# GLOBAL MAPPING OF PROTEIN-NUCLEIC ACID INTERACTION

# OF THE RLE LINE R2Bm

by

SANTOSH DHAMALA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Biology at
The University of Texas at Arlington
May, 2022

Arlington, Texas

Supervising Committee:

Dr. Shawn Christensen, Supervising Professor
Dr. Clay Clark
Dr. Matthew Fujita
Dr. Mark Pellegrino
Dr. Saiful Chowdhury

# ABSTRACT

# GLOBAL MAPPING OF PROTEIN-NUCLEIC ACID INTERACTION OF THE RLE LINE R2Bm

Santosh Dhamala

The University of Texas at Arlington, 2022

Supervising Professor: Shawn Christensen

The R2 Long Interspersed Nucleotide Elements (LINEs) is a widely distributed site-specific non-LTR retrotransposons that integrates exclusively into host genome's 28s rRNA genes and replicate by a process called Target Primed Reverse Transcription (TPRT) which is a "copy-out, copy-in" mechanism whereby element protein binds to the mRNA from which they are translated, forming ribonucleic acid protein particles (RNPs).

The investigation of nucleic acid-protein interactions using chemical modification of surface amino acid residues on a protein in the presence and absence of nucleic acids is a relatively new application/technique in the field of Mass Spectrometry and particularly significant in studying protein-nucleic acid communication in nucleoprotein complexes. Here we exploit similar technique to modify the lysine surface residues in presence and absence of nucleic acids to gain insights on which lysine residues are involved in DNA/RNA binding. Upon study using selective modification of lysine residues and subsequent mass spectrometry, we find some new residues  in R2 protein which are potentially involved in the DNA binding function during retrotransposition.

# ACKNOWLEDGEMENTS

The completion of this study could not have been possible without the expertise of my beloved supervisor, Dr. Shawn Christensen. I would like to express my special appreciation to him for his constant motivation, direction and support over the years. His guidance on experimental design and data analysis has always been key for me to develop better understanding of the subject.

I am also sincerely grateful to my committee members: Dr. Clay Clark, Dr. Matthew Fujita, Dr. Mark Pellegrino and Dr. Saiful Chowdhury for their constructive comments and suggestion during my research.

I am highly indebted to the past and present graduate students from Dr. Chowdhury's lab (Jayanta Kishor Chakrabarty, Zixiang Fang and Adway Ouseph Zacharias) who provided the technical help with nano-LC-ESI mass spectrometer.

I am obliged to my labmates who have willingly helped me out with their abilities and provided constant cooperation and encouragement during my research.

Last but not the least, I thank all the staffs of Department of Biology for providing a prosperous atmosphere for learning and research.

# TABLE OF CONTENTS

## LIST OF ILLUSTRATIONS

Figure Page

# CHAPTER 1

# INTRODUCTION

## 1.1. Transposable elements

Transposable elements, or transposons, are discrete mobile segments of DNA that can move around the genome and insert into different DNA target sites using a specialized type of recombination called transposition. Transposable elements are found in all living organism; indeed, they are often the largest components of the genomes of multicellular organisms (for instance, about 45% of the human genome derives from transposable element sequences) (1,2). Although the transposable elements can be considered to be genome parasites, with the selfish function of simply propagating themselves, they also have many effects on genome structure and function, and they contribute to the genetic variation that sis the substrate for evolution (3).

## 1.2. Retrotransposons

The ancient genetic mobile elements that utilize reverse transcription of RNA templates to make additional copies of themselves are generally referred to as Retrotransposons, or also called retroelements. On the basis of mobility mechanism, retrotransposons can be grouped into two major classes (4).

i) *LTR retrotransposons:* The LTR elements, which include retroviruses and retroviral-like elements, are characterized by long terminal repeats of hundreds of base pairs at each end of the element. These LTRs are required for reverse transcription by the element-encoded reverse transcriptase, which converts the RNA form of the element to the double-stranded form (4,5).

This DNA is then inserted into the target site by an element encoded integrase; the mechanism if insertion is identical to that used by DNA mediated mobile elements (transposons)

*ii) Non-LTR retrotransposons:* The non-LTR elements also move via an RNA intermediate but lack long terminal repeats, and simply reverse transcribe a cDNA copy of their RNA transcript directly into the chromosomal target site (6). These elements use a different transposition mechanism called target-primed reverse transcription (TPRT).

### 1.3. Non-LTR retrotransposons

Although DNA transposons and LTR-elements are widespread, it is the non-LTR retrotransposons that make up a large proportion of the DNA of many eukaryotic genomes. Strikingly, These elements comprise about 35% of the human genome (7). In eukaryotes, one autonomous family of non-LTR elements called long interspersed elements (LINEs), which are prominent in the human genome, encode proteins that mediate their own transposition. Non-autonomous non-LTR elements that do not encode transposition proteins (and so rely on ‘non-self’ proteins for their movement) are called short interspersed elements (SINEs) and cane be mobilized by LINE-encoded proteins (8,9).

In LINEs, Transcription of the full length element is either driven by an internal RNA Polymerase II promoter at the 5’ end of the element or transcribed using an upstream promotor near the insertion site (10). The element encoded proteins display a strong *cis preference* to their encoding mRNAs to form an element specific ribonucleoprotein particle (RNP) that is involved in the retrotransposition process. The RNP is transported into the nucleus or interacts with the DNA during nuclear envelope breakdown. The element encoded endonuclease nicks the target

DNA. The resulting free 3'-OH us used to prime reverse transcription (11). This process of integration is called Target Primed Reverse Transcription (TPRT). Tpically, a non-LTR element has one or two open reading frames encoding a reverse transcriptase domain, a non-LTR specific cysteine-histidine rich motif (CCHC), an endonuclease domain, DNA and RNA binding domains and/or a ribonuclease H domain.

These non-LTR retrotransposons have been classified into two large groups on the basis of their structural and phylogenetic features. Based on the endonuclease the element contains, non-LTR retrotransposons can be divided into Apurinic Apyrimidinic endonuclease (APE) containing elements and restriction like endonuclease (RLE) containing elements. RLE bearing LINEs appear to be the earlier branching of the two groups, while APE containing elements are appear to be more recent (12).

### 1.3.1. APE containing non LTR retrotransposons

These later branching clades generally have two non-overlapping open reading frames. The role of ORF1p has been hard to understand since its amino terminal end appears to have no sequence homology to any known protein sequence (13). However, studies conducted with mouse and human L1s have revealed several domains in ORF1p that are critical to retrotransposition like nucleic acid binding and RNP formation (14,15). The ORF2p of APE non-LTRs is a multi-domain protein with an APE endonuclease domain, a reverse transcriptase domain and the CCHC domain (16,17) *(Figure 1.1)*. The APE LINEs tend to be non site specific. APE LINEs include the L1, RTE, Jockey, I factor, Tras, etc. Among the APE LINEs, L1 elements are best studied.

LINE1 transposes via an RNA intermediate like the other retrotransposons. Once in the cytoplasm, L1 transcript serves as a template for the translation of ORF1 and ORF2 protein. The two encoded proteins show very strong cis-preference and associate with the RNA from which they were translated to form the ribonucleoprotein (RNP). RNP then interacts with the target DNA and inserts a copy of the transposon using a mechanism known as target primed reverse transcription (TPRT) (13,15,18).

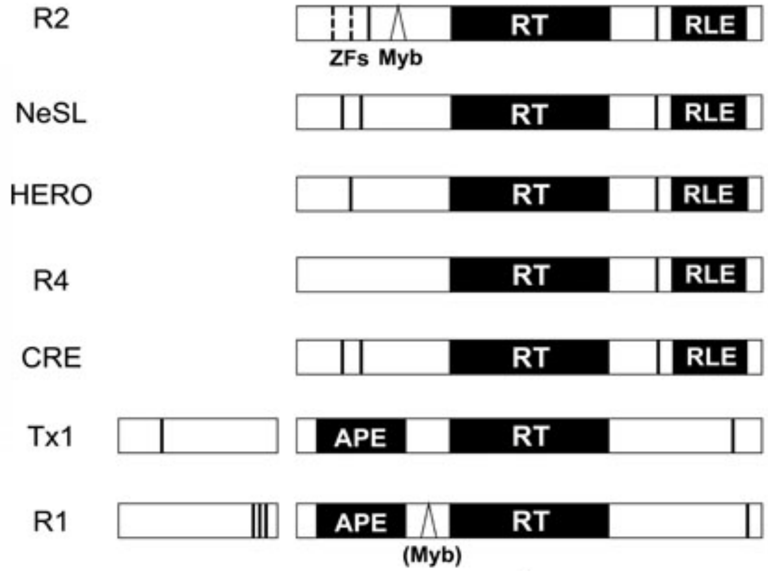

**Figure 1.1: Schematic structure of RLE-encoding *(first five)* and APE-encoding non-LTR retrotransposons *(last two*).** In APE-encoding elements, two ORFs are shown; ORF1 shown as small rectangle and ORF2 as large rectangle. Zinc-finger (ZF) like structure is shown as vertical bold line. Some R2 clade elements have addition ZFs represented by dotted vertical line. Figure adapted from 22.

### 1.3.2. RLE containing non LTR retrotransposons

The RLE-encoding non-LTR elements are a phylogenetically ancient class and are further categorized into five clades: Genie, CRE, R2, R4 and NeSL clades and tend to be site-specific in their integration. Members of the R2 and R4 clades insert specifically into the rRNA genes of their host. In case of the CRE and NeSL clades, the target sites are the tandemly repeated spliced leader exons of trypanosomes and nematodes, respectively (19-21). All of these non-LTRs contain a single open reading frame encoding a central reverse transcriptase, a cysteine-histidine motif (CCHC), and basic RH patch and a restriction like endonuclease at the carboxyl terminus. ORF structure of some RLE along with some APE bearing LINES from different clades is shown in *Figure 1.1.*

R2 element, which will be discussed below in detail, is one of the most studied site-specific non-LTR elements and has served as a model for understanding RLE bearing elements as a whole.

## 1.4. R2 element

R2 is a site-specific LINE that insert into a specific site in the 28S rDNA. R2 was first identified as an insertion sequence in the fruit fly, *Drosophila melanogaster,* and the domestic silkworm, *Bombyx mori* (22). The major features of the R2 group of elements is a single ORF *(Figure 1.2)* and the presence of a C-terminal restriction-like endonuclease (EN) domain and a large central Reverse Transcriptase (RT) domain (23,24). In addition, all R2 elements have an extensive N- terminal region that usually contains zinc-finger (ZF) and myb domains which are used by the R2 protein to bind the DNA downstream of integration site (22).

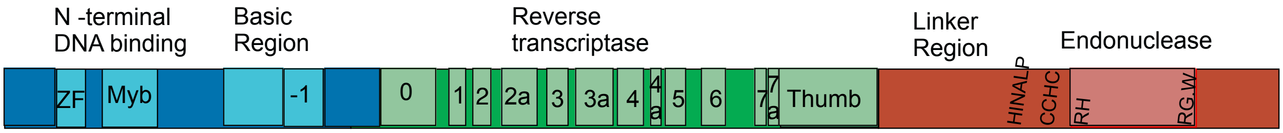

**Figure 1.2. ORF structure of R2 protein.** The N-terminal region (blue) contains zinc finger (ZF) and Myb motifs for DNA binding. In green is the reverse transcriptase (RT) region with standard RT motifs (RT0 through RT7 and thumb region). In red is the C-terminal region that has endonuclease (EN) activity as well as HINALP and CCHC motifs.

R2 are subdivided into four clades: R2-A, R2-B, R2-C and R2-D. In the amino terminal region, R2-A elements encodes three zinc fingers, R2-C elements encodes two, and R2-D elements encodes one zinc finger motif. The amino terminal structure of R2-B clade has not yet been determined. Apart from the variability of the N-terminal ZF motifs, all the R2 clades share same functional domains. (11)

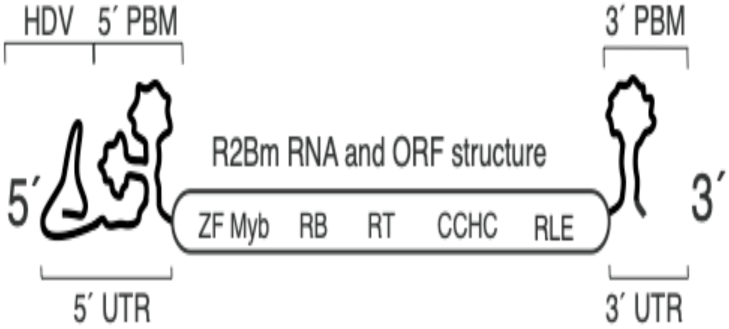

**Figure 1.3: ORF structure of R2 protein and the Protein Binding motifs RNAs.** R2 element RNA is special in a sense that it contains two specific protein binding motifs, one near the start of open reading frame which is called 5’ PBM RNA and the other located at the 3’ untranslated region, the 3’ PBM RNA.

### 1.4.1. R2 RNA and R2 Ribonucleoprotein complex (RNP) formation

Transcription of R2 element is controlled by upstream cellular promoter that transcribes their RNA as a co-transcript. R2 elements rely on the transcription of host rRNA gene using RNA polymerase I to be transcribed as 28S/R2 co-transcript (23). 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 (24,25). The 5’ ribozyme encoded by R2 RNA enables processing of R2 RNA from the 28S/R2 co-transcript. After transcription of LINEs, the RNA transcript is transported to the cytoplasm where it gets translated into protein. The protein translated from R2 has a strong *cis* preference and binds to the mRNA from which they were translated to form a Ribonucleoprotein (RNP) complex essential for Target Primed Reverse Transcription (TPRT). The 5’ and 3’ Untranslated Regions (5’ and 3’  UTR) 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) *(Figure 1.3)*. 5’ PBM is formed by a 319 nt segment that starts within the 5’ and ends just before the N-terminal ZF while the 3’ PBM constitute the 248 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 (26,27) *(Figure 1.4)*. R2 protein bound to 3’ PBM adopts a conformation that favors the binding 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. And on binding to the binding 5’ PBM, the R2 protein attains another conformation which allows it to binds downstream of the 28S rDNA and this downstream subunit is responsible for second strand cleavage and possibly second strand synthesis to complete R2 integration (28).

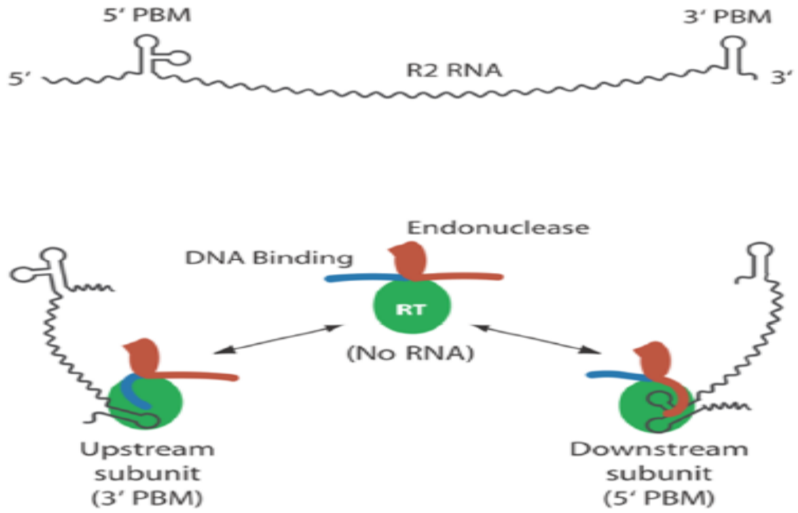

**Figure 1.4: Interaction of R2 protein with 3’ and 5’ PBM RNA to form upstream and downstream DNA binding RNP complex.** The binding of R2 protein subunit to the 5’ PBM and 3’ PBM RNAs dictates the protein’s conformation and in turn the role in the integration reaction. Figure adapted from 27.

1.4.2. R2 integration mechanism

During the integration of R2 element, an RNA copy of the element associated with the target site and is then used as a template for reverse transcription *in situ* at the site of insertion. In R2 RLE LINEs, the integration process is catalyzed by upstream and downstream protein subunits *(Figure 1.5)*. Protein bound to 3' 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 pairs downstream of the insertion site in DNA foot-printing assay (29). According to the basic and most well understood model, R2 is believed to integrate into target DNA via TPRT in four steps. *(Figure 1.5, steps 1-4)*. (Step-1) The endonuclease from upstream subunit nicks the target DNA exposing a 3'-OH at the insertion site; (Step-2) The exposed 3'-OH is used as a primer by the upstream subunit's reverse transcriptase for TPRT; (Step-3) The protein subunit bound to 5' PBM RNA binds to  target DNA downstream of the insertion site and cleaves  the second (top) strand. (Step-4) the 3'-OH generated by the cleavage event is used as the primer for second strand DNA synthesis of the element. Recently, a new and comprehensive integration model *(Figure 6)* has been developed which addressed the open question of whether R2 protein could

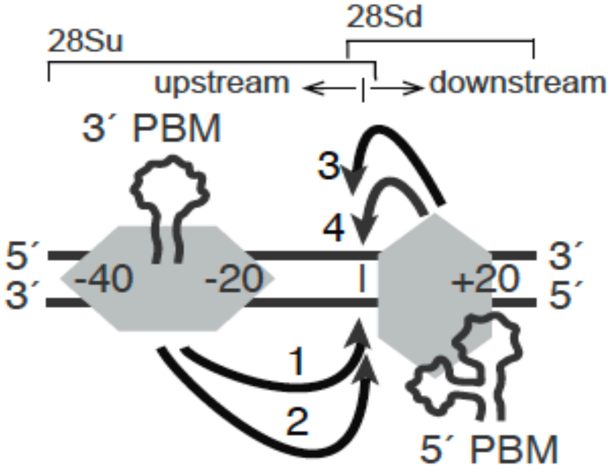

**Figure 1.5: R2 integration mechanism.** 1. DNA cleavage of the first/ antisense strand. 2. First strand DNA synthesis by TPRT. 3. DNA cleavage of the second/ sense strand. 4. Second strand DNA synthesis. Figure adapted from 30

function on branched DNA molecules like pseudo 3-way junction and pseudo 4-way junction. The first half of the integration reaction is identical to steps 1 and 2 *(in Figure 1.5)*. However, differing from the previous model in the timing and substrate of second strand cleavage (31), this model establishes a template jump or recombination event where the cDNA from the 5' end of the R2 RNA becomes associated with the upstream target DNA sequences to form a 4-way junction *(Step iii in figure 1.6)*. The open 4-way junction thus formed upon second-strand DNA cleavage *(Step iv)* resolves into a natural primer-template that is then used in second-strand DNA synthesis *(Step v)*.

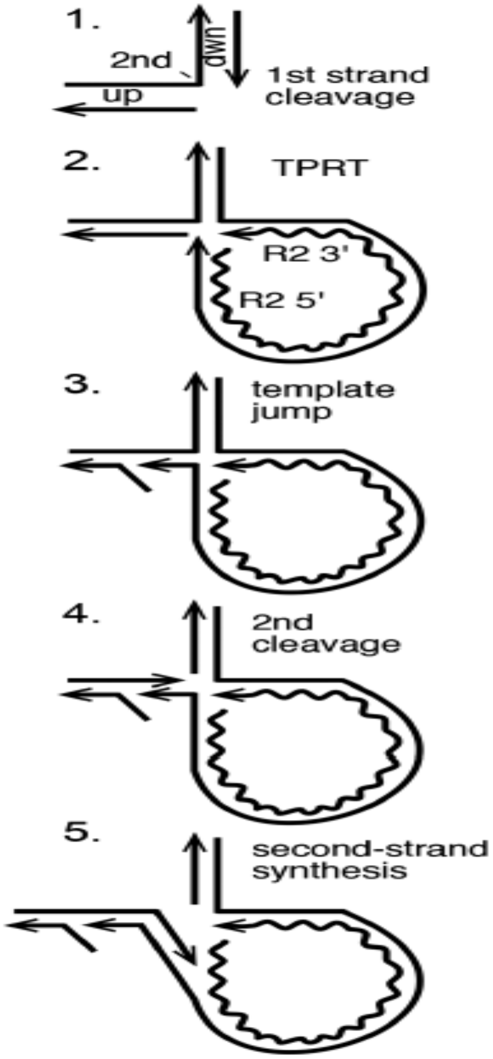

**Figure 1.6: New model of R2 integration.** The initial steps of the integration reaction (i, ii) are as in Figure 5 except that the target site is bent 90º near the second strand insertion site for better illustration. Step iii is template jump/recombination event near the second-strand cleavage site that generates the 4-way junction. Step iv outlines second-strand cleavage. Finally, step v depicts the second-strand DNA synthesis. Figure adapted from 31.

## 1.5. Scope of the Thesis

The knowledge of R2 integration mechanism has come a long way down the years, especially upon the characterization of the important functional domains of the protein; N-terminal domain with the DNA binding region, the reverse transcriptase domain and the endonuclease domain. This understanding is even more so important because R2Bm is been extensively used as the model system not only to understand its integration mechanism, but also to facilitate the significant understanding of integration mechanism of other RLE LINEs as well as APE LINEs. However, there are still some outstanding issues about know-how of protein domains/regions involved in nucleic acid binding.

The protein bound to the 3‘ PBM binds upstream of the insertion site whereas the protein bound to the 5‘ PBM binds downstream of the insertion site (29). And, it has been shown that the amino terminal zinc finger and myb motifs are responsible for downstream DNA binding activity (32). However, the DNA binding domain involved in upstream DNA binding is yet to be determined. Likewise as far as RNA is concerned, the binding ability has been previously mapped to domain -1 and domain 0 just upstream to Reverse Transcriptase (33). However, it is highly probable that RNA contacts is not limited to just these two regions. The CCHC domain of of ORP2p in LINE-1 has been shown to potentially be involved in RNA-protein interactions (13,32). Since the CCHC is highly conserved and is found in APE as well RLE elements like R2, it would be interesting to see if it plays a similar role in the R2 element. RNA binding plays a major role in determining protein subunit conformation and activity in the integration reaction. It is highly likely that RNA binding might completely or to some degree block the binding regions on the R2 protein as a mechanism to force binding to either upstream or downstream DNA sequences in the presence of 3’ PBM and 5’ PBM RNA, respectively.

The major goal of my research is to identify the regions in R2 protein that are involved in nucleic acid binding by the use of lysine residue modification followed by mass spectrometric analysis.

# CHAPTER 2

# PROTEIN FOOT-PRINTING: USE OF MASS SPECTROMETRY FOR GLOBAL MAPPING OF PROTEIN-NUCLEIC ACID INTERACTION IN R2Bm INTEGRATION MECHANISM

## 2.1. Background and Significance

It has always been interesting to know which part of the R2 protein interacts with the nucleic acid (DNA or RNA) during the integration mechanism. Reports have shown that ZF and myb motifs of R2 protein is used to bind DNA sequences downstream of target site while it is still unclear which part of protein has the same role in binding upstream DNA sequences. Protein footprint analysis on R2 RLE LINE from *Bombyx mori* (R2Bm) has shown DNA binding site being separate from the cleavage site, as the protein was found to bind the upstream and downstream sequences from the insertion site (29,34). 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 (21). 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 (21).

Also, it is now well established that R2 protein adopts different distinct conformations upon binding to RNA and the ribonucleoprotein (RNP) drives the integration mechanism. R2 protein bound to 3’ PBM RNA adopts a conformation that allows it to bind 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 (35,36). 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 (37). RNA is considered to contact R2 protein via this conserved region of the transcript. Similarly, R2 protein bound to 5’ PBM RNA binds downstream of the insertion site and this downstream subunit is responsible for second strand cleavage and possibly second strand synthesis to complete R2 integration. Recently, couple of RNA binding motifs in R2 protein has been mapped. Both motifs located immediately N-terminal of Reverse transcriptase (pre-RT region); one in -1 region and another in 0 region, have been identified to have both 3’ and 5’ PBM binding ability (33). Sequence similarity in 0 motifs has been found in all lineages of LINEs (38) implying a similar RNA binding domain across lineages.

The large globular domain of R2Bm, shares structural as well as sequence similarities to the large fragment of prp8, a highly conserved eukaryotic splicing factor. 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 are 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 which protrudes out over the reverse transcriptase. It is by analogy to the α-finger in Prp8 that the corresponding region of the RLE LINEs is called the ‘presumptive α-finger’. In Prp8 the non-zinc knuckle, the α-finger, and the RT thumb work together to bind the splice sites and spliceosomal RNAs. Cryo-EM structure of the Prp8 and RNA from the spliceosome complex is shown in *Figure 2.1 D,* the prp8 is buried in intronic RNA and we believe that similar could be the case for R2 protein given that great degree of structural and sequence similarity to each other.

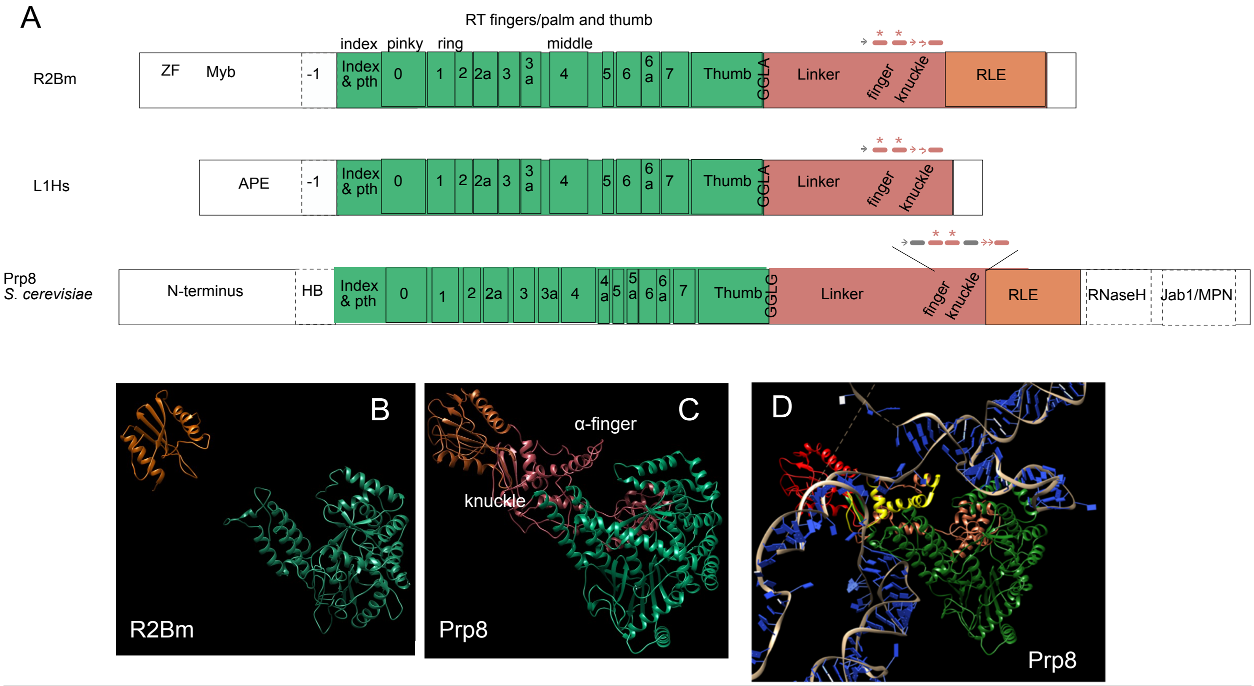

**Figure 2.1: Similarities between R2Bm and Prp8.** A. The ORF structure of R2Bm, human L1 (L1Hs), and *Saccharomyces cerevisiae* Prp8 are presented as color block diagram. The RT is green, the linker is maroon, and the RLE is orange. In the linker region, the sequences of the orange colored α-helices (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 (23). The ribbons have been colored as in the corresponding color block diagram. C. Cryo-Em structure of the large fragment of Prp8 . Ribbon color is matching the corresponding color block diagram.  D. Cryo-EM structure of the Prp8 and RNA from the B spliceosome complex (24). 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. Figure adapted form 39,40.

It is now thought that the R2 encoded protein sits in the middle of nest of complicated sets of element RNA and target DNA segments, much like the spliceosomal protein prp8. In addition to the extensive RNA contacts, R2 must recognize linear DNA, open 3-way junction

DNA and open 4-way junction during different stages of the integration reaction. Also as the integration reaction proceeds, the DNA recognition becomes more shape driven and the protein-nucleic acid complexes becomes void of element RNA during reverse transcription of the element RNA. So, it is highly likely that possibility of RNA binding or say nucleic acid binding in general is more extensive than previously envisioned, it is thought to be widely distributed across the protein surface, rather than just being limited to these regions.

The concept of using modification reagents to map protein structure has been around for at least 40 years, but the more recent ability of Mass Spectrometry (MS) to quickly, sensitively, accurately, and precisely map protein modification sites has made this approaches much more powerful method to obtain protein structural information (41). Combination of chemical modification of lysines and multiple-reaction monitoring mass spectrometry was done to identify putative substrate-contacting residues in *Arabidopsis thaliana* PRORP1 (AtPRORP1), confirming protein- and RNA-based forms of RNase P have distinct modules for substrate recognition and cleavage (42). Similarly, MS-based protein footprinting by NHS-Biotin modification and limited proteolysis was used to locate the lysine residues in DNA binding domain and C-terminal domain that were involved in XPA-DNA interaction (43). Using two-stage amino acid probing combined with MS/MS, the reactivity of lysine side chains of the proteins in spliceosome complexes was assessed to identify that the Prp8 position of U5 snRNA being linked to 5' splice site recognition (44).

Analysis of structural information for proteins or protein-ligand complexes with amino acid residue modification approach typically relies on the differential reactivity of amino acids upon exposure to a particular label. The implied assumption in these experiments is that amino acids that are exposed to solvent and, therefore, accessible to a modifying chemical reagent will be modified, whereas buried amino acids will be modified slowly or not at all. The amino acids that become more or less accessible to the reagent will react to greater or lesser extents respectively; tracking that particular difference gives an idea of their involvement in a conformational change and/or their presence at a ligand-binding site. Nano-LC-ESI-MS/MS is used to identify the location and extent of two different lysine modifications (Acetylated vs Duetero-acetylated). Analysis of mass spectrum for PSM values of peptides with differentially modified lysine residues and the ratio of second to first acetylation in different combinations provided information on which lysine residues are otherwise protected from chemical modification because of binding to the nucleic acids.

## 2.2. Results

Different combinations of R2 protein-nucleic acid complexes were designed such that they mimic different stages of the integration reaction. The complexes were then subjected to two-step acetylation reactions to chemically modify lysine resides in presence and absence of nucleic acid *(Figure 2.2)* using the reagents: i) Sulfo-NHS Acetate and ii) Deuterated Acetic Anhydride. Before actually subjecting the protein or RNP complex to chemical modification, effect of reagent modification on nucleic acid binding is determined by gel mobility assays to optimize the concentration of reagent needed for lysine modification *(Figure 2.3 A).* This optimization is important to ensure the reagent concentration sufficient for modification of solvent-accessible lysine residues of R2 protein to prevent binding while having minimal effects

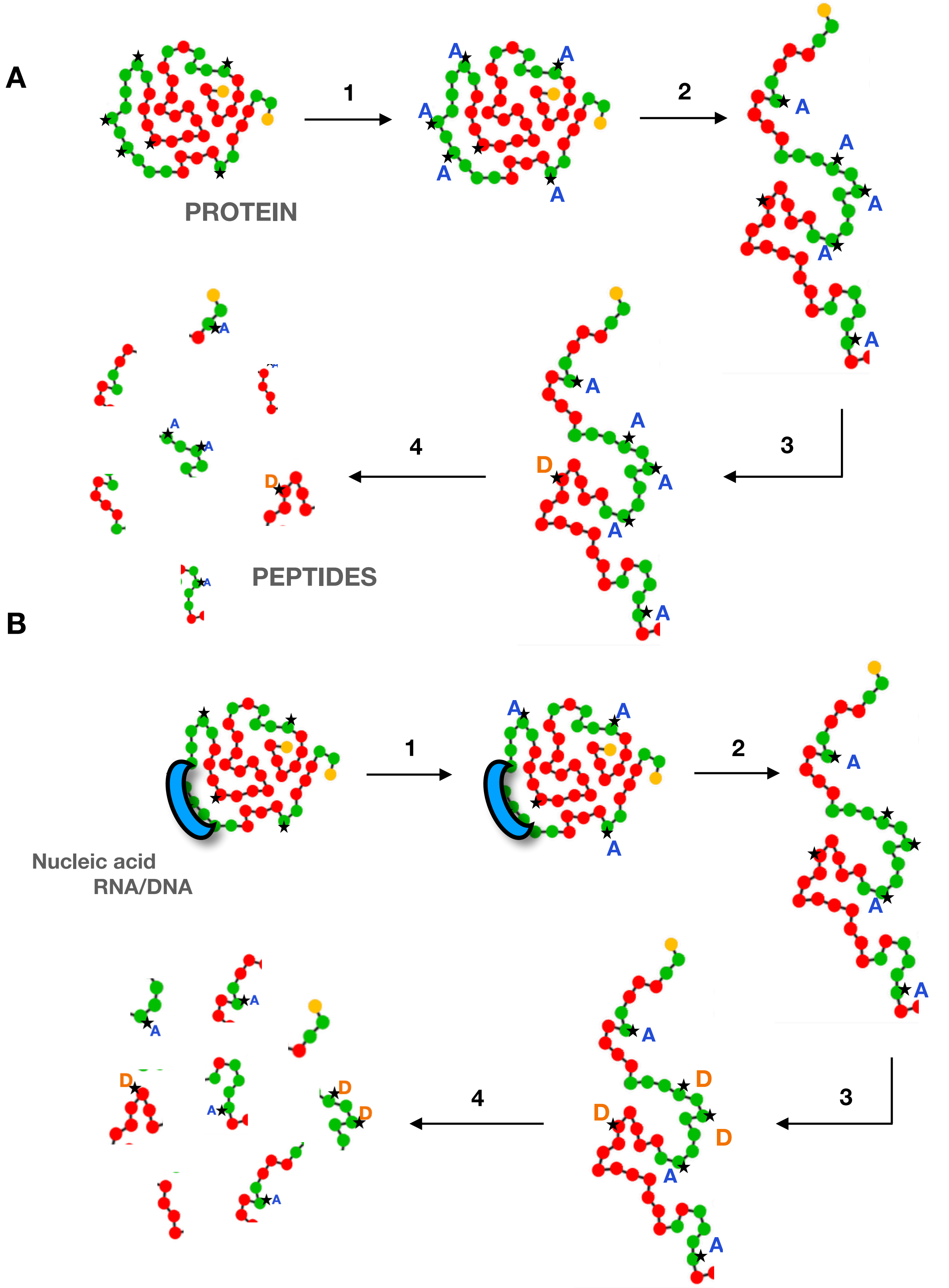

**Figure 2.2: Schematic approach of protein footprinting by modification of lysine residues (marked with stars):** A. In absence of nucleic acid, and B. in presence of nucleic acid. Step 1: the surface lysines are acetylated by Sulfo-NHS Acetate (acetylated lysines marked with blue colored 'A'). Step 2: the protein is denatured. Step 3: the originally buried residues and/or the residues protected by nucleic acid binding are modified by Dueterated Acetic Anhydride (duetero-acetylated lysines marked with orange colored 'D'). Step 4: Cleavage with trypsin to generate peptides which are then purified and forwarded to mass-spectrometric analysis (not shown here)

on complex integrity. From the result of EMSA gel performed to optimize the concentration of acetate reagent and time of acetylation reaction *(Figure 2.3 B),* 15mM final acetate concentration and 30 min incubation time was chosen to carry out the reaction. The least concentration of acetate that began to hamper the RNA and DNA binding was chosen.

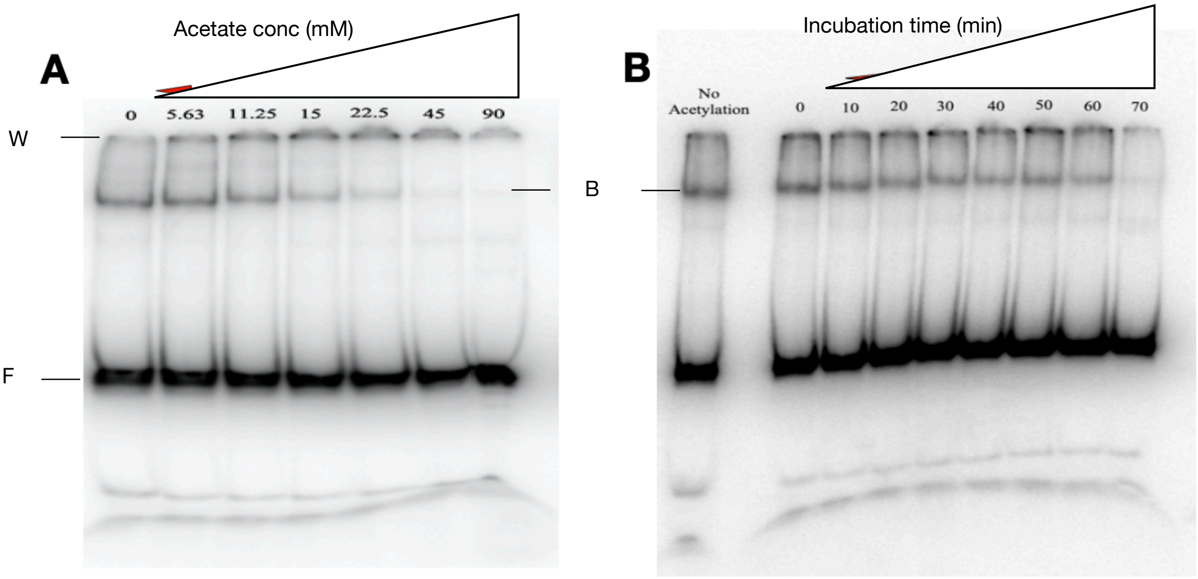

**Figure 2.3: Effect of acetate concentration on RNA binding activity of R2Bm protein**

A. EMSA for checking effect of increasing acetate reagent concentration in the preservation of binding activity in the presence of nucleic acid. Reaction Ratio: 5 pmol R2, 100 pmol 5' PBM RNA, 25 pmol linear DNA (Spiked with bottom strand radiolabeled DNA), acetate concentration as labeled in the gel
B. EMSA for checking acetylation efficiency at 15 mM final concentration vs time of acetylation. Reaction Ratio: 5 pmol R2, 100 pmol 5' PBM RNA, 25 pmol linear DNA (Spiked with bottom strand radiolabeled DNA) 15 mM final acetate concentration.

Abbreviations: W- Well complex, F- Free DNA, B- Bound/ shifted DNA

Upon first acetylation, the surface exposed lysine residues gets modified leaving the buried residues unmodified or very low degree of modification. The second acetylation done after  denaturation of protein modifies all those lysine residues which were buried in the first place but now exposed and accessible to the reagent. However, in case of combination of RNPs (R2+RNA/DNA/TPRT reaction intermediates), the lysine residues which are potentially involved in binding to nucleic acids are protected from chemical modification during the first acetylation, making it only possible to get modified by the second stage of modification.

A control set (A) protein alone, no nucleic acid and following two different combination of protein-nucleic acid complex were formed *in vitro* from purified R2 protein; (B) R2 protein complexed to 3’ PBM RNA, and (C) R2 protein complexed to 3’ PBM RNA and linear target DNA.

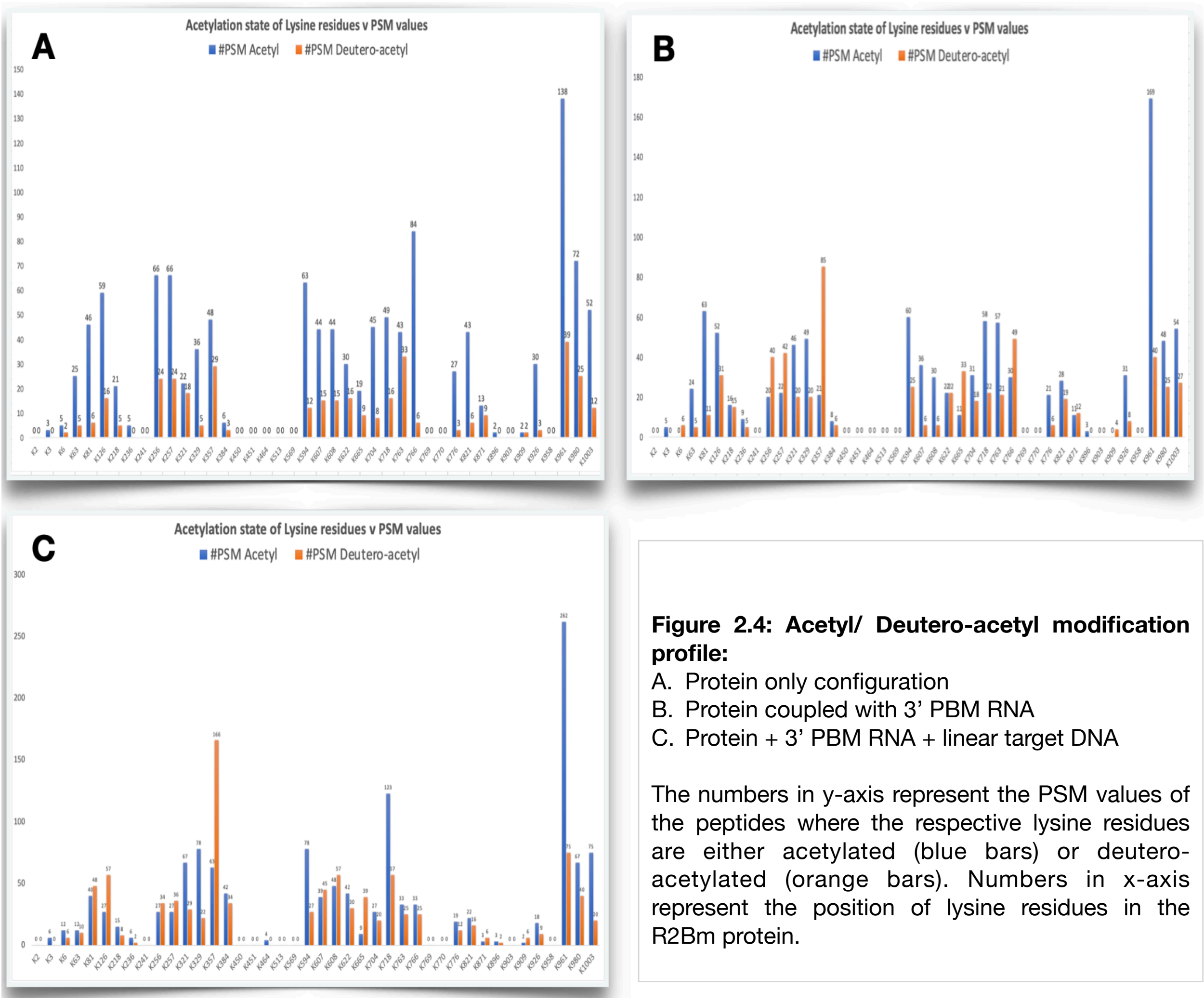

**Figure 2.4: Acetyl/ Deutero-acetyl modification profile:**

A. Protein only configuration
B. Protein coupled with 3’ PBM RNA
C. Protein + 3’ PBM RNA + linear target DNA

The numbers in y-axis represent the PSM values of the peptides where the respective lysine residues are either acetylated (blue bars) or deutero-acetylated (orange bars). Numbers in x-axis represent the position of lysine residues in the R2Bm protein.

Upon two-step acetylation, digestion into peptides and mass spectrometry run, number of peptides were obtained from all three sets. The PSM values of all the peptides of interest (peptides including the lysines) were assigned to the modified lysine residues (either acetylated or deutero-acetylated or both) and the differential pattern of modification was tracked *(Figure 2.4)*. These raw PSM values were taken to calculate the fraction of acetylation and deutero-acetylation for any particular lysine residues, represented by *f*A and *f*D, respectively. And the change in this fraction as we go from one set of reaction to the other is what we are particularly interested in.

As the state changed from “no RNA” to ‘protein with 3’ RNA”, the change in fraction of acetylated and deutero-acetylated lysine residues is noticed quite significantly in most of the lysine residues as represented by the height of bars (blue and orange) *(Figure 2.5 A).* This finding suggests that the RNA binding function is probably widely distributed across the surface of protein rather than just limited to some regions, as hypothesized. However, when the reaction state changes from “protein with 3’ RNA” to “ protein with 3’ RNA and linear target DNA”, the significant change in *f*A and *f*D is observed in handful of residue only *(Figure 2.5 B)*. Six residues in particular, K63, K81, K126, K607, K608 and K896 show high fraction change in duetero-acetylation which is second acetylation.

This implies that upon binding to linear target DNA, the second acetylation was highly dominant for these six residues providing promising evidence that those residues are much more likely to be involved in DNA binding which is why they were protected during first acetylation and exposed for the second.

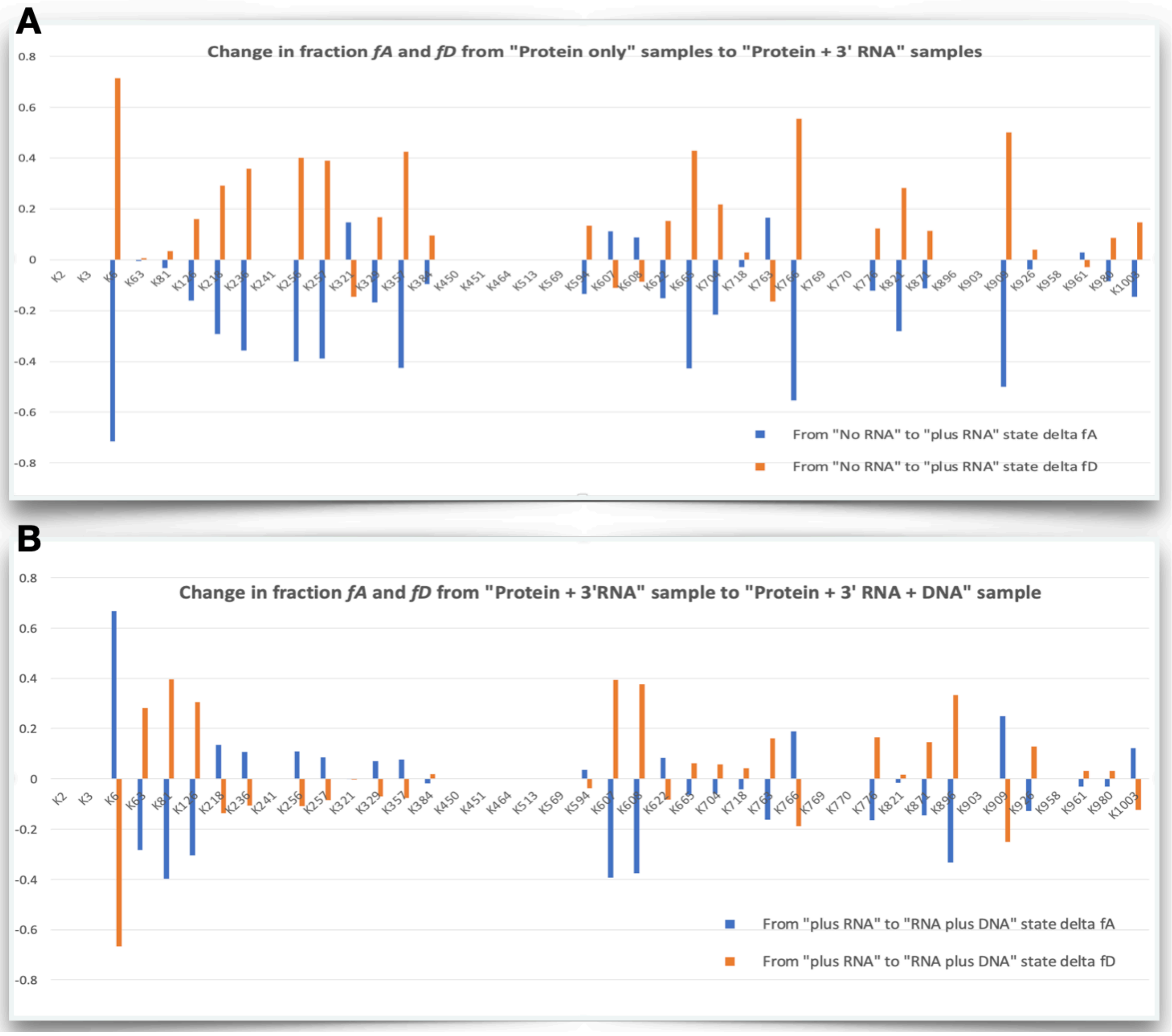

**Figure 2.5: Bar graph representing the change in fraction Acetylated (fA) and fraction Duetero-acetylated (fD) lysine residues:**

A. as the R2 Protein sample goes from “no RNA” state to “protein + 3’PBM RNA” state

B. as the 3’ PBM RNA bound reaction condition goes from “without target DNA” state to “with target DNA” state

Numbers in x-axis represent the position of lysine residues in the R2Bm protein.

[*f*A and *f*D are reciprocal to each other ]

## 2.4. Discussion

Out of the six above mentioned residues with significant changes in fraction of deutero-acetylation, K896 is not taken into consideration because of the very low PSM values of peptides containing that residue which would translate to bigger changes in fraction even with small change in acetylation state. Three lysine residues in particular, K63 and K81 and K126 are actually present in the N-terminal zinc finger (ZF) and myb-DNA binding domain and they are found to be dominated by duetero-acetylation in presence of target DNA, thus providing a confirmation that these residues are involved in DNA binding. This observation is in line with the well established fact that the ZF and Myb motifs are the major motifs used by downstream subunit to bind the target DNA  (21,27). The other two lysine residues, K607 and K608, located in the Reverse Transcriptase domain are the unique residues identified from this set of experiment which might have a role in DNA interaction.

The Mass Spectrometric analysis returned a fairly good data with peptide coverage in and around 75-80%. However, there were some regions with no peptide and some of those non-hits area were significant because they contained the amino acid residue of our interest (lysines). A good chunk of region in Reverse Transcriptase domain and one major area in the linker region just before the Endonuclease domain were seen to be missing. To counter this problem, two alternative strategies can be applied.

Use of different protease(s): Either GluC or mixture of protease (Trypsin + GluC) can be used to digest the protein, instead of the trypsin that we used. In particular, the co-digestion strategy, theoretically should give an excellent proteome coverage such that almost all the peptides of our interest (containing lysines) should be generated (45). R2 protein has 42 Asp residues and 122 Arg residues, providing enough cut-sites for the enzymes to generate the peptides along the entire length of protein.

Arginine modification: There are almost three times the number of Arginine residue than the Lysines in R2 protein, so one strategy would be to modify the Arginines and look for the nucleic-acid binding surfaces. Whole scale chemical modification of Arginine residues can be done by 2, 3-Butanedione. This reagent apparently has the edge over several available reagents because of its relatively higher specificity to Arginine and minimal side reaction, and also the adducts can be stabilized by the use of Borate (39).

## 2.4. Materials and methods

*Protein expression and purification:* To express the proteins, 500 mL expression cultures containing the appropriate R2Bm expression construct were grown in LB broth supplemented with 50 µg/ml kanamycin in an incubator-shaker (37 C, 280 rpm). At an OD 600 of between 0.8-1.0, cells were induced with 0.1mM IPTG and grown for an additional hour at 37 °C. Cells were then harvested by centrifugation at 4000 rpm for 20 minutes at 4° C. The cells were resuspended in 50 ml 10mM Tris pH 7.5 and centrifuged again at 4000 rpm for 10 minutes. The rinsed cell pellets were stored at -80° C. (30). R2Bm protein purification was carried out for wild Type R2 protein as well as endonuclease mutant (KPD/A) as previously published (30). The induced and protein expressed cells pellets were, resuspended, and gently lysed in a HEPES buffer containing lysozyme and triton X-100. The cellular DNA and debris were spun down by ultracentrifugation (33,000 rpm for 20 hours) and the supernatant containing the R2Bm protein was purified over Talon resin (Clontech #635501). The R2Bm protein was then 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). The purified 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.

*Nucleic acid and constructs preparation:* The 120 bp target DNA substrate was generated by PCR using primers GCTCTGAATGTCAACGTGAAGAAATTCAA and TAATCCATTCATGCGCGTCACTAATT that annealed to sites to either side of the R2 target site in plasmid pB108. Various oligonucleotides (oligos) containing specific target DNA (i.e. 28S R2 target DNA) and non-specific DNA (i.e. non-target DNA) were ordered from Sigma-Aldrich. R2 integration intermediates and analogs of presumptive integration intermediates, including the full length target DNA, pre-cleaved DNA target DNA, TPRT product and 4-way junctions were specifically engineered by annealing the oligos. To make each constructs, 20 pmol of one of the component-oligos is annealed with 30 pmol of each of the remaining oligos together. Annealing reaction was carried out in a 1× TPRT buffer (10 mM Tris–HCl (pH 8.0), 5 mM $MgCl_2$, 200 mM NaCl) for 2 min at 95° C, followed by 10 min at 65° C, 10 min at 37° C and finally, 10 min at room temperature. 5′ PBM RNA and 3′ PBM RNA were made by *in vitro* transcription of 1.5 ug PCR products, each derived from appropriate plasmids bearing R2 sequences corresponding to 248 bp of 3' PBM and 319 bp of 5' PBM regions. The DNA substrates were trancribed in vitro at 37°C for 4 hours with T7 RNA polymerase (ThermoFisher # EP0111) in 1mM rNTPs. Template DNA was removed with 10 units of DNase (Promega) for 30 minutes, and the newly synthesized RNA was column purified, ethanol precipitated and resuspended in 1x HEPES TPRT Buffer (10 mM HEPES pH 8.0, 200 mM NaCl, no $MgCl_2$). The RNA was refolded in 5mM MgCls before using in binding reaction.

*Protein-nucleic acid reaction conditions:* 20 pmol of purified R2 protein in protein storage buffer was allowed to bind with 400 pmol of RNA and/or 200 pmol of DNA (linear or any reaction intermediate constructs) in 1x TPRT with $MgCl_2$ buffer for 30 minutes at room temperature. The following combination of protein-nucleic acid complex were prepared; A) R2 protein only (no nucleic acid), B) Protein + 3'PBM RNA, C) Protein + 3'PBM RNA + linear uncleaved target DNA

*Two-stage acetylation:* The RNP complex prepared were incubated in 15mM sulfo-NHS-acetate (Pierce Technology, MW 259.17 g/mol) for 30 min. Acetylation reaction was then quenched by 1/10th volume of chilled 1M Tris pH 7.9, and then mixture subjected to SDS-PAGE gel followed by staining and destaining of gel using Colloidal Blue staining kit (ThermoFisher #LC6025). After the gel run, the bands were excised into eight to ten slices of ~1mm size. Series of washing steps for gel slices included (i) water for 10 min at 37°C, (ii) 10 mM ammonium bicarbonate at room temperature, (iii) 50 mM ammonium bicarbonate:acetonitrile (1:1) twice for 45 min at 37°C, (iv) water twice, and (v) 100% acetonitrile twice at 37°C for 5 min. Second acetylation was performed with D6-Acetic Anhydride (Acros Organics, MW 108.13 g/mol) for 1 hour in presence of Ammonium Bicarbonate buffer pH adjusted to 7-8.

*Protein digestion and MS analysis:* The gel pieces after second acetylation were washed and subjected to reduction and alkyaltion. Disulfide bonds were reduced using 10 mM dithiothreitol in 25 mM ammonium bicarbonate for 30 min at 60°C, then free sulfhydryls were alkylated using 20 mM iodoacetamide in 25 mM ammonium bicarbonate for an hour. The gel slices were then digested overnight using 200 ng Trypsin (Promega #V5111). Peptides generated by digestion were extracted using 50% v/v Acetonitrile/1% v/v Formic acid. Extracted peptides were dried down by vacuum centrifugation, resuspended in 0.1% Formic acid, zip-tip purified, dried again and resuspended in 0.1% Formic Acid prior for mass spectrometric analysis. Peptides were run on Nano LC-ESI-MS/MS and analyzed on Protein Discoverer to identify the location and extent of lysine modification.

# CHAPTER 3

# CONCLUSION

## 3.1. Summary

Global analysis of amino acid residues in the R2Bm protein involved in nucleic acid interaction is possible with mass spectrometric foot-printing analysis. Acetylation of lysine residues in presence and absence of R2Bm interacting nucleic acid can show us a footprint of which lysine residues are protected from acetylation in the presence of nucleic acid. Adding target DNA, 3' PBM RNA, 5' PBM RNA, branched DNA in combinations and separately will lead to the identification of lysine residues involved in interacting with corresponding nucleic acid. These unique residues of interest, in addition to data from other different stages of integration reaction can be collected and analyzed for potential footprints. The residues then can be superimposed in R2Bm model structure to identify the regions/pockets of R2 protein that interact with specific nucleic acid (DNA, RNA) or reaction intermediates, and develop a comprehensive picture of integration mechanism. The findings in this experiment revealed some residues which were previously reported to have been involved in DNA binding, and also brought forward some novel residues which could potentially be involved in the binding.

## 3.2. Future Direction

In the experiment discussed in Chapter 2, only three sets of reaction samples were prepared and subjected to mass spectrometric analysis. The binding activity of 3’ PBM RNA was studied along with linear target DNA and the findings are promising. However, this finding alone is not sufficient to address the outstanding issues related to nucleic acid binding and to present the overall picture of retrotransposition mechanism. To picture the overall retrotransposition mechanism, the binding activity during the different steps of reaction *(from Figure 9)* should be studied and thus different constructs (in addition to the three we used in the experiment) should be prepared and similar modification experiments should be done. These additional constructs that represent the various stages of integration reaction could be:

✦ 1) Protein + 3’PBM RNA + pre-cleaved target DNA

✦ 2) Protein + 5’PBM RNA, 3) Protein + 5’PBM RNA + linear uncleaved target DNA, 4) Protein + 5’ RNA + TPRT analog

✦ 5) Protein + linear uncleaved target DNA 6) Protein + TPRT analog & 7) Protein + 4-way junction DNA

Once we obtain the similar data from all the above mentioned constructs, we can completely map the amino acid residues involved in binding to the various nucleic acid components of the integration reaction. And, the finding thus obtained could also be extrapolated to the study of integration mechanism of other LINEs.

Beside this, following experiments can be performed to support and validate this foot-printing experiment.

### 3.2.1. Protein-DNA crosslinking

Use of crosslinkers to map the interaction of R2Bm protein with linear target DNA cross linking reagents or cross linkers are used to covalently bind two or more protein molecules to facilitate the identification of relationships between near-neighbor proteins, ligand-receptor interactions, three-dimensional protein structures, and molecular associations in cell membranes. Protein crosslinking reagents typically contain two or more chemically reactive groups that will connect themselves to the functional groups (e.g. primary amines, sulfhydryls, carbonyls, carbohydrates and carboxylic acids) found in proteins and other molecules. These reactions make the molecules stable enough to allow for intensive scientific analysis.

Synthetic methodologies for DNA-protein crosslinks formation are based on solid phase synthesis of oligonucleotide strands containing protein-reactive unnatural DNA bases. This approach allows for a wider range of protein substrates to be conjugated to DNA and affords a greater flexibility for the attachment sites within DNA (46)

There are four functional groups; primary amines, carboxyls, sulfhydryls and carbonyls most widely used as cross-linkers. Because the reactivities of most cross linkers favor protein-protein cross linking over Protein-DNA cross linking, this crosslinking of DNA (or RNA) to protein is often difficult (46). To favor the crosslinking between R2 protein and target DNA, DNA probes will be constructed. For this, some select bases across the target DNA are synthesized with primary amines or thiols attached to specific bases. After insertion of the bases into the DNA, one of four groups of reactive cross-linkers can be used to conjugate to proteins. This cross-linking makes the protein-DNA complex stable enough to allow for further procedures.

Two set of reactions; (1) control set with protein only (no crosslink with target DNA) and (2) experimental set with R2 protein cross-linked to target DNA can be carried out simultaneously. Mass spectrometry data of peptides generated from both sets can then be compared for change in mass which equivalents to the mass from cross-linker attached to specific base of target DNA. This mass change, if detected will provide information about which region of R2 protein is involved in binding the target DNA. The result of this experiment coupled with the results from footprinting by residues modification will give a comprehensive insight on nucleic acid binding patterns during the integration reaction.

3.2.2. Mutational analysis of identified residues:

Mutant designing and testing for loss of function will be helpful for further validating the data and for investigating nucleic acid interactions functions. The two lysines identified as unique residues, K607 and K608 can be mutated and tested for the loss of binding functions and likewise for the residues identified by the full set footprinting and also from crosslinking experiment.

# REFERENCES

1. Gogvadze, E. & Buzdin, A. Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66, 3727-3742 (2009).

2. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41, 331-368 (2007).

3. Kazazian, H. H. J. Mobile elements: drivers of genome evolution. Science 303, 1626-1632 (2004).

4. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973-982 (2007).

5. Eickbush, T. H. & Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res (2008).

6. Fujiwara, H. (2015). Site-specific non-LTR retrotransposons. *Mobile DNA III*, 1147-1163.2083-2088 (1998).

7. Kramerov DA, Vassetzky NS (2005). Short retroposons in eukaryotic genomes. *Int Rev Cytol* **247**: 165–221.

8. Ohshima K, Okada N (2005). SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res* **110**: 475–490.

9. 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* 448, 207-213 (2009).

10. 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).

11. Kojima, K. K., & Fujiwara, H. (2005). Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Molecular biology and evolution*, *22*(11), 2157-2165.

12. Martin, S. L. The ORF1 Protein Encoded by LINE-1: Structure and Function During L1 Retrotransposition. *J Biomed Biotechnol* 2006, 45621 (2006).

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

14. 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, 731-736 (2009).

15. Feng, Q., Moran, J.V., Kazazian, H.H., Jr. and Boeke, J.D. (1996) Human L1 retrotransposonencodes a conserved endonuclease required for retrotransposition. Cell, 87, 905-916.

16. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* 87, 917-927 (1996).

17. 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 (2011).

18. 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).

19. 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).

20. Malik, H. S. & Eickbush, T. H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. *Genetics* 154, 193-203 (2000).

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. Burke, W. D., Malik, H. S., Jones, J. P., & Eickbush, T. H. (1999). The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Molecular biology and evolution*, *16*(4), 502-511.

23. Eickbush, T. H. (1992). Transposing without ends: the non-LTR retrotransposable elements. *The new biologist*, *4*(5), 430-440.

24. Chen, J. H., Yajima, R., Chadalavada, D. M., Chase, E., Bevilacqua, P. C., & Golden, B. L. (2010). A 1.9 Å crystal structure of the HDV ribozyme precleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry*, *49*(31), 6508-6518.

25. Webb, C. H. T., Riccitelli, N. J., Ruminski, D. J., & Lupták, A. (2009). Widespread occurrence of self-cleaving ribozymes. *Science*, *326*(5955), 953-953.

26. Eickbush, D. G., Burke, W. D., & Eickbush, T. H. (2013). Evolution of the R2 retrotransposon ribozyme and its self-cleavage site. *PloS one*, *8*(9), e66441.

27. Christensen, S. M., Ye, J., & Eickbush, T. H. (2006). RNA from the 5′ end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proceedings of the National Academy of Sciences*, *103*(47), 17602-17607.

28. Feng, Q., Schumann, G., & Boeke, J. D. (1998). Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proceedings of the National Academy of Sciences*, *95*(5), 2083-2088.

29. 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).

30. 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).

31. Khadgi, B. B., Govindaraju, A., & Christensen, S. M. (2019). Completion of LINE integration involves an open '4-way'branched DNA intermediate. *Nucleic Acids Research*, *47*(16), 8708-8719.

32. 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, e19672 (2011).

33. 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).

34. 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).

35. 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).

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

37. 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).

38. 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

39. 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).

40. 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

41. Mendoza, V. L., & Vachet, R. W. (2009). Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass spectrometry reviews*, *28*(5), 785-815.

42. Chen, T. H., Tanimoto, A., Shkriabai, N., Kvaratskhelia, M., Wysocki, V., & Gopalan, V. (2016). Use of chemical modification and mass spectrometry to identify substrate-contacting sites in proteinaceous RNase P, a tRNA processing enzyme. *Nucleic acids research*, *44*(11), 5344-5355.

43. Hilton, B., Shkriabai, N., Musich, P. R., Kvaratskhelia, M., Shell, S., & Zou, Y. (2014). A new structural insight into XPA–DNA interactions. *Bioscience reports*, *34*(6).

44. MacRae, A. J., Mayerle, M., Hrabeta-Robinson, E., Chalkley, R. J., Guthrie, C., Burlingame, A. L., & Jurica, M. S. (2018). Prp8 positioning of U5 snRNA is linked to 5′ splice site recognition. *Rna*, *24*(6), 769-777.

45. Giansanti, P., Tsiatsiani, L., Low, T. Y., & Heck, A. J. (2016). Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nature protocols*, *11*(5), 993-1006.

46. Tretyakova, N. Y., Groehler IV, A., & Ji, S. (2015). DNA–protein cross-links: formation, structural identities, and biological outcomes. *Accounts of chemical research*, *48*(6), 1631-1644.