVYVVDVTVRYENN-DSL
R2Dr CRRCSSRLESSSHILGKCPAV-QGARIRRHNKICDLLKAEAETRG---WEVRREW--AFRTPAGELRRLDLVLILGD-EALVIDVTVRYEFAPDTL
R8Hm-A CRRCGLTAETDGHILAGCHTS-SDAYSKRHNMLCDKLAKELKLNGGPNRRVWRER--TCFTSTGRRYRPDIIVKDDS-KITVIDMTCPYEKSEGHL
R9Av-1 CRRCHVADESLMHVLQCCSST-KGLRYSRHHKICAKVANKLVMNG---YGVFREKSYPDPNNSGSYLRPDIIAVKNG-HVIVLDVTVVYEVTGATF

Figure 4-1: R2Bm structure, integration model and multiple sequence alignment. **A.** R2Bm RNA structure. Box in the middle represents open reading frame (ORF) with conserved motifs. Abbreviations: zinc finger (ZF), Myb (Myb), RNA binding (RB), reverse transcriptase domain (RT), a conserved predicted α-helices (HINALP motif), a gag-like zinc knuckle (CCHC motif) and a PD-(D/E)XK type II restriction-like endonuclease (RLE). R2Bm RNA segments corresponding to the 5' and 3' untranslated regions known to adopt distinct structures and bind the R2Bm protein are labelled within the brackets as 5' and 3' protein binding motifs (PBMs), respectively. **B.** R2 target site, 28S rDNA, and insertion model. R2 protein associated with the 3' PBM RNA binds 20 to 40 bases upstream (28Su) of the insertion site (vertical line) and protein associated with the 5' PBM RNA binds to 20 bases downstream of the insertion site [12,23]. Insertion occurs in five steps. **C.** Clustal alignment of R2 sequences showing the R-box region from diverse arthropods (12 sequences), a vertebrate *Danio rerio* (R2Dr), R8 from *Hydra magnipapillata* (R2Hm-A) and R9 from

Adineta vaga (R9Av-1). Red star represents the Arginine residues mutated in this study. The point mutations generated in this study are: R/AILRH and RILR/AH. Abbreviations: R2Bm = Bombyx mori, R2Dm = Drosophila melanogaster, R2Dana = Drosophila ananassae, R2Dwil = Drosophila willistoni, R2Dsim = Drosophila simulans, R2Dpse = Drosophila pseudoobscura, R2Fauric = Forficula auricularia, R2Amar = Anurida maritima, R2Nv-B = Nasonia vitripennis, R2Lp = Limulus polyphemus, R2Ci = Ciona intestinalis, R2Amel = Apis mellifera, R2Dr = Danio rerio, R8Hm-A = Hydra magnipapillata, R9Av-1 = Adineta vaga

4.3 Results

4.3.1 Single mutations in the R-box region do not seem to decrease binding to linear target DNA

In order to determine whether the arginine residues in the R-box region were involved in target DNA binding, the residues were mutated to alanine by site directed mutagenesis. The single point mutations generated were R/AILRH and RILR/AH. The linear DNA binding ability of the two R-box mutants were assayed relative to WT protein using Electrophoretic Mobility Shift Assays (EMSAs) (Figure 4-2A, B, and C). The DNA was pre-cleaved on the first strand and labeled on the second/top strand to track the protein-DNA/ protein-RNA-DNA complexes formation in EMSA gel. The EMSA gel was used to quantify the fraction of target DNA bound by R2 protein and the DNA binding efficacy is presented in the bar graph. DNA binding activity of WT protein (*f*WT activity) was set to 1 and relative activity of mutant proteins were calculated. Duplicate binding reactions were prepared and duplicate lanes were run in the gel. Vector control (Pet28a) and DNA only lanes represent the negative controls.

The upstream DNA binding ability of R/AILRH and RILR/AH mutants were tested in the presence of 3' PBM RNA and the downstream DNA binding ability was tested in the presence of 5' PBM RNA. Both upstream and downstream DNA binding ability for R/AILRH and RILR/AH mutant proteins were found to be very similar to that of WT protein. R/AILRH and RILR/AH mutant proteins decreased the upstream DNA binding activity of R2 protein by only 4% and 2%, respectively (Figure 4-2A). R/AILRH mutant protein reduced binding to the downstream DNA sequences by 12% while R/AILRH mutant protein showed the reduction of only 1% (Figure 4-2B). In the absence of RNA, both the mutant proteins were as efficient as WT in binding the target DNA, with only a decrease of about 3% to 6% (Figure 4-2C). Overall, both R/AILRH and RILR/AH mutations did not significantly affect the DNA binding ability of the R2 protein.

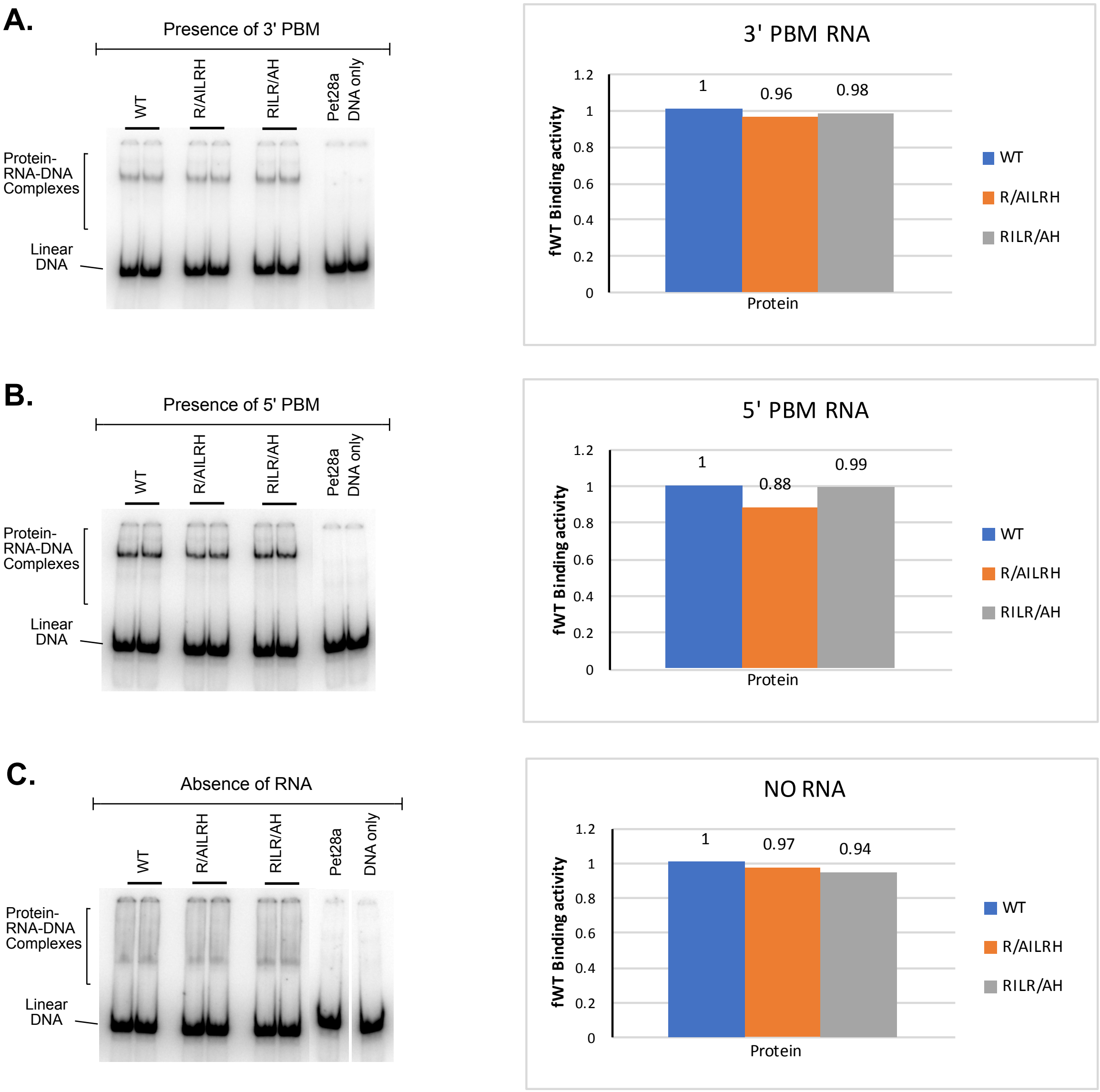


Figure 4-2: Target DNA binding activity of R-box mutants. In the bar graph, WT DNA binding activity (*f*WT activity) was set to 1 and relative activity of mutants were calculated. **A.** Electrophoretic mobility shift assay (EMSA) gel and bar graph showing upstream target DNA binding ability of R-box mutants in the presence of 3' PBM RNA. **B.** EMSA gel and bar graph showing downstream target DNA binding ability of R-box mutants in the presence of 5' PBM RNA. **C.** EMSA gel and bar graph showing target DNA binding ability of R-box mutants in the absence of RNA. DNA binding activity of WT, R/AILRH, and RILR/AH mutants are reported as blue, orange and grey bars, respectively.

4.3.2 Single mutations in the R-box region do not decrease first strand cleavage and first strand synthesis

The ability of the R-box mutants to cleave the first strand and conduct first strand synthesis by TPRT were assayed. The R2 proteins were pre-bound with 3' PBM RNA and incubated with linear target DNA labeled on bottom strand to track cleavage and first strand synthesis. In synthesis assay, dNTPs were also added. Aliquots of the cleavage and synthesis assay reactions were run in EMSA and denaturing gels.

The first strand cleavage activity is represented in a scatterplot of fraction of target DNA bound by R2 protein (*f*bound) plotted as a function of fraction of target DNA cleaved at the first strand (*f*cleaved) (Figure 4-3A). *f*bound was quantitated from EMSA gel and *f*cleaved from denaturing gel. Both R/AILRH and RILR/AH mutant proteins do not show any effect on the ability of the protein to cleave the first strand. Cleavage on target DNA was observed only at the insertion site.

The first strand synthesis activity is shown in a scatter plot of fraction of the cleaved target DNA that undergoes TPRT reaction (*f*TPRT) plotted as a function of fraction of target DNA cleaved at the first strand (*f*cleaved) (Figure 4-3B). Both *f*cleaved and *f*TPRT were quantitated from denaturing gel. Both R/AILRH and RILR/AH mutant proteins were as efficient as WT at synthesizing the first strand (Figure 4-3B).

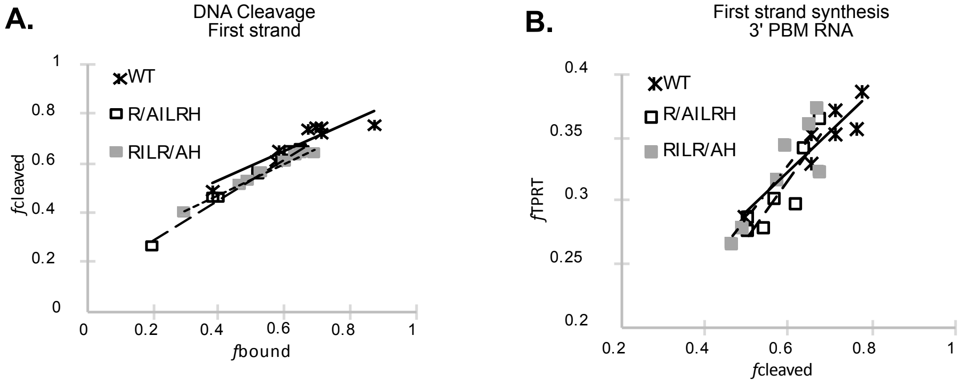


Figure 4- 3: First strand target DNA cleavage and first strand synthesis activity of R-box mutants. The reactions were prepared for a range of protein titrations and aliquots were loaded onto EMSA and urea denaturing gels. **A.** Scatter plot for first strand cleavage activity. The fraction of target DNA cleaved at the first strand (*f*cleaved) plotted as a function of fraction of target DNA bound by protein (*f*bound) at each protein concentrations. *f*bound is calculated from the EMSA gel and *f*cleaved is quantitated from denaturing gel. Data points for WT, R/AILRH, and RILR/AH are represented by asterisk, open square, and grey square, respectively. **B.** Scatter plot for first strand synthesis activity. The fraction of the cleaved target DNA that undergoes first strand synthesis by TPRT (*f*TPRT) is plotted as a function of fraction of target DNA cleaved at the first strand (*f*cleaved) at each protein concentrations. *f*TPRT and *f*cleaved are calculated from the denaturing gel. Data points for WT, R/AILRH, and RILR/AH are as indicated before.

4.3.2 Single mutations in the R-box region decrease second strand cleavage on linear target DNA

In the presence of 5’ PBM RNA, the R2 protein binds to the downstream sequences of the target DNA and cleaves the second strand. To test if R/AILRH and RILR/AH mutant proteins affect the second strand cleavage activity of R2 protein, cleavage assay was carried out in which the protein were pre-bound with 5’ PBM RNA and incubated with target DNA. The linear target DNA was labeled on the second strand to track cleavage. The reactions were run in EMSA and denaturing gel for a series of protein titrations.

In EMSA gel, the R/AILRH and RILR/AH mutant proteins could form the typical protein-DNA-RNA complexes as seen for WT (Figure 4-4A). In WT protein, there is an additional complex below the typical band that corresponds to the RNP complex bound to the double strand cleaved target DNA which is not seen for both the mutants. In denaturing gel, R/AILRH and RILR/AH mutants showed significant reduction in the formation of second strand cleaved target DNA (Figure 4-4B). The second strand cleavage activity of R-box mutants is shown in a scatterplot where fraction of the target DNA that undergoes second strand cleavage (*f*cleavage) is plotted as a function of fraction of linear target DNA bound by R2 protein (*f*bound). Both R/AILRH and RILR/AH mutant proteins were found to drastically reduce second strand cleavage activity of R2 protein on a linear target DNA when compared to WT protein (Figure 4-4C).

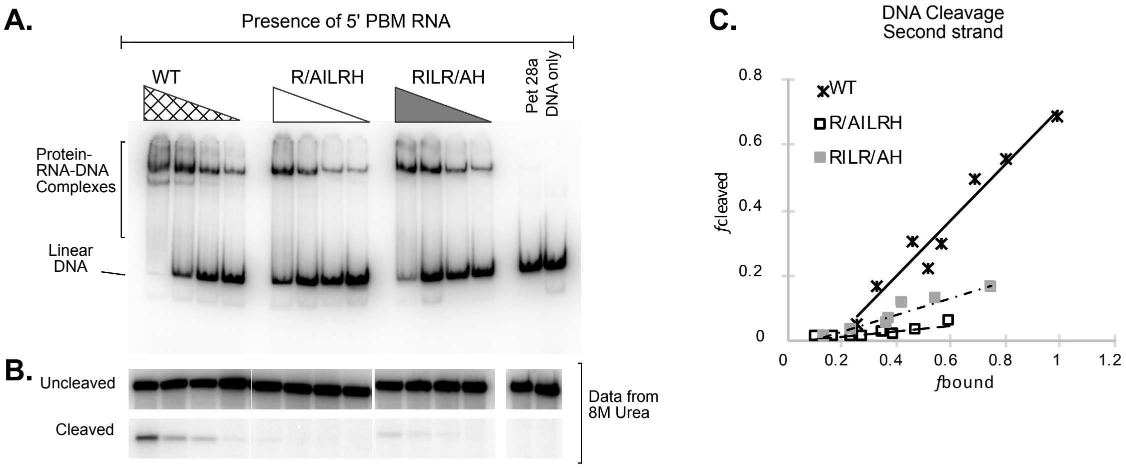


Figure 4- 4: Second strand cleavage activity of R-box mutants. Aliquots of the reactions were prepared for a series of protein titrations (represented by triangles) and were loaded in the EMSA and denaturing gels. **A.** EMSA gel used to calculate the fraction of target DNA bound by R2 protein (*f*bound). **B.** Denaturing gel used to calculate the fraction of target DNA cleaved by the R2 protein (*f*cleaved). **C.** Scatter plot of second strand cleavage activity. Symbols and abbreviations are as in previous figures.

4.4 Discussion

R2 element encoded RLE belongs to PD-(D/E)XK superfamily of endonucleases which also includes most bacterial restriction enzymes and certain Holiday junction resolvases. Protein structure prediction and threading programs like Phyre2 and HHPRED have identified structural similarity of R2 endonuclease with Fok I restriction endonuclease [11,19,24]. A part of the R2 endonuclease from *Drosophila melanogaster* has been modeled using the crystal structure of Fok I restriction endonuclease [24]. Structural database has shown Holiday junction resolvases from archaea as a top match to model the R2 endonuclease. Very recently, R2 endonuclease from *Bombyx mori* has been extensively modeled using the crystal structures of archaeal Holliday junction resolvases E and C (Hje and Hjc) and have shown to share the αβββαβ core fold along with the similar placement of PD-(D/E) catalytic residues [19]. The arginine rich R-box region with consensus sequence RxxR or NXRXXR is identified at the beginning of the first α-helix in restriction endonuclease. In restriction enzymes like Mbol, SsoII, and PspGI, mutating two arginine to alanine in this region has shown to catalytically inactivate the enzyme and drastically reduce the ability to bind to DNA [21,22,25]. In the crystal structure of Holiday junction resolvase C from (Hjc) from *Pyrococcus furiosus,* the R-box region consists of arginine and lysine near the start of the α-helix. Mutating the RK residues to alanine reduced binding of the enzyme to the holiday junction, but it did not affect the ability to cleave the DNA [26]. Deleting MYRKG residues in the R-box region of *Pyrococcus furiosus* Hjc completely knocked out the DNA binding ability of the enzyme [26]. The RK residues are conserved in most of the members of Hjc resolvase. A Glycine residue is also found to be conserved in the R-box region of Hjc and Hje Holliday junction which is thought to aid conformational change of the N-terminal segment of the enzyme once bound to a Holiday junction [26]. The Glycine residue is also found in some restriction enzymes with R-box.

In R2Bm, the R-box region is located immediately after the presumptive α-finger and the gag-like zinc knuckle (CCHC motif) of the liker region (Figure 4-1C). The R2Bm R-box consists of "RILRH" motif. Mutating the conserved RH residues that is located at the start of the first α-helix of R2 RLE decreases the overall DNA binding ability of R2 protein [19]. The RH double mutant also showed about 40% decrease in binding to a non-specific DNA in the presence of non-specific RNA, indicating they are involved in non-sequence specific Protein-DNA contacts [19]. In this study, mutating only the arginine residue as a single mutant (RILR/AH) did not show any decrease in target DNA binding activity (Figure 4-2A, B, and C). In addition, the previously unexplored first arginine residue in the RILRH motif was mutated (R/AILRH) in this study and subjected to DNA binding assay. R/AILRH mutant protein also did not show any reduction in DNA binding activity of R2 protein (Figure 4-2A, B, and C). The binding assay was conducted using a target DNA pre-nicked on the first strand because DNase I footprint has shown that after first strand gets cleaved the protein footprint extends on target DNA [23].

Given that the RH double mutation has previously shown to affect DNA binding, the lack of similar phenotype when single R was mutated might be because mutating only a single arginine residue is not strong enough to show any drastic change. For future extension of this work, a double point mutant having both the arginine residues mutated will be generated and tested to better characterize their DNA binding role. The ZF and myb motifs is recognized as downstream DNA binding domain of R2 protein and we speculate similar distinct upstream DNA binding domain [27]. The non-specific DNA interacting ability of the R-box region of RLE could work together to cleave the non-palindromic sequences of the DNA target.

While R/AILRH and RILR/AH mutations did not show the expected effect on DNA binding, the mutations significantly affected the ability of R2 protein to cleave the second strand (Figure 4-4B, C). The first strand cleavage and first strand synthesis were unaffected by the mutations (Figure 4-3A, B). This indicated that the mutations interfered with the catalytic activity of endonuclease to specifically make cleavage on the second strand possibly due to protein conformational change. In the 3D model of R2Bm endonuclease, the second arginine of R-box is located at the edge of the active site cleft and sits on the surface of the protein [19]. The R-box arginine residues could also be involved in coordinating the target DNA to correctly position it to the active site of the endonuclease required for the second strand cleavage. The cleavage assay was done using a linear DNA in this study. It has been recently identified that a 4-way junction is the correct intermediate target DNA structure essential for second strand cleavage and second strand synthesis (Chapter 2). Cleavage assays for these mutants should also be conducted using a 4-way junction to make final conclusions about their second strand cleavage activity. In addition, the single and double mutants in R-box should also be assayed for 4-way junction binding activity to see if they play a role in recognizing the 4-way intermediate DNA structure which is key to the second half of the integration mechanism.

4.5 Methods

The methods for protein and nucleic acid preparation and R2Bm reactions and analysis were carried out as described in chapter 3.

4.6 References

1. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
2. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol. Biol. Evol.* **22,** 2157–2165 (2005).
3. Burke, W. D., Malik, H. S., Jones, J. P. & Eickbush, T. H. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* **16,** 502–11 (1999).
4. Jakubczak, J. L., Burke, W. D. & Eickbush, T. H. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 3295–3299 (1991).
5. Kojima, K. K., Kuma, K. I., Toh, H. & Fujiwara, H. Identification of rDNA-specific non-LTR retrotransposons in cnidaria. *Mol. Biol. Evol.* **23,** 1984–1993 (2006).
6. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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).
7. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21,** 5899–5910 (2002).
8. Christensen, S. M. & Eickbush, T. H. 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).
9. Moran, J., Holmes, S. & Naas, T. High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
10. Burke, W. D., Calalang, C. C. & Eickbush, T. H. 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.* (1987). doi:10.1128/MCB.7.6.2221.Updated
11. Yang, J., Malik, H. S. & Eickbush, T. H. 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,** 7847–52 (1999).
12. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).
13. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res.* **42,** 8405–8415 (2014).
14. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: Similarities to eukaryotic splicing factor Prp8. *Mob. DNA* **8,** 1–15 (2017).
15. Christensen, S. M., Ye, J. & Eickbush, T. H. 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).
16. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. & Ginalski, K. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks382
17. Kosinski, J., Feder, M. & Bujnicki, J. M. 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* (2005). doi:10.1186/1471-2105-6-172
18. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1242
19. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: Loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res.* **44,** 3276–3287 (2016).
20. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. Type II restriction endonucleases: Structure and mechanism. *Cellular and Molecular Life Sciences* **62,** 685–707 (2005).
21. Pingoud, V. *et al.* 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.* (2005). doi:10.1074/jbc.M409020200
22. Pingoud, V. *et al.* 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.* (2003). doi:10.1016/S0022-2836(03)00523-0

23. Christensen, S. & Eickbush, T. H. Footprint of the Retrotransposon R2Bm Protein on its Target Site before and after Cleavage. *J. Mol. Biol.* **336,** 1035–1045 (2004).

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

25. Pingoud, V. *et al.* Evolutionary relationship between different subgroups of restriction endonucleases. *J. Biol. Chem.* **277,** 14306–14 (2002).

26. Nishino, T., Komori, K., Ishino, Y. & Morikawa, K. Dissection of the Regional Roles of the Archaeal Holliday Junction Resolvase Hjc by Structural and Mutational Analyses. *J. Biol. Chem.* (2001). doi:10.1074/jbc.M104460200

27. Shivram, H., Cawley, D. & Christensen, S. M. 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 (2011).

Chapter 5

Identification of 3’ RNA Binding Motif in the RT-1 Region of Protein Encoded by R2 LINE.

Monika Pradhan and Shawn M. Christensen

5.1 Abstract

Long Interspersed Elements (LINEs) prevalent in eukaryotic genome integrate by a process called Target Primed Reverse Transcription (TPRT). R2 is a site-specific restriction like endonuclease (RLE) bearing LINE that inserts into a specific target site of 28S rRNA gene and is used as a model system to study integration mechanism by TPRT. The R2 protein recognizes the distinct RNA structures within the 5' and 3' untranslated regions of its own RNA, also known as Protein binding motifs (PBMs). Two R2 protein subunits are believed to be required for integration, one bound to the 3' PBM and one bound to the 5' PBM of the element RNA. RNA binding determines final protein conformation and dictates the role of each protein subunit in the integration reaction. Two motifs located in the N-terminal RT-1 and RT0 regions of the R2 protein are identified to bind both 3' and 5' PBM RNAs. In this study, mutation generated in the RT-1 region results in R2 protein that retains the ability to bind to 5' PBM RNA but reduces the ability to bind to 3' PBM RNA. The RT-1 mutation also affects the ability of R2 protein to conduct TPRT reaction and to cleave the top strand target DNA. There appears to be additional RNA binding residues within R2 protein open reading frame specifically recognizing and binding 3' PBM RNA and 5' PBM RNA with some extent of overlap.

5.2 Introduction

Long INterspersed Elements (LINEs) or non-Long Terminal Repeat retrotransposons (non-LTRs) are an important class of retrotransposons that have persisted in the eukaryotic genome for hundreds of millions of years [1,2]. LINEs integrate into the host DNA by a process called Target Primed Reverse Transcription (TPRT) in which the element encoded protein creates a nick on the target site exposing 3'-OH group that in turn primes reverse transcription of their RNA template [3,4]. R2 is a clade of site-specific LINE that inserts solely into the 28S rRNA genes of its host [5–7]. R2 encodes a single ORF with N-terminal DNA binding zinc finger (ZF) and a myb motif and RNA binding domain; centrally located is a reverse transcriptase (RT) domain; and a C-terminal area containing linker region with presumptive alpha helix (HINALP motif) and gag-knuckle like zinc finger (CCHC motif), and a type II restriction like endonuclease (RLE) domain (Figure 5-1A) [6,8–13]. The RT, linker, and RLE of R2 encoded protein forms a larger globular domain while the N-terminal region forms a smaller globular domain [13].

Current model of R2 retrotransposition (Figure 5-1B) requires two ribonucleoprotein (RNP) complexes which are formed due to the strong *cis* preference of R2 encoded protein to R2 RNA (Figure 5-1B) [12]. When R2 protein is bound to 3’ protein binding motif (PBM) of R2 RNA, it adopts a conformation that allows it to bind to the upstream sequences of 28S rDNA (28Su) relative to the insertion site and when R2 protein is bound to 5’ PBM, it attains another conformation that binds to the downstream sequences (28Sd) of the insertion site (Figure 5-1B) [14]. The upstream and downstream protein subunits catalyzes the integration mechanism [14]. Endonuclease from the upstream subunit nicks the first (antisense) strand exposing 3’-OH on the target site which then acts as a primer for cDNA synthesis by reverse transcriptase of the same subunit (Figure 5-1B, step 1 and 2) [3]. The reverse transcriptase removes the 5’ PBM from downstream subunit during cDNA synthesis and then jumps template form 5’ PBM end of RNA to 28S upstream sequences (Figure 5-1B step 3) [12] (Chapter 2). The endonuclease from downstream subunit cleaves the second (sense) strand(Figure 5-1B step 4) [12]. Finally, reverse transcriptase from downstream subunit catalyzes the synthesis of second strand thus completing element integration (Figure 5-1B step 5) [14].

Binding of R2 protein to 3’ and 5’ PBM RNA forms two RNP complexes that are crucial for successful element integration. 3’ PBM corresponds to the 250 nt of the 3’ untranslated region (UTR) of R2 RNA while 5’ PBM covers a 300 nt segment that starts within the 5’ UTR and ends just before the N-terminal ZF. The 3’ and 5’ PBM regions of R2 transcript are folded into distinct structures and have shown to be involved in binding R2 protein [15–17]. In addition to structural difference between the two PBMs, they are also functionally different. While contacts of R2 protein with 3’ PBM induces the protein to bind to upstream DNA sequences and lets the RT utilize R2 RNA for TPRT, binding to the 5’ PBM changes the specificity of R2 protein to bind to the downstream sequences from the insertion site [12,18]. R2 protein has high specificity to bind to RNA and the only RNA utilized in TPRT reaction are those that contain R2 3’ PBM [19]. Two motifs, one in RT-1 region and another in RT0 region N-terminal to RT domain have been identified to bind both 3’ and 5’ PBMs of R2 RNA [11] (Figure 5-1C residues with blue open circles on top). These regions are called RT-1 and RT0 regions since they lie upstream to the conserved 1-7 domains of reverse transcriptase. It is still unclear how one region of the protein could bind both PBMs, yet 3′ PBM is positioned to be used as a template for TPRT, while the 5′ PBM changes the specificity of protein for binding to the downstream DNA sequences. Given the different secondary structures adopted by 3’ and 5’ PBMs of R2 transcript and the different conformational changes they induce when bound to R2 protein, we hypothesize that there are additional unidentified residues that distinguish and specifically binds to either 3’ or 5’ PBM, while there are residues that bind both. This chapter reports a Tyrosine (Y) residue in the RT-1 region that is solely involved in binding 3’ PBM of R2 RNA. Mutating the Y residue did not affect the ability of R2 protein to bind to 5’ PBM and cleave the first strand, but it did affect the ability to bind to 3’ PBM RNA, use the cleaved target DNA strand to synthesize first strand, and cleave the second target DNA strand.

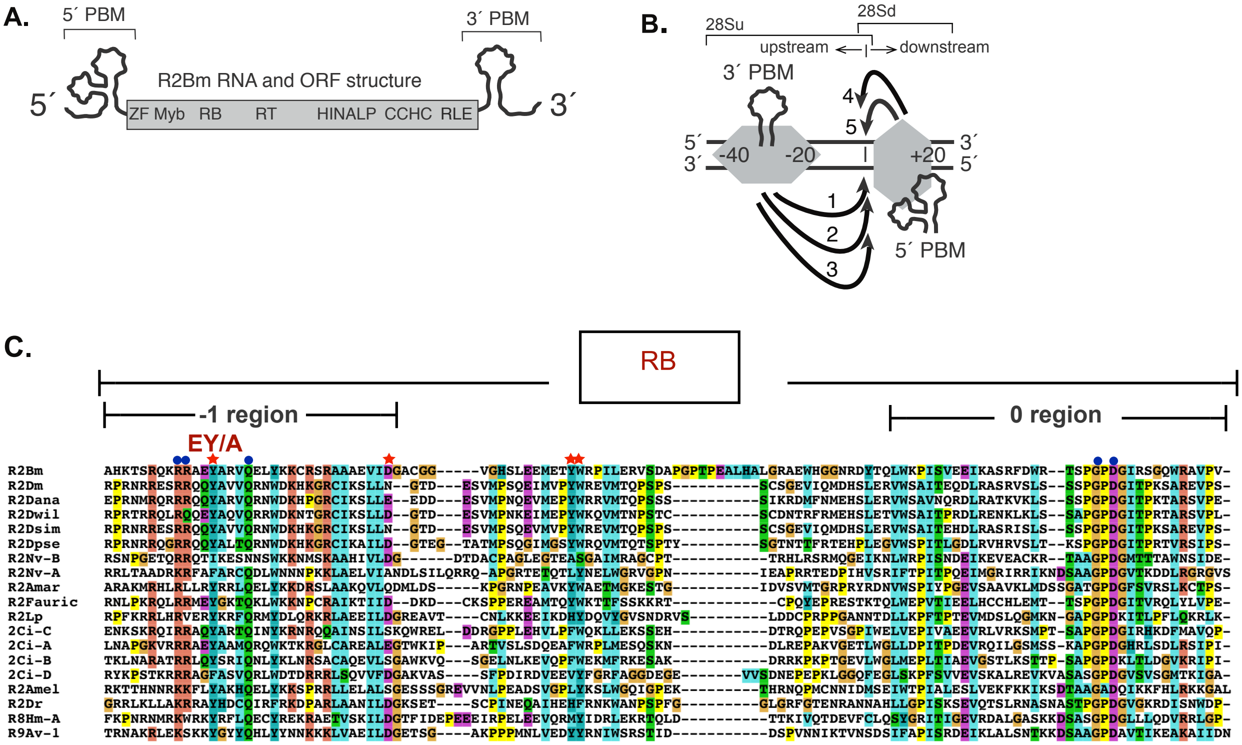

A.
5´ PBM
3´ PBM
5´
3´
R2Bm RNA and ORF structure
ZF Myb RB RT HINALP CCHC RLE
B.
28Su
28Sd
upstream
downstream
3´ PBM
4
5
5´
3´
-40
-20
+20
3´
5´
1
2
3
5´ PBM
C.
RB
-1 region
0 region
EY/A
R2Bm AHKTSRQKRRAEYARVQELYKKCRSRAAAEVIDGACGG-----VGHSLEEMETYWRPILERVSDAPGPTPEALHALGRAEWHGGNRDYTQLWKPISVEEIKASRFDWR--TSPGPDGIRSGQWRAVPV-
R2Dm RPRNRRESRRQQYAVVQRNWDKHKGRCIKSLLN--GTD---ESVMPSQEIMVPYWREVMTQPSPS----------SCSGEVIQMDHSLERVWSAITEQDLRASRVSLS--SSPGPDGITPKSAREVPS-
R2Dana EPRNMRQRRRQQYARVQRNWDKHPGRCIKSLLE--EDD---ESVMPNQEVMEPYWRRVMTQPSSS----------SIKRDMFNMEHSLERVWSAVNQRDLRATKVKLS--SSPGPDGITPKTARSVPE-
R2Dwil RPRTRRQLRQQEYAQVQRRWDKNTGRCIKSLLD--GTD---ESVMPNKEIMEPYWKQVMTNPSTC----------SCDNTRFRMEHSLETVWSAITPRDLRENKLKLS--SAPGPDGITPRTARSVPL-
R2Dsim RPRNRRESRRQQYAVVQRNWDKHKGRCIKSLLN--GTD---ESVMPSQEVMVPYWREVMTQPSPS----------SCSGEVIQMDHSLERVWSAITEHDLRASRISLS--SSPGPDGITPKSAREVPS-
R2Dpse RPRNRRQGRRQQYALTQRNWDKHKGRCIKAILD--GTEG--TATMPSQGIMGSYWRQVMTQTSPTY---------SGTNTTFRTEHPLEGVWSPITLGDLRVHRVSLT--KSPGPDGITPRTVRSIPS-
R2Nv-B RSNPGETQRRQTIKESNNSWKKNMSKAAHIVLDG------DTDACPAGLEGTEASGAIMRAGCPT-----------TRHLRSRMQGEIKNLWRPISNDEIKEVEACKR--TAAGPDGMTTTAWNSIDE-
R2Nv-A RRLTAADRKRFAFARCQDLWNNNPKKLAELVIANDLSILQRRQ-APGRTETQTLYNELWGRVGPN----------IEAPRRTEDPIHVSRIFTPITPQEIMGRIRRIKNDSAAGPDGVTKDDLRGRGVS
R2Amar ARAKMRHLRLLRYRRLQELYKKDRSLAAKQVLQDMLDS-----KPGRNPEAVKYWAETMGKESTG----------IDVSVMTGRPRYRDNVWSPIYPGEVSAAVKLMDSSGATGPDGFSVRSLKCTPS-
R2Fauric RNLPKRQLRRMEYGKTQKLWKKNPCRAIKTIID--DKD---CKSPPEREAMTQYWKTTFSSKKRT-----------CPQYEPRESTKTQLWEPVTIEELHCCHLEMT--TSPGPDGITVRQLYLVPE-
R2Lp RPFKRRLHRVERYKRFQRMYDLQRKRLAEEILDGREAVT----CNLKKEEIKDHYDQVYGVSNDRVS---------LDDCPRPPGANNTDLLKPFTPTEVMDSLQGMKN-GAPGPDKITLPFLQKRLK-
2Ci-C ENKSKRQIRRAQYARTQINYKRNRQQAINSILSKQWREL--DDRGPPLEHVLPFWQKLLEKSSEH-----------DTRQPEPVSGPIWELVEPIVAEEVRLVRKSMPT-SAPGPDGIRHKDFMAVQP-
2Ci-A LNAPGKVRRRAEYAAMQRQWKTKRGLCAREALEGTWKIP---ARTVSLSDQEAFWRPLMESQSKN-----------DLREPAKVGETLWGLLDPITPDEVRQILGSMSS-KAPGPDGHRLSDLRSIPI-
2Ci-B TKLNARATRRLQYSRIQNLYKLNRSACAQEVLSGAWKVQ---SGELNLKEVQPFWEKMFRKESAK-----------DRRKPKPTGEVLWGLMEPLTIAEVGSTLKSTTP-SAPGPDKLTLDGVKRIPI-
2Ci-D RYKPSTKRRRAGFASVQRLWDTDRRRLSQVVFDGAKVAS----SFPDIRDVEEVYFGRFAGGDEGE-------VVSDNEPEPKLGGQFEGLSKPFSVVEVSKALREVKVASASGPDGVSVSGMTKIGA-
R2Amel RKTTHNNRKKFLYAKHQELYKKSPRRLLELALSGESSSGREVVNLPEADSVGPLYKSLWGQIGPEK---------THRNQPMCNNIDMSEIWTPIALESLVEKFKKIKSDTAAGADQIKKFHLRKKGAL
R2Dr GRRLKLLAKRRAYHDCQIRFRKDPARLAANILDGKSET----SCPINEQAIHEHFRNKWANPSPFG---------GLGRFGTENRANNAHLLGPISKSEVQTSLRNASNASTPGPDGVGKRDISNWDP-
R8Hm-A FKPNRNMRKWRKYRFLQECYREKRAETVSKILDGTFIDEPEEEIRPELEEVQRMYIDRLEKRTQLD-----------TTKIVQTDEVFCLQSYGRITIGEVRDALGASKKDSASGPDGLLLQDVRRLGP-
R9Av-1 TRNAKRLEKSKKYGYYQHLYYNNKKKLVAEILDGETSG----AKPPPMNLVEDYYRNIWSRSTID-----------DSPVNNIKTVNSDSIFAPISRDEIKLALSNTKKDSAAGPDAVTIKEAKAIIDN

Figure 5-1: R2Bm structure, integration model and multiple sequence alignment. **A.** R2Bm RNA structure. Box in the middle represents open reading frame (ORF) with conserved motifs. Abbreviations: zinc finger (ZF), Myb (Myb), RNA binding (RB), reverse transcriptase domain (RT), a conserved predicted α-helices with HINALP residues (HINALP), a gag-like zinc knuckle with Cysteine/Histidine rich motif (CCHC) and type II restriction-like endonuclease (RLE). R2Bm RNA segments corresponding to the 5’ and 3’ untranslated regions known to adopt distinct structures and bind the R2Bm protein are labelled within the brackets as 5’ and 3’ protein binding motifs (PBMs), respectively. 5’ PBM and 3’ PBM are used to generate data in the paper. **B.** R2 target site, 28S rDNA, and insertion model. R2 protein associated with the 3’ PBM RNA binds 20 to 40 bases upstream (28Su) of the insertion site (vertical line) and protein associated with the 5’ PBM RNA binds to 20 bases downstream of the insertion site [10,20]. Insertion occurs in five steps. **C.** Clustal alignment of the N-terminal region spanning from RT-1 to RT0 region from diverse arthropods (12 sequences), a vertebrate Danio rerio (R2Dr), R8 from Hydra magnipapillata (R2Hm-A) and R9 from Adineta vaga (R9Av-1). Blue closed circle represents the motifs already identified to be involved in binding 3’ and 5’ PBM RNAs [11] and red star represents the residues mutated in this study. The point mutations generated in this study are: GAEE/A, EY/A, VID/A, YW/A, and YWR/A. Abbreviations: R2Bm = Bombyx mori, R2Dm = Drosophila melanogaster, R2Dana = Drosophila ananassae, R2Dwil = Drosophila willistoni, R2Dsim = Drosophila simulans, R2Dpse = Drosophila pseudoobscura, R2Fauric = Forficula auricularia, R2Amar = Anurida maritima, R2Nv-B = Nasonia vitripennis, R2Lp = Limulus polyphemus, R2Ci = Ciona intestinalis, R2Amel = Apis mellifera, R2Dr = Danio rerio, R8Hm-A = Hydra magnipapillata, R9Av-1 = Adineta vaga

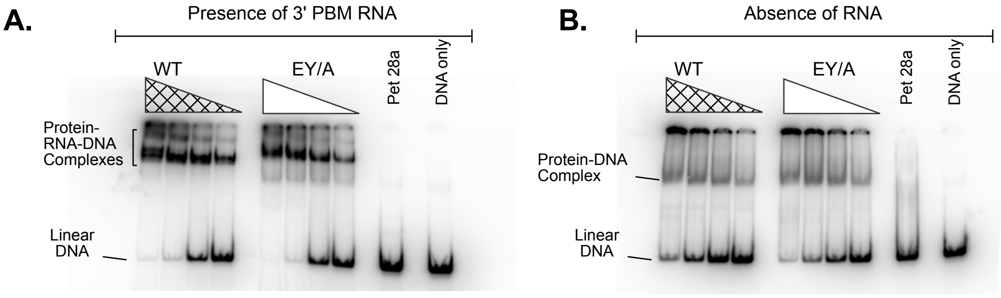
5.3 Results

5.3.1 Single point mutants were generated in the N-terminal region.

To determine the residues involved in independently binding 3’ or 5’ PBM of R2 RNA, additional residues in the N-terminal region were mutated (Figure 5-1C). The mutations generated were GAEE/A, EY/A, VID/A, YW/A, and YWR/A. The EY/A and VID/A mutations were located within the RT-1 region, GAEE/A mutation was located upstream of the RT-1 region, and YW/A and YWR/A mutations were located between the RT-1 and RT0 regions. The VID/A and YW/A mutations did not yield enough soluble protein as compared to wild type (WT) protein to carry out functional assays, so were dropped out from the study. The GAEE/A and YWR/A mutations showed wild-type like phenotype, so they will not be further discussed.

5.3.2 EY/A mutation reduces binding to 3’ PBM RNA but does not affect first strand cleavage

To investigate if the RT-1 tyrosine (Y) residue is involved in binding 3’ PBM RNA, the EY/A mutant protein was tested against WT protein for its ability to form protein-RNA-DNA complexes. In the presence of 3’ PBM RNA, the WT protein formed typical protein-RNA-DNA complexes, whereas the EY/A mutant protein was less efficient in forming 3’ PBM bound complexes on EMSA gel on linear DNA (Figure 5-2A). Unlike WT protein, the EY/A mutant protein formed an additional smeary band below the typical complex that corresponds to protein-DNA only (no RNA) complex. The analogous protein-DNA complex could only be observed in EMSA gel in the absence of RNA for both WT and EY/A mutant (Figure 5-2B). A bar graph was plotted in which the WT protein activity to form protein-RNA-DNA complex (*f*WT activity) was set to one and the relative activity of EY/A mutant was quantitated. The EY/A mutant reduced the ability of the R2 protein to form 3’ PBM RNA bound complex by ~38% as compared to WT protein.



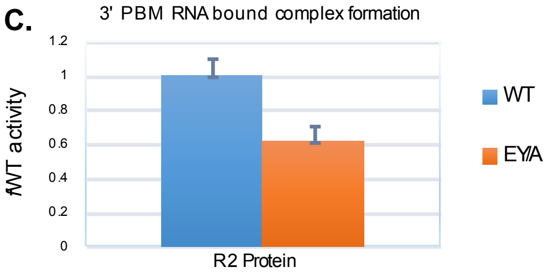


Figure 5-2: EY/A mutation affects the 3' PBM RNA binding activity of R2 protein. **A.** EMSA gel showing the ability of R2 protein to bind to linear target DNA and 3' PBM RNA. The triangles above the EMSAs represent a protein titration series. **B.** EMSA gel showing the ability of R2 protein to bind to the target DNA in the absence of RNA. **C.** Bar graph of WT vs EY/A mutant protein's ability to form 3' PBM RNA bound complexes in EMSA gel shown in A.

The first strand cleavage activity of mutant protein was also tested against WT protein and presented in a scatterplot (Figure 5-3B). The fraction of the target DNA cleaved at the first strand (*f*cleaved) was quantitated from denaturing gel (Figure 5-3A) and the fraction of target DNA bound by the R2 protein (*f*bound) was calculated from EMSA gel (Figure 5-2A). The EY/A mutant protein was found to be as efficient as WT protein to conduct first strand cleavage. No additional cleavages beyond the R2 cleavage site was observed for WT and mutant proteins.

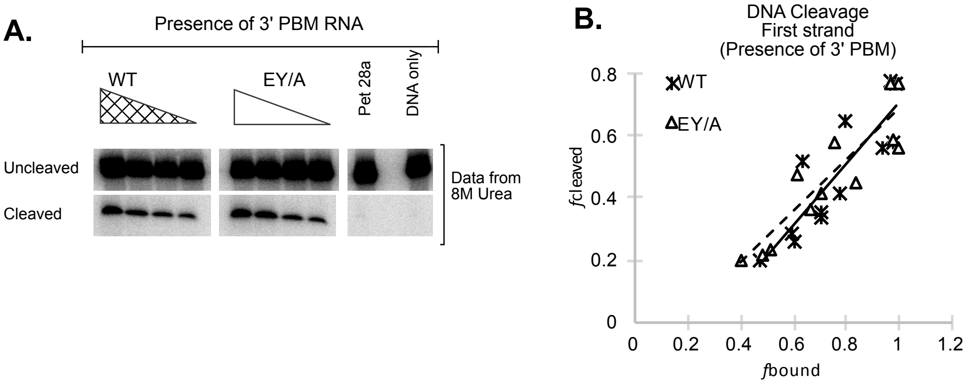


Figure 5-3: First strand cleavage activity of EY/A mutant protein. **A.** Denaturing gel of the same reactions in Figure 5-2A. showing the signals for uncleaved and first strand cleaved target DNA. Fraction of target DNA that undergoes first strand cleavage (fcleaved) is quantitated from denaturing gel. **B.** Scatter plot of fraction of cleaved target DNA (fcleaved) plotted as a function of fraction of target DNA bound by protein (fbound) at each protein concentrations. Data points for WT, and EY/A are represented by asterisk and open triangle, respectively.

5.3.3 EY/A mutation reduces the ability of R2 protein to conduct first strand synthesis

After first strand cleavage, R2 protein utilizes R2 RNA as a template to synthesize first strand. In order to investigate the ability of the EY/A mutant protein to conduct TPRT reaction, R2 protein was incubated with 3' PBM

RNA and linear target DNA in the presence of dNTPs. The target DNA was labeled on the bottom strand to track the first strand cleavage and TPRT product formation on a denaturing gel (Figure 5-4A). The efficacy of WT vs EY/A mutant protein to synthesize first strand by TPRT are presented as a scatter plot (Figure 5-4B). While the WT protein generated robust TPRT products at each protein titrations in denaturing gel, the EY/A mutant seems to be comparatively less efficient (Figure 5-4A). Overall, the EY/A mutation decreased the ability of R2 protein to utilize the cleaved target DNA for TPRT product formation.

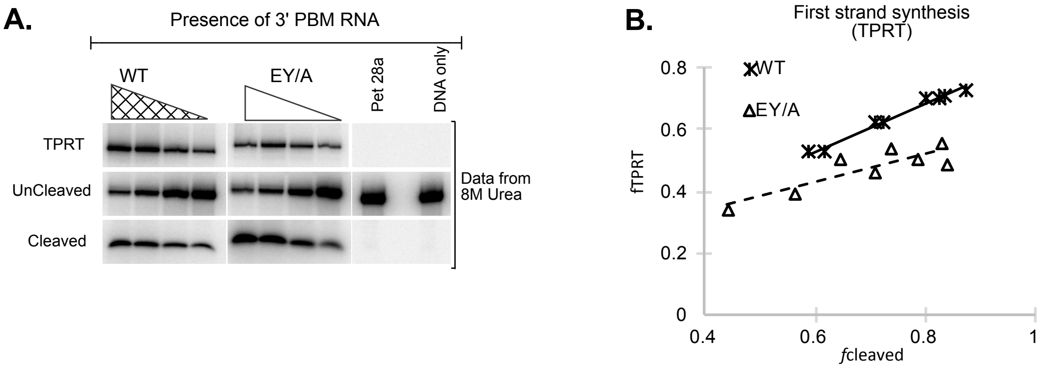


Figure 5-4: First strand synthesis by TPRT activity of EY/A mutant. **A**. Denaturing gel showing uncleaved, first strand cleaved and TPRT product formation. Fraction of target DNA that undergoes first strand cleavage (fcleaved) and fraction of cleaved target DNA that undergoes TPRT (fTPRT) were calculated from the denaturing gel. **B.** Scatter plot of fraction of the target DNA that undergoes first strand synthesis by TPRT (fsynthesis) as a function of fraction DNA cleaved by the R2 protein (fbound) at each protein concentrations. Symbols and abbreviations are as in previous figures.

5.3.4 EY/A mutation does not reduce binding to 5’ PBM RNA but affects second strand cleavage on linear target DNA

In the presence of 5’ PBM RNA, the R2 protein binds to the downstream sequences from the insertion site and cleaves the second strand. The ability of EY/A vs WT protein to bind to the 5’ PBM RNA was assayed by the formation of a typical Protein-RNA-DNA complex in the EMSA gel (Figure 5-5A) and the ability to cleave the second strand was assayed using denaturing gel (Figure 5-5B). Linear target DNA labeled on the second strand was used in the assay. The EY/A mutant protein was found be as efficient as WT protein in binding 5’ PBM RNA as they formed the typical protein-RNA-DNA complex in EMSA gel (Figure 5-5A). The WT protein forms an additional band below this typical band that corresponds to RNP bound to the double strand cleaved DNA. A graph of the fraction of target DNA that underwent second strand cleavage (*f*cleavage) as a function of fraction of target DNA bound by R2 protein (*f*bound) is reported in Figure 5-5C. EY/A mutation affects the ability of R2 protein to cleave the second strand of linear DNA as compared to WT.

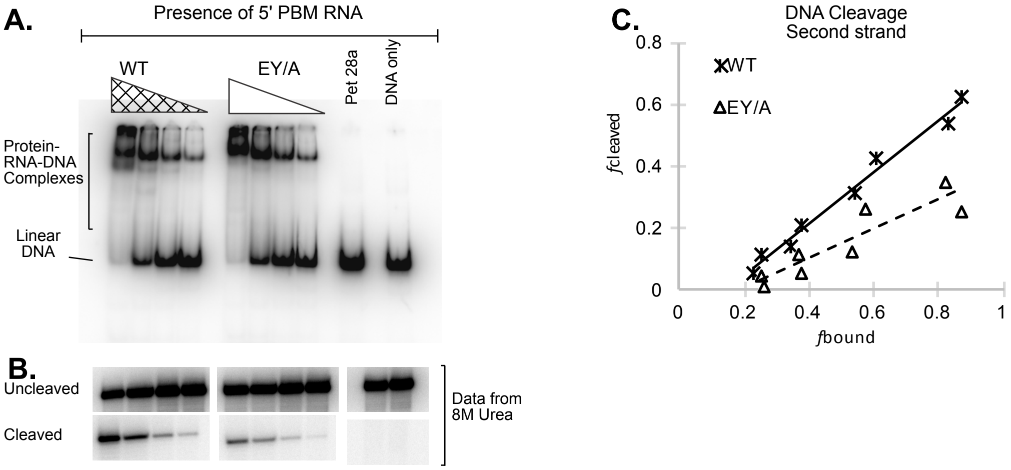


Figure 5-5: Second strand cleavage activity of EY/A mutant protein. A. EMSA gel showing the Protein-RNA-DNA complexes formed in the presence of 5' PBM RNA and linear target DNA. B. Denaturing gel of the same reactions in A. showing the signals for uncleaved and second strand cleaved target DNA. Fraction of target DNA that undergoes second strand cleavage (fcleaved) is quantitated from denaturing gel. C. Scatter plot of fraction of cleaved target DNA (fcleaved) plotted as a function of fraction of target DNA bound by protein (fbound) at each protein concentrations. Symbols and abbreviations are as in previous figures.

5.4 Discussion

Mutational analysis has revealed that the N-terminal RT-1 and RT0 regions forms a part of RNA binding domain of R2 protein [11]. Motifs identified in the -1 and 0 regions bind both 3' and 5' PBM RNAs [11]. However, it is not made clear how the two regions of RNA could be bound to a single area of the protein, yet 3' PBM RNA could induce R2 protein to adopt upstream DNA binding protein conformation and position itself to initiate first strand synthesis, whereas 5' PBM RNA could prompt the R2 protein to adopt downstream DNA binding conformation. The 5' PBM region of R2 RNA adopts a distinctive pseudoknot structure that is found to be conserved across silk moths [17,21]. The 3' PBM of R2 transcript forms a unique secondary structure that differs from that of 5' PBM RNA [16]. It remains ambiguous how two structurally different regions of RNA could be recognized and be bound by the same region of R2 protein. With these observation, we predict that while there are some common 3' and 5' PBM RNA binding residues in R2 protein, there should also exist additional residues that discretely binds either 3' or 5' PBM RNA. In order to identify the distinct set of RNA binding motifs, a high resolution alanine scanning of the N-terminal region was done and mutational analysis within the RT-1 region has identified a Tyrosine (Y) motif exclusively involved in 3’ PBM RNA binding. Mutating the residue also affected the protein’s ability to conduct TPRT and to cleave the second strand in gel shift assays.

In the presence of 3’ PBM RNA, WT forms a typical 3’ PBM RNA bound complex but the EY/A mutant showed about 38% reduction in forming the RNA bound complex (Figure 5-2C). In addition, EY/A mutant protein formed a smeary band below this typical complex that represented to a Protein-DNA complex or 3’ PBM RNA free complex (Figure 5-2A). This clearly indicated that the Y residue in the -1 region plays an important role in binding 3’ PBM RNA. Any such RNA free complex was not observed in EMSA gels in the presence of 5’ PBM RNA. The EY/A mutant protein was efficient in binding 5’ PBM RNA and formed distinct protein-RNA-DNA complex like that of WT (Figure 5-5A) implying that Y residue does not take any part in binding 5’ PBM RNA. This shows that as speculated, we have identified at least one residue in the RT-1 region that identifies as a motif that discretely binds 3’ PBM RNA.

The EY/A mutation did not affect the ability of the protein to cleave the first DNA strand at the target site (Figure 5-3B). However, the mutation reduced the efficiency of R2 protein to form TPRT products (Figure 5-4B). Since the mutant protein could readily cleave the first strand target DNA, the observed reduction in first strand synthesis should be because of the reduced ability of the mutant protein to bind to 3’ PBM RNA which is utilized as a template for first strand synthesis. In addition, that the mutant protein might also have lost the ability to properly position the 3’ PBM region of R2 RNA to allow priming or once primed, the protein may not be able to perform reverse transcription. However, it is more apparent that the loss of 3’ PBM RNA binding has led to the reduction in TPRT product formation for the EY/A mutant.

The EY/A mutant reduced the second strand cleavage activity of EY/A mutant when compared to WT on a linear target DNA (Figure 5-5C). The inefficiency of EY/A mutant protein to cleave the second strand explains the lack of double strand broken complex for the mutant in EMSA gel (Figure 5-5A). This indicates that EY/A and WT could bind the RNA with similar efficacy but the mutation hinders second strand cleavage. This could occur when the protein bound 5’ PBM RNA could not be removed from the downstream subunit only after which second strand cleavage occurs. In other words, the mutation could possibly affect proper endonuclease conformation required for second strand cleavage to occur. The cleavage assay, however was done using linear target DNA. Although linear DNA was used for cleavage assays historically, the target DNA intermediate identified to be essential for second strand cleavage is a 4-way junction (Chapter 2). So, further investigation using a four-way junction is essential to make conclusions about second strand cleavage activity of the mutant protein.

While we have identified a single RNA binding residue that specifically recognizes and binds 3' PBM RNA, extensive characterization of RNA binding by R2 protein still needs to be done. In LINE-1 elements that are APE bearing LINEs, RNA binding domain was found to be located in ORF1 which is separate from ORF2 that encodes for endonuclease and reverse transcriptase [22]. Mobile group II introns and Telomerase which are phylogenetically related to LINEs, have their RNA binding residues limited not only to the regions N-terminal to reverse transcriptase but also involve residues in finger, palm, and thumb regions of reverse transcriptase [23,24]. *In vivo* studies in human LINE-1 elements have shown reduction of ORF2 protein in RNP complex when the first two cysteines of the zinc knuckle (CCHC motif) were mutated, implying its possible role in RNA binding [25]. In human LINE-1 elements, a recombinant protein containing 180 amino acid residues of the C-terminal region of ORF2 have shown to be involved in RNA binding [26]. It is evident that in elements similar to R2, RNA binding surface is not limited to few N-terminal residues but they are spread over the central RT domain and the C-terminal domain. Investigation on additional domains of R2 protein needs to be done to identify the spread of these RNA interacting motifs. While a rigorous alanine scanning of R2 protein could yield information on discrete 3' and 5' PBM RNA binding motifs, mutational studies on the entire R2 protein will be too time consuming. So, targeted site directed mutagenesis study should be coupled with other global approaches like mass spectrometric protein footprinting assay for determination of amino acid residues of R2 protein that makes RNA contacts [27]. Elucidating a complete picture of RNA binding motifs of R2 protein would be a large step forward to understanding the insertion mechanism of this important class of retrotransposable element.

5.5 Methods

The methods for protein and nucleic acid preparation and R2Bm reactions and analysis were carried out as described in chapter 3.

5.6 References

1. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
2. Eickbush, T. H. & Malik, H. S. in *Mobile DNA II Edited by: Craig NL, Craigie R, Gellert M, Lambowitz*

AM. Washington, DC: ASM Press **93,** 1111–1144 (2002).

3. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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).
4. Eickbush, D. G., Luan, D. D. & Eickbush, T. H. Integration of Bombyx mori R2 sequences into the 28S ribosomal RNA genes of Drosophila melanogaster. *Mol. Cell. Biol.* **20,** 213–23 (2000).
5. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol. Biol. Evol.* **22,** 2157–2165 (2005).
6. Jakubczak, J. L., Xiong, Y. & Eickbush, T. H. 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.* (1990). doi:10.1016/0022-2836(90)90303-4
7. Kojima, K. K., Kuma, K. I., Toh, H. & Fujiwara, H. Identification of rDNA-specific non-LTR retrotransposons in cnidaria. *Mol. Biol. Evol.* **23,** 1984–1993 (2006).
8. Burke, W. D., Calalang, C. C. & Eickbush, T. H. 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.* (1987). doi:10.1128/MCB.7.6.2221.Updated
9. Yang, J., Malik, H. S. & Eickbush, T. H. 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,** 7847–52 (1999).
10. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).
11. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res.* **42,** 8405–8415 (2014).
12. Christensen, S. M. & Eickbush, T. H. 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. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: Similarities to eukaryotic splicing factor Prp8. *Mob. DNA* **8,** 1–15 (2017).
14. Christensen, S. M., Ye, J. & Eickbush, T. H. 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. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3,** 1–16 (1997).
16. Ruschak, A. M. *et al.* Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* **10,** 978–987 (2004).
17. Kierzek, E. *et al.* 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 (2009).
18. Yang, J. & Eickbush, T. H. RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol. Cell. Biol.* **18,** 3455–65 (1998).
19. Luan, D. D. & Eickbush, T. H. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell. Biol.* **15,** 3882–91 (1995).
20. Christensen, S. & Eickbush, T. H. Footprint of the Retrotransposon R2Bm Protein on its Target Site before and after Cleavage. *J. Mol. Biol.* **336,** 1035–1045 (2004).
21. Kierzek, E. *et al.* 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.* (2008). doi:10.1093/nar/gkm1085
22. Martin, S. L. The ORF1 protein encoded by LINE-1: Structure and function during L1 retrotransposition. *J. Biomed. Biotechnol.* **2006,** 1–6 (2006).
23. Gu, S.-Q. *et al.* Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16,** 732–747 (2010).
24. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat. Struct. Mol. Biol.* **17,** 513–518 (2010).

25. Doucet, A. J. *et al.* Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet.* **6,** 1–19 (2010).

26. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3,** 433–437 (2013).

27. Kvaratskhelia, M. & Grice, S. F. J. Le. Structural analysis of protein-RNA interactions with mass spectrometry. *Methods Mol. Biol.* **488,** 213–9 (2008).

Chapter 6

Conclusions

6.1 Second half of the integration mechanism

The RLE encoded by R2 elements is reported to have sequence and structural homology with the archaeal Holliday junction resolvases and a 3D model of R2 RLE has been built using the crystal structure of Holliday Junction resolvases C and E (Hjc and Hje) [1]. This had put forward the question if the R2 RLE could bind and cleave Holliday-junction like target DNA structures. Chapter 2 identifies that a Holliday-junction-like DNA structure is a critical component for second half of the integration mechanism. R2 protein would preferentially bind a nonspecific 4-way junction over a nonspecific linear DNA but would not cleave them symmetrically. However, the four-way junction containing 28S rDNA and R2 sequences shows robust site specific second strand cleavage which is almost symmetrical with the first-strand cleavage pre-engineered in the 4-way junction structure. The formation of the 4-way junction DNA is structurally relevant to the template jumping by the reverse transcriptase that has been hypothesized/shown to occur for both APE LINEs and RLE LINEs [2–7]. The identified 4-way junction integration-intermediate DNA structure is expected to be formed upon a post-TPRT template-jump by reverse transcriptase from the RNA to the target DNA (Chapter 2, Figure 3-8a step iii). The downstream 28S DNA sequence, the covalent linkage by template jump and the presence of the preexisting bottom-strand cleavage all appeared to be important sequence and structural components. As we expect for second strand cleavage, the endonuclease cleaved the 28S DNA second strand on the 4-way junction only in the absence of 5' PBM and 3' PBM RNA. The second strand cleavage of the 4-way junction seemed to create a primer-template setup between the cleaved target DNA and the first strand cDNA which led to the second strand synthesis product being observed for the first time in our *in vitro* assays. The protein that forms the downstream subunit was observed to be primarily involved in catalyzing the second strand synthesis. These findings have led us to an updated and detailed model of R2Bm integration mechanism (Chapter 2, Figure 3-8). The first two steps of the integration mechanism that includes first strand cleavage and first strand synthesis by TPRT, is same as that of the previous model. After TPRT, a template jump or a recombination event has been proposed as the third step in the new integration model. At the end of first strand synthesis by TPRT, when reverse transcriptase reaches the 5’ end of R2 RNA, the reverse transcriptase jumps from 5’ end of R2 RNA to the top strand of 28S rDNA upstream of the R2 insertion site. The template jump/recombination creates a 4-way junction intermediate DNA structure. In the fourth step of the integration, the endonuclease from downstream subunit recognizes the 4-way junction target DNA and cleaves the second strand. Finally, the reverse transcriptase from downstream subunit performs the second strand synthesis, completing element integration.

The template jump step as observed for R2 RT can also be functionally analogous to L1 RT (Chapter 2, Figure 2-8b). In L1 LINEs, second strand cleavage by APE occurs 14- 25 bp downstream relative to the first strand cleavage site, creating staggered breaks in the target DNA that are later filled in to form direct repeats flanking the newly inserted element (termed target site duplications or TSDs) [8]. After second strand cleavage, the 14-25 nucleotides on the bottom strand between the first and second strand nicked site melts away from the top strand. This enables the reverse transcriptase to template jump from 5' end of L1 RNA to the upstream DNA sequences from the top strand nicked site, thus forming a 4-way junction intermediate DNA structure shown to be essential for second strand cleavage and second strand synthesis in R2 elements. The proposed template jump step in the new integration model has provided a testable model of integration mechanism not only for other RLE LINEs but also for APE LINEs that bear completely different endonuclease.

6.2 Linker region

The reverse transcriptase (RT), linker and restriction like endonuclease (RLE) is known to form a large globular domain in R2 elements from *Bombyx mori* (R2Bm) [9]. While the RT and RLE are extensively characterized, the linker region remained as one of the understudied regions of the R2 encoded protein. Of particular interest in the linker region are the gag- like zinc knuckle (CCHC motif) and the predicted upstream α-helices (α-finger) which are found to be universally conserved not only in RLE LINEs but also across APE LINEs [9,10]. These regions in human L1 elements have long been speculated to be involved in RNA binding [11,12]. Our studies using R2Bm have shown that RNA binding does not seem to be the major role of the zinc knuckle and α-finger motifs. The mutations in these regions did not lead to the formation of RNA-free complexes in the EMSA gel when compared to WT protein. Given that the RNA binding surface of the protein could be broadly spread over the R2 protein surface, it is possible that our assay with double point mutations may not be sensitive enough to detect any effects. There remains the need for further exploration of RNA binding residues within R2 protein using protein-wide expansive approaches.

The CCHC motif of zinc knuckle and the residues in the upstream α-helices have been identified to be indispensable for successful retrotransposition of L1 and SART1 LINE elements *in vivo,* however, none of the studies have determined the actual biochemical role of these conserved motifs in the linker [11–14]. Our studies have now identified that the linker region is indeed biochemically crucial in coordinating nucleic acid and protein functions both in the first and the second half of the integration mechanism. In R2Bm, the core residues of zinc knuckle (CCHC motif) and α-finger (HINALP motif) appears to be structurally important for all protein functions. Mutations in the core residues lead to structural collapse of the local protein structure which in turn unconstrains the endonuclease domain promoting off target DNA cleavages and abolishes the ability of R2 protein to stably bind target DNA. Four arginine residues located in the α-finger have been identified to be important to correctly position the target DNA and/or the endonuclease for successful first-strand and second-strand DNA cleavages. Additionally, α-finger residues are involved in positioning the reverse transcriptase and/or the nucleic acid components required for effective first strand synthesis by TPRT. Zinc knuckle residues seem to have similar function especially in the second half of the integration. Both zinc knuckle and α-finger are equally essential for second strand synthesis as they properly locate the cleaved second strand plus the reverse transcriptase for primer extension. Both zinc knuckle and α-finger seem to be coordinating together to induce protein conformational changes required for proper activity of endonuclease and reverse transcriptase at each step of integration mechanism. Also, they appear to correctly present the bound nucleic acid components to the active sites of reverse transcriptase and endonuclease during the integration process. The α-finger, in addition, was found to be specifically recognizing and binding the 4-way junction intermediate DNA structure which is a gateway to the second half of the integration mechanism.

Given the universal conservation of linker region motifs, our findings can be directly implied to other RLE LINEs and also to the group of APE bearing LINEs. Although the location of endonuclease differs between the RLE and APE LINEs, the α-finger and zinc knuckle of linker region could still modulate the reverse transcriptase and endonuclease activity in APE LINEs by controlling how they position nucleic acids to the protein's active sites during integration process. It remains possible that these regions in APE LINEs could also induce protein conformational changes essential for protein function. Formation of a 4-way junction post-TPRT has been proposed for other groups of LINEs. Since our study has put forward α-finger to be a major determinant in recognizing the 4-way junction, similar functionality could hold true for α-finger motif across the groups of LINEs.

The similarity of the RT, linker, and RLE of R2 LINE with the large fragment of Prp8, a eukaryotic splicing factor, is very much apparent. The crystal structure of the large fragment of Prp8 was a major template to build the 3D model of R2Bm RT [9]. Also, Prp8 RLE is very similar to the R2Bm RLE sharing the unique spacing of the catalytic core residues [9]. Joining the RT and RLE in Prp8 is a linker region. Towards the end of the linker region in Prp8 lies 1585 loop-helices (α-finger) and ββα non-zinc knuckle which is found to align very well with the predicted α-finger helices and zinc knuckle region of both RLE and APE LINEs [9]. Different complexes are formed during spliceosome assembly and activation, and at each complex formation step the α-finger and non-zinc knuckle of Prp8 play major roles. At U4/U6.U5 tri-snRNP and B complex formation steps, the residues in the non-zinc knuckle and the thumb of Prp8 is shown to form a binding pocket for intron and they coordinate the positioning of the Guanine base (G) at the 5' splice site [15,16]. The α-finger of Prp8 is found to be highly dynamic during spliceosome rearrangements [17]. The interaction of 1585 loop with U2/U6 duplex, Cwc24, and Prp11 has shown to cage the guanine (G) nucleotide at the 5' splice site and is important to maintain catalytic dormancy of B^{act} complex, while the removal of Cwc24, and Prp11 from 1585 loop catalytically activates the B* complex [17,18] In C* complex, which is a second catalytically active stage of spliceosome, the 1585 loop is found to be essential to stabilize the active site conformation [17–19]. In post catalytic P complex, the 1585 loop interacts with the intron at 3' splice site and stabilizes the conformations [20]. The Prp8 1585 loop and non-zinc knuckle are found to interact with important branched RNA structures near the splice sites [15–17,20]. The findings from our experiments on the linker region are indicative that these regions of Prp8 could also be functionally very similar to LINEs.

6.3 DNA binding and R-box region

Two R2 protein subunits are believed to be required for integration, one bound to the upstream 28S DNA sequences and one bound to the downstream 28S DNA sequences from the insertion site. The N-terminal ZF and Myb motif of R2 protein have been identified to bind to the downstream sequences of 28S target DNA [21]. Mutations in the α-finger of linker has shown to induce moderate decrease in both upstream and downstream target DNA binding (Chapter 3, Figure 3-3). However, the α-finger and zinc knuckle have shown to be important in positioning the target DNA for endonuclease and reverse transcriptase activity. Our studies have identified α-finger of the linker region to have specificity to recognize the 4-way junction intermediate target DNA structure which is a major step required to proceed to the second half of the integration process.

R2 elements encode a variant of the PD-(D/E)XK superfamily of endonucleases [22]. A subset of type II restriction enzymes and holiday junction resolvases that share a αβββαβ restriction endonuclease-like fold similar to R2 RLE, consists of an arginine rich R-box region within or in front of the first α-helix [23–25]. The arginine and lysine residues make contacts with target DNA, and mutation in these residues leads to lack of DNA binding and hence cleavage [24–27]. The R2 endonuclease has similar R-box region with "RILRH" motif and mutating the RH have shown to reduce DNA binding [1]. In our study, the single mutation of the two arginine residues did not seem to be strong enough to produce any observable effect (Chapter 4). We expect that mutating both the arginine residues of R-box as a double mutant will show a reduction in DNA binding activity of protein. An additional phenotype shown by the single mutants was their inability to cleave the second strand when using a linear DNA in our *in vitro* assay. Additional experiments with a 4-way junction target DNA is required to make conclusions about their binding specificity and cleavage activity.

6.4 RNA binding and RT-1 region

RNA binding determines final protein conformation and dictates the role of each protein subunit in the integration reaction. Involvement of domains of R2 protein in RNA binding is just beginning to be understood. The N-terminal RT-1 and RT0 motifs have recently been recognized to bind both 3' PBM RNA and 5' PBM RNA [3]. Our study has identified a "Y" residue in the RT-1 region that specifically recognizes and binds 3' PBM RNA (Chapter 5). This finding has shown that there are additional residues within R2 protein that can differentiate and distinctively bind to either 3' PBM RNA or 5' PBM RNA. Identification of these residues will help us understand how R2 protein is able to distinguish between 3' PBM, 5' PBM, and non-specific RNAs and upon binding, how they regulate protein's DNA binding modules. In telomerase and group II intron, that share close evolutionary relationship with R2 LINEs, the RNA binding residues are not limited to the N-terminal regions but are distributed across the fingers, palm, and thumb domains of RT [28–30]. Additional high resolution scanning of the residues within the N-terminal, RT, linker, and endonuclease domains of R2 needs to be done. Also, a more global approach requires to be undertaken along with point mutational analysis for fast and accurate identification of areas of R2 protein interacting with RNA template.

6.5 Limitations

The α-finger of linker region have been identified to play a critical role in recognizing 4-way target DNA structure. But at this point, we do not know which arm(s) of the 4-way junction is an important determinant for being recognized by the α-finger. We still do not have a 3D model for the linker region nor we have a protein-nucleic acid structure to understand the details of how linker region modulates protein conformational change and positioning of nucleic acids at different steps of integration process.

The study with the RT-1 region and the R-box region have provided some important insights into nucleic acid binding activity of R2 protein. But additional studies on each of these regions remains to be done. Our studies did not identify the domain of R2 protein that binds to the upstream sequences of 28S target DNA from the insertion site. A larger study encompassing the point mutations in conjunction with global footprinting approach to track down the DNA/RNA binding surfaces of the protein needs to be carried out.

In hot DNA experiments that we conducted, we have tracked DNA binding and made inferences about RNA binding. Hot DNA experiments do give useful information in a boarder view as we can also track target DNA cleavages and new strand synthesis. However, hot DNA experiments may not be able to detect very subtle changes in binding affinity of R2 protein to 5' PBM and 3' PBM RNAs.

6.6 Future directions

The limitation of our studies will be addressed as a part of the future extension of current work. To determine which arm(s) of the 4-way junction target DNA are important for being recognized by the α-finger of the linker region, different 4-way junctions will be strategically built by swapping R2 specific sequences with non-specific sequences on each arm. Testing the ability of WT and α-finger mutant protein to bind the new set of junctions will help us identify the sequence and the structural component of the 4-way junction DNA that drives the binding specificity of R2 protein to the 4-way junction.

In addition, linear and 4-way junction target DNA will also be footprinted to investigate sequence specificity of the linker region of R2 protein. Hydroxyl radical/ Missing nucleoside footprinting techniques will be used to map the loss of DNA contacts in R2 mutants that have been shown to have impaired DNA binding activity [31,32]. This DNA footprinting technique utilizes Fenton's reaction to generate hydroxyl radicals that removes DNA bases and breaks the DNA backbone. The hydroxyl radical modification of DNA (pre and post protein binding) gives nucleotide level footprinting data. Hydroxyl radical footprinting is a protection assay in which R2 protein will first be bound to ^{32}P labeled target DNA followed by hydroxyl modification of the DNA. The DNA will be protected from modification where the protein is in close contact [31]. On the other hand, missing nucleoside footprinting is an interference assay in which ^{32}P labeled target DNA is first modified by hydroxyl radical cleavage and then assayed for R2 protein binding

[32]. In both the assays, the resulting footprint of the protein bound to DNA will be analyzed by electrophoresis on a denaturing DNA sequencing gels and visualized by autoradiography. The DNA footprint pattern of WT R2 protein bound to DNA will be compared to the DNA footprint pattern of mutant protein bound to DNA. If the mutation abolishes/reduces binding of R2 protein to specific nucleotides of target DNA, then the footprint for mutants will be different from that of WT. These techniques will point out the nucleotides on the target DNA that are contacted by linker region of R2 protein.

Protein footprinting will help us map out the yet unknown domain of the R2 protein involved in binding 28S DNA sequences upstream of the insertion site. Also, the RNA binding residues spread over the large surface of R2 protein can be established. Mass spectrometric footprinting is a technique that allows fast and accurate identification of amino acids in a protein that makes DNA/RNA contacts. The method utilizes a primary amine modifying reagent that only modifies available residues in free protein but not when the residues are masked due to protein-nucleic acid interactions [33]. There are 44 lysines and 123 arginines in the full-length R2 protein, many of which are likely to be on the surface of the R2 protein and involved in binding to nucleic acids [9]. Whole scale chemical modification of the lysine and arginine surface residues in the presence and absence of 3' PBM RNA, 5' PBM RNA, linear target DNA, and branched target DNA will provide a footprint of which of the lysine and arginine residues bind to or are otherwise blocked from chemical modification upon binding these nucleic acids. One of the modifying reagents that can be used is N-hydroxysuccinimide (NHS)-biotin that biotinylate lysine residues. The modifying agent exposed protein-nucleic acid complexes will be subjected to SDS gel electrophoresis followed by in-gel proteolysis which produces short peptide fragments that can be analyzed by mass spectrometry [33]. The biotinylated peptide peak can be identified and assigned to a particular lysine residue. Comparison of peaks obtained from free protein vs RNA/DNA bound protein readily shows the residues modified in free but protected in protein-nucleic acid complexes. These protected residues are the ones involved in RNA/DNA interactions [33]. In a similar way, the arginine residues can be modified by using p-hydroxyphenylglyoxal (HPG) / 1,2-cyclohexanedione (CHD) and mapped by mass spectrometry for nucleic acid interactions [34,35]. Amino acids identified from the footprinting assay will be mutated to alanine and tested *in vitro* for loss of nucleic acid binding function. Additionally, RNA binding assays will also be conducted using radiolabeled 3' PBM and 5' PBM RNAs.

Finally, to get a high-resolution structure of the R2 protein, R2 RNP, and R2 DNA complexes, single particle reconstruction techniques like cryo-electron microscopy (cryo-EM)[36] will be used which would be a huge step forward in understanding the overall protein structure and nucleic acid interactions at near atomic level.

6.7 References

1. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: Loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res.* **44,** 3276–3287 (2016).
2. Bibiłło, A. & Eickbush, T. H. The reverse transcriptase of the R2 non-LTR retrotransposon: Continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* (2002). doi:10.1006/jmbi.2001.5369
3. Jamburuthugoda, V. K. & Eickbush, T. H. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res.* **42,** 8405–8415 (2014).
4. Szafranski, K., Dingermann, T., Glöckner, G. & Winckler, T. Template jumping by a LINE reverse transcriptase has created a SINE-like 5S rRNA retropseudogene in Dictyostelium. *Mol. Genet. Genomics* (2004). doi:10.1007/s00438-003-0961-9
5. Bibillo, A. & Eickbush, T. H. End-to-End Template Jumping by the Reverse Transcriptase Encoded by the R2 Retrotransposon. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M310450200
6. George, J. A., Burke, W. D. & Eickbush, T. H. Analysis of the 5′ junctions of R2 insertions with the 28S gene: Implications for non-LTR retrotransposition. *Genetics* (1996).
7. Han, J. S. & Shao, S. Circular retrotransposition products generated by a LINE retrotransposon. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks859
8. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Mol. Cell. Biol.* (2005). doi:10.1128/MCB.25.17.7780-7795.2005
9. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: Similarities to eukaryotic splicing factor Prp8. *Mob. DNA* **8,** 1–15 (2017).
10. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
11. Doucet, A. J. *et al.* Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet.* **6,** 1–19 (2010).
12. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3,** 433–437 (2013).
13. Moran, J., Holmes, S. & Naas, T. High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
14. Takahashi, H. & Fujiwara, H. Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* (2002). doi:10.1093/emboj/21.3.408
15. Wan, R. *et al.* The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science (80-.).* (2016). doi:10.1126/science.aad6466
16. Nguyen, T. H. D. *et al.* Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530,** 298–302 (2016).
17. Shi, Y. The Spliceosome: A Protein-Directed Metalloribozyme. *J. Mol. Biol.* **429,** 2640–2653 (2017).
18. Yan, C., Wan, R., Bai, R., Huang, G. & Shi, Y. Structure of a yeast step II catalytically activated spliceosome. *Science (80-.).* (2017). doi:10.1126/science.aak9979
19. Fica, S. M. *et al.* Structure of a spliceosome remodelled for exon ligation. *Nature* (2017). doi:10.1038/nature21078
20. Bai, R., Yan, C., Wan, R., Lei, J. & Shi, Y. Structure of the Post-catalytic Spliceosome from Saccharomyces cerevisiae. *Cell* **171,** 1589–1598.e8 (2017).
21. Christensen, S. M., Bibillo, A. & Eickbush, T. H. 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).
22. Yang, J., Malik, H. S. & Eickbush, T. H. 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,** 7847–52 (1999).
23. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. Type II restriction endonucleases: Structure and mechanism. *Cellular and Molecular Life Sciences* **62,** 685–707 (2005).
24. Pingoud, V. *et al.* 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.* (2005).

doi:10.1074/jbc.M409020200

25. Pingoud, V. *et al.* 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.* (2003). doi:10.1016/S0022-2836(03)00523-0

26. Nishino, T., Komori, K., Ishino, Y. & Morikawa, K. Dissection of the Regional Roles of the Archaeal Holliday Junction Resolvase Hjc by Structural and Mutational Analyses. *J. Biol. Chem.* (2001). doi:10.1074/jbc.M104460200

27. Nishino, T., Komori, K., Tsuchiya, D., Ishino, Y. & Morikawa, K. Crystal structure of the archaeal Holliday junction resolvase Hjc and implications for DNA recognition. *Structure* (2001). doi:10.1016/S0969-2126(01)00576-7

28. Gu, S.-Q. *et al.* Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16,** 732–747 (2010).

29. Rouda, S. & Skordalakes, E. Structure of the RNA-Binding Domain of Telomerase: Implications for RNA Recognition and Binding. *Structure* **15,** 1403–1412 (2007).

30. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat. Struct. Mol. Biol.* **17,** 513–518 (2010).

31. Jain, S. S. & Tullius, T. D. Footprinting protein-DNA complexes using the hydroxyl radical. *Nat. Protoc.* **3,** 1092–1100 (2008).

32. Hayes, J. J. & Tullius, T. D. The missing nucleoside experiment: A new technique to study recognition of DNA by protein. *Biochemistry* **28,** 9521–9527 (1989).

33. Kvaratskhelia, M. & Grice, S. F. J. Le. Structural analysis of protein-RNA interactions with mass spectrometry. *Methods Mol. Biol.* **488,** 213–9 (2008).

34. Wanigasekara, M. S. K. & Chowdhury, S. M. Evaluation of chemical labeling methods for identifying functional arginine residues of proteins by mass spectrometry. *Anal. Chim. Acta* (2016). doi:10.1016/j.aca.2016.06.051

35. Bailey, G. S. (Humana P. I. *The Protein Protocols Handbook*. *The Protein Protocols Handbook* (2009). doi:10.1007/978-1-59745-198-7

36. Zhang, X. *et al.* Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc. Natl. Acad. Sci.* (2008). doi:10.1073/pnas.0711623105