TARGETING OF SITE-SPECIFIC NON-LTR RETROTRANSPOSONS: ROLE OF AMINO-TERMINAL DOMAINS

by

HARIDHA SHIVRAM

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN BIOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON
DECEMBER 2011

ACKNOWLEDGEMENTS

I would like to thank the members of my masters thesis committee for their contributions to the work performed during the completion of my masters degree. Dr. Shawn Christensen co-principal investigator, Dr. Ellen Pritham, and Dr. Michael Roner. This work was funded by The University of Texas at Arlington and the National Science Foundation.

I would also like to thank the Christensen lab members- Dr Kimberly Bowles, Dr Brad Reveal, Athena Jagdish and Dillon Cawley.

Finally I wish to thank the University of Texas at Arlington Biology department for its support during my education.

November 11, 2011

ABSTRACT

TARGETING OF SITE-SPECIFIC NON-LTR RETROTRANSPOSONS: ROLE OF AMINO-TERMINAL DOMAINS

Haridha Shivram, M.S

The University of Texas at Arlington, 2011

Supervising Professor: Shawn Christensen

Restriction-like endonuclease (RLE) bearing non-LTR retrotransposons are site-specific elements that integrate into the genome through target primed reverse transcription (TPRT). RLE-bearing elements have been used as a model system for investigating non-LTR retrotransposon integration. R2 elements target a specific site in the 28S rDNA gene. We previously demonstrated that the two major sub-classes of R2 (R2-A and R2-D) target the R2 insertion site in an opposing manner with regard to the pairing of known DNA binding domains and bound sequences—indicating that the A- and D-clades represent independently derived modes of targeting that site. Elements have been discovered that group phylogenetically with R2 but do not target the canonical R2 site. Here we extend our earlier studies to show that a separate R2-A clade element, which targets a site other than the canonical R2 site, does so by using the amino-terminal zinc fingers and Myb motifs. We further extend our targeting studies beyond R2 clade elements by investigating the ability of the amino-terminal zinc fingers from the

nematode NeSL-1 element to target its integration site (This work was done by Dillon Cawley). Our data are consistent with the use of an amino-terminal DNA binding domain as one of the major targeting determinants used by RLE-bearing non-LTR retrotransposons to secure a protein subunit near the insertion site. This amino-terminal DNA binding domain can undergo modifications, allowing the element to target novel sites. The binding orientation of the amino-terminal domain relative to the insertion site is quite variable.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

1.1 Transposable elements and their impact on host genome

Transposable elements(TEs) are mobile genetic elements that have the ability to integrate at a new site in the cell of origin. TEs impact genomes in a myriad of ways. In some plants they occupy ~90% of the genome thereby influencing their genome size. The maize genome is five times larger than the rice genome as a result of ~85% of maize genome being occupied by TEs [1]. Close to 65% of the *Trichomonas vaginalis* genome is made up of DNA transposons [2]. TEs are shown to be involved in the origin of structural variations (SVs). About 50% of the human genome comprises of TEs of which 10% are associated with SVs larger than 100bp [3]. Structural variations have been recognized as a major cause of genomic variation among individuals. These structural variations (SVs) include insertions, deletions, duplications, translocations and inversions. Such structural rearrangements have been implicated to be the cause of many diseases, including cancer [4].

Apart from being responsible for overall genome size and for producing structural variations, TEs can be a source of new coding sequences. There is an ever growing list of examples where transposon derived sequences have been used as new genes or new parts of genes (e.g. exons or gene fusions) (for a fairly recent list see [5] [6]. A couple of notable examples include: duplicated *AMAC* gene from SVA transduction event and, transposase derived RAG1 and RAG2 genes that are involved V(D)J recombination.

Transposons are also a source of conserved noncoding sequences (e.g., regulatory sequences). Indeed, close to 16% of the conserved non-coding elements in eutherian genomes are derived from transposons [5]. Transposons have also been shown to provide promoters for cellular genes (Alu and L1 elements serve as promoter for *FUT5* gene) enhancers (Retrotransposon ERV9 in beta-globin locus region), insulators (B2SINE element in murine growth hormone locus) and transcriptional silencer (Part of *Alu* serves as silencer for human BRCA gene) [7] [8] [9] [10] [6]. TEs can also serve as binding sites for regulatory elements. For example, *Alu* elements serve as binding sites for retinoic acid and thyroid hormone receptors [11].

Transposons are a source of attraction for the host regulatory machinery which may sometimes lead to spreading of heterochromatic region. L1 element for example can trigger an RNAi pathway which can suppress L1 retrotransposition. Apart from containing an internal promoter to drive its transcription, L1 also possesses an antisense promoter which sometimes can make bi-directional transcripts. These bi-directional transcripts can be processed to produce siRNAs which trigger the RNAi pathway [12].

A number of classification systems have been used and adapted based on different factors and with different emphases. Transposable elements have been classified into class I elements that transpose via a DNA intermediate and class II elements that transpose via an RNA intermediate [13] [14]. A more expansive classification system further divides DNA transposons and retrotransposons using additional mechanistic descriptors such as copy-out/paste-in, cut-out/paste-in, and copy-out/copy-in (Figure 1.1). These broad mechanistic descriptors can then be further divided based upon additional specific criteria such as ORF structure, type of transposase encoded, type of endonuclease encoded, or reverse transcriptase phylogeny.

For example, DNA transposons are further subdivided into DDE transposons, Y-transposons, S-transposons and Y2 transposons based on the transposase they contain.

Retrotransposons can be subdivided further based on the enzyme they contain in addition to the reverse transcriptase into- Y-retrotransposons (Contains tyrosine transposase), Long terminal repeat (LTR) retrotransposons (Contains DDE transposase) and non long terminal repeat retrotransposons (Contains an endonuclease) [15].

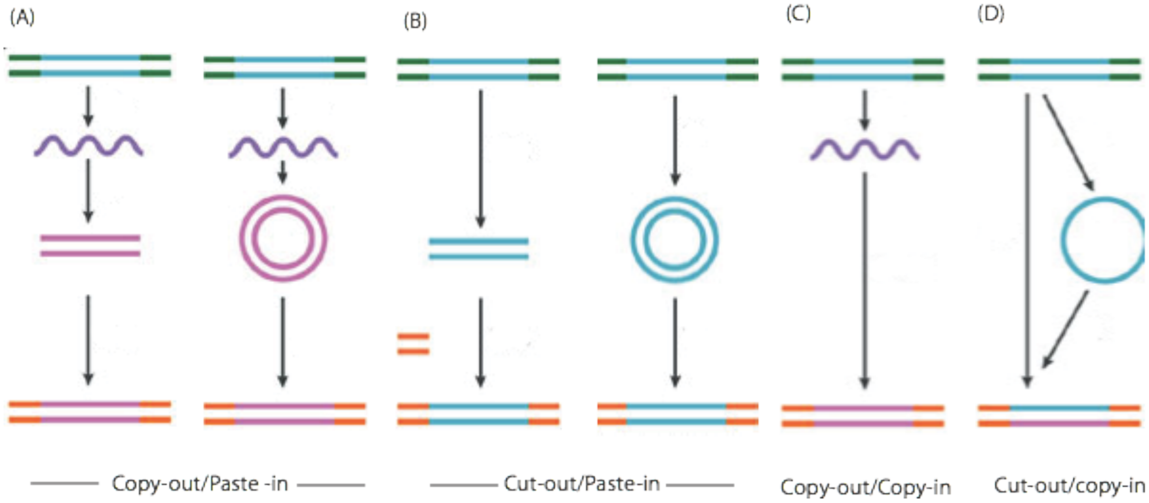

Figure 1.1: Types of transposition mechanism. DNA transposons utilize the cut-out mechanisms while retrotransposons utilize the copy-out mechanisms. (A) This mechanism works via an RNA intermediate which involves reverse transcription of the transcript into linear or circular cDNA. This cDNA then integrates into the target site by a recombination reaction. (B) Here the transposon is excised from the host DNA as a linear or circular DNA and then integrated into the target site through recombination. (C) The element is first transcribed and then integrated into the target site by TPRT. (D) The element first excises out and then pastes one strand at the target site, which it then uses as a template for replication. Figure adapted from [15]

1.2 Retrotransposons

Retrotransposons populate genomes by transposing via an RNA intermediate (copy-out) mechanism. The RNA copy is then either copied into DNA at the new genomic site during reverse transcription (copy-in) or is pasted into the genome as dsDNA post reverse transcription (paste-in). Retrotransposons are subdivided into penelope-like elements, non long terminal repeat (non-LTR) retrotransposons, Tyrosine recombinase (YR) retrotransposon and long terminal repeat (LTR) retrotransposon (Figure 1.2) [16]. LTR retrotransposons use copy-out/paste-in mechanism while Non-LTR retrotransposons use a copy-out/copy-in mechanism called target primed reverse transcription (TPRT) (Figure 1.1) [16]. Penelope elements are also thought to use the TPRT. Non-LTR retrotransposons (and penelope-like elements) are sometimes called target primed (TP)retrotransposons [16]. Reverse transcriptase phylogenies using the seven conserved domains of the reverse transcriptase (common to all retrotransposons) recapitulate the above classifications [13 17].

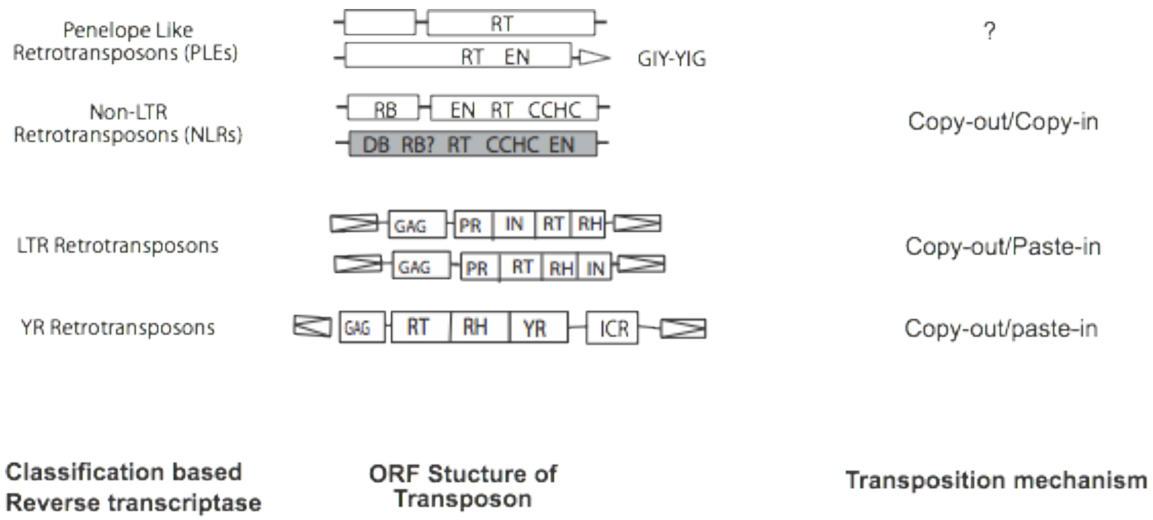

Figure 1.2: Classification and domain structure of retrotransposons. This figure gives the classification of retrotransposons based on the reverse transcriptase domain and the additional enzyme. Each rectangular box represents an open reading frame. The rectangular box with an triangle inside represents terminal repeats. RT- Reverse transcriptase, DB- DNA binding domains, RB- RNA binding domains, EN- endonuclease domain, CCHC- cysteine/histidine motif, RT-Reverse transcriptase, PR- Protease, IN-Integrase, RH- RNaseH, YR- Tyrosine recombinase, ICR- Internal complementary repeat.

1.2.1 Long terminal repeat retrotransposons with DDE transposases : DDE-retrotransposons

LTR retrotransposons or DDE retrotransposons are flanked by long-terminal direct repeats that start with 5'-TG-3' and end with 5'-CA-3' [13]. These elements use copy-out/paste-in mechanism to integrate into a new site which upon integration creates a 4-6 bp target site duplication [13]. LTR retrotransposons and retroviruses share certain amount structural similarity. It is possible that retroviruses evolved from LTR retrotransposons by capturing the *env* protein [18]. Gypsy and ZAM are examples of LTR retrotransposons that contain an *env* gene [19] [20]. However, a retrovirus can also give rise to retrotransposons by losing or inactivating the *env* gene. Such elements that carry an inactivated *env* gene are no longer capable of replicating. These elements are vertically transmitted from one generation to another are called endogenous retrovirus [21]. LTR retrotransposons encodes *gag* which forms the virus-like particle and *pol* that encodes products like protease, reverse transcriptase, RNase H, and integrase (DDE transposase). The RT encoded by the *pol* gene of LTR retrotransposons utilizes the tRNA annealed to the template RNA to initiate the cDNA synthesis which then is bound by the integrase to form the pre-integration complex. The integrase generates a nick at the 3' end of both cDNA strands releasing two terminal nucleotides and exposing 3'-CA. LTR retrotransposons also contains an RNaseH domain which the element uses to get rid of the RNA as the cDNA is being synthesized [22] [23]. Though *gag* and *pol* genes are conserved among LTRs, the level of expression of these genes differ. Some hosts experience higher level of expression of *gag* than *pol* genes (*copia from drosophila melanogaster* and *Ty5 f*rom *Saccharomyces cerevisiae)* which they require for productive VLP (virus-like protein) formation and replication. The expression of *gag-po*l is regulated by mechanisms like ribosomal frameshifting and differential protein degradation[24]. There are also cases where non-coding sequences make a major part of the element. For example, LARD element in barley is comprised of 3.5 kb of non-coding sequence [25].

LTR retrotransposons can be grouped into three major subclasses based on reverse transcriptase phylogeny- Ty3/Gypsy, Ty1/Copia and Bel/Pao elements.

1.2.1.1 Ty1/Copia LTR retrotransposons

According to the international committee on the taxonomy of virus (ICTV), Ty1/Copia group has been classified into *Metaviridae* and *Errantiviridae* genera [26]. Ty1/copia elements are the least abundant group of LTR retrotransposons. The general domain structure of these elements consists of long terminal repeats flanking the open reading frame encoding *gag* and *pol* proteins. The pol gene consists of integrase, RT and the RNaseH domain. The integrase is located on the amino-terminal of RT and the RNaseH domain. Based on the RT phylogeny, Ty1/Copia elements are the oldest lineage of LTR retrotransposons [16]. An example of a Ty1/copia element that underwent an *env* gene acquisition is SIRE1. Elements related to SIRE1 have now been found in numerous plant species [16].

1.2.1.2 Ty3/Gypsy LTR retrotransposons

Based on the classification by ICTV, Ty3/Gypsy group have been classified into three genera- *Pseudovirus, Hemivirus* and *Sirevirus* [26]. This group is the most abundant group of LTR retrotransposons and are widely distributed in the genomes of plants, animals and fungi [26]. Ty3/Gypsy elements contain the integrase domain on the carboxy- terminal of the RT and RNaseH. *Gypsy* element found in the genome of *Drosophila melanogaster* is the best-studied example of *env* gene acquisition. The *gypsy* element because of the presence of *env* gene has shown to be infections [16]. These elements do not replicate using the RT priming by tRNA mechanism used by the other LTR retrotransposons [16].

1.2.1.3 Bel/Pao LTR retrotransposons

This group of LTR retrotransposons have been found only in metazoan genomes [27] [26]. The abundance of these elements falls in between Ty1/Copia and Ty3/Gypsy. According to the classification by ICTV, these elements belong to the genera *Semotivirus* [26]. These elements have the integrase domain on the carboxy-terminal of the RT, as in case of Ty3/Gypsy elements [16].

1.2.2 Tyrosine recombinase containing retrotransposons: YR-retrotransposons

Tyrosine recombinases are a type of site-specific recombinase that catalyze recombination between two specific DNA sequences via the formation of holliday junction. These recombinases use tyrosine residue as the nucleophile to cleave the phosphodiester linkage [28]. YR-retrotransposons (PAT, kangaroo) or DIRS1 are elements that encode a tyrosine recombinase (Y-transposase) in addition to the reverse transcriptase. YR-retrotransposons are placed under LTR retrotransposon if the classification is based on the RT domain. They use the copy-out/paste-in mechanism for their transposition. Excised circular cDNA generated by reverse transcription are integrated into the genome by the Y-transposase. These retrotransposons either contain inverted long terminal repeats (DIRS1) or split direct repeats [15]. It is likely that these repeats play a role in the reverse transcription of an RNA transcript into a DNA intermediate and also serve as recombination sites. They also consist of a region called Internal complementary repeats (ICR), that are thought to play a role in the replication cycle [29].

1.2.3 Penelope-like retrotransposons

Penelope-like retrotransposons (PLE) are a very diverse group of retrotransposons with an RT domain that shows a high sequence divergence compared to other retrotransposons. Their RT domain is more closely related to telomerase in comparison to LTR or non-LTR elements.

Penelope elements are thought to transpose using TPRT [16]. The endonuclease domain of these elements show sequence similarity to Uri endonucleases of bacterial group I introns and UvrC bacterial DNA repair endonucleases [16 30]. These elements are found to occupy sites in or near the telomeric region. This element causes hybrid dysgenesis in various *drosophila* species [31].

1.2.4 Non-LTR retrotransposons

As the name suggests, these elements are not flanked by direct or inverted repeats. Instead, these elements contain a poly(A) tail on their 3' end and truncations on their 5' end [32]. Non-LTR retrotransposons (NLRs) are sometimes referred to as target-primed (TP) retrotransposons. Both NLRs and group II introns integrate into a new site using the mechanism called target primed reverse transcription [33]. The Mobile group II introns are retroelements that consist of catalytic RNA that carries out RNA splicing and reverse splicing to integrate into a new site.

NLRs can be grouped into 28 clades based on the phylogenetic analyses of the RT domain(Figure 1.2) [34]. Sequence alignments were done using 11 blocks from the RT region- 8 from the catalytic fingers/ palm subdomain and 3 from the thumb subdomain. NLRs contain one or two ORFs that encode an endonuclease domain, Ribonuclease H and/or nucleic acid binding domains. Based on the endonuclease the element contains, NLRs can be divided into Apurinic/Apyrimidinic endonuclease (APE) containing elements and restriction like endonuclease (RLE) containing elements (Figure 1.2). RLE bearing NLRs appear to be the earlier branching of the two groups, while APE containing elements are appear to be more recent [35]. An exception to this classification is an element called DUALEN (part of the RandI clade) which was found to contain both APE and RLE [35].

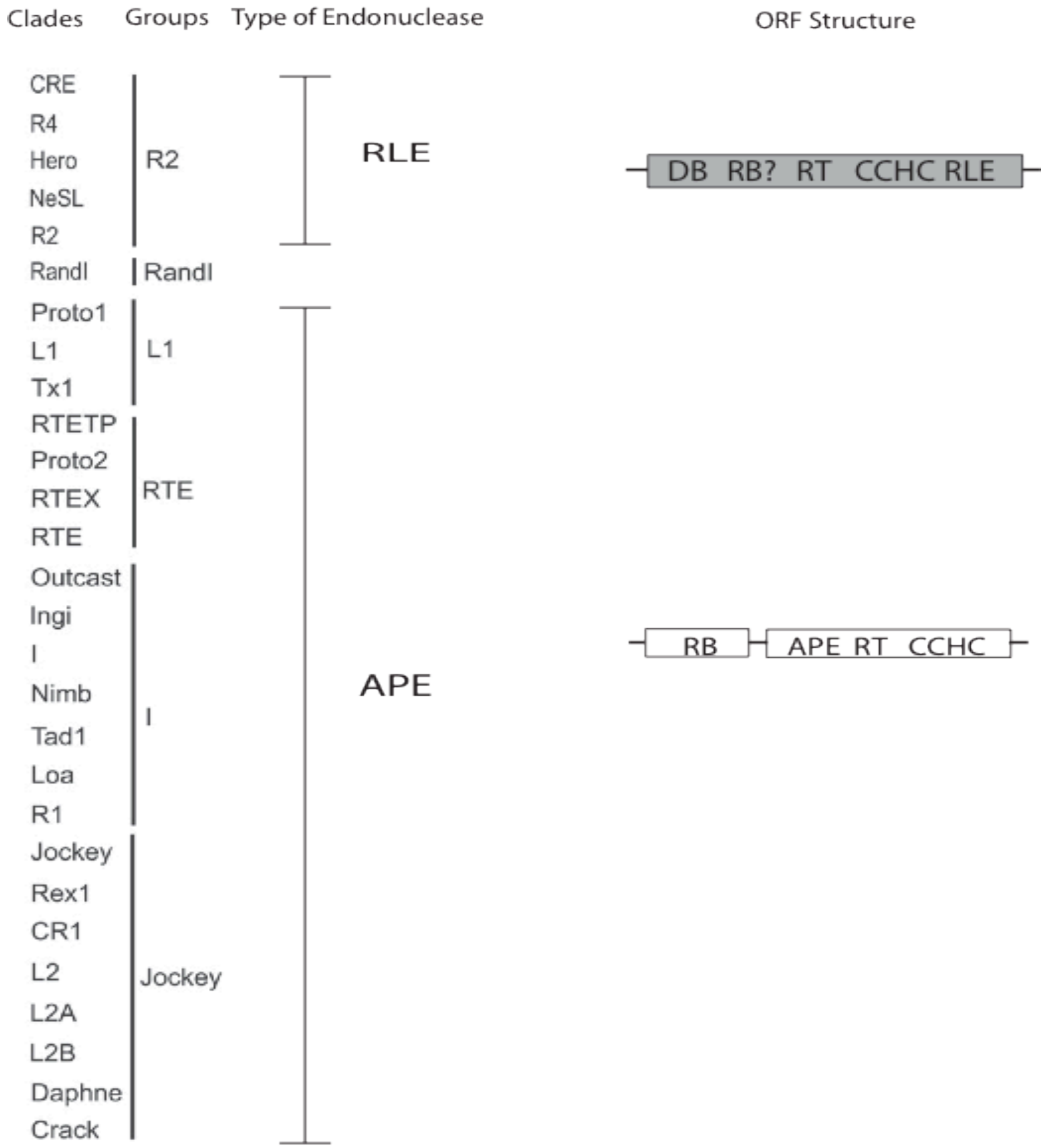

Figure 1.3: Classification of Non-LTR retrotransposons. Different clades of RLE Non LTR retrotransposons based on RT phylogeny are shown with their further classification into different groups. These groups are further grouped together based on the type of endonuclease they contain. Figure adapted from [34].

1.2.4.1 Apurinic/Apyrimidinic endonuclease containing non LTR retrotransposons

Twenty-two out of the twenty-eight clades of NLRs are Apurinic/Apyrimidinic endonuclease (APE) bearing NLRs, which are further grouped into the L1, Jockey, RTE and I goupings. APE bearing NLRs typically encode two ORFs. The first ORF consists of RNA binding domains.

LINE1 elements are the most extensively studied members of APE bearing NLRs. Human L1 consists of 5' untranslated region (5' UTR), two ORFs separated by intergenic spacer region, followed by 3' UTR and then the poly (A) tail. [36] ORF1 of LINE1 contains a coiled-coil domain (C-C) on its N-termina end, a RNA recognition motif (RRM), and a C-terminal domain (CTD). The RRM and CTD domains are necessary for the LINE1 element to strongly bind to nucleic acids. The C-C domain helps in the trimerization of the protein and also to organize the relative orientations of RRM and CTD domains within the trimer [37] [38]. LINE1 lineages are very diverse in the coiled coil region. Human L1 coiled coil region are similar to leucine zipper while mouse L1 coiled coil region is similar to keratin. The ORF protein of L1 is similar to retroviral nucleocapsid proteins in that they have nucleic acid chaperone activity. Nucleic acid chaperones facilitates the formation of most thermodynamically stable form of nucleic acids via annealing and melting. L1s use this chaperone activity during the process of replication. [39] [40] [41]

The second ORF contains the APE domain, RT domain, and a CCHC domain with the endonuclease domain preceding it [16]. The APE and RT are used during the cleavage reactions and reverse transcription steps of the retrotransposition cycle, respectively. Although the role of CCHC domain is not known yet, they have been shown to be critical for retrotransposition. ORF2p is translated less number of times than ORF1p and therefore hard to detect in culture assays [42].

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). Despite cis preference, the L1 machinery can transpose other RNAs to varying degrees (e.g., SINE elements and cellular mRNAs). The enzymatic machinery of L1 is used by *alu* and SVA elements for their transposition. [38 39 43]

1.2.4.2 Restriction-like endonuclease containing NLRs

Most RLE bearing NLRs are site specific elements that target repetitive sequences with in their host genome. RLE containing elements have an endonuclease that is similar to type IIs restriction enzymes [44]. RLE bearing NLRs have been classified into five clades CRE, R2, R4, NeSL and Hero clades. NLRs bearing restriction-like endonuclease usually contain one ORF, though there are exceptions (R5 has two ORFs) [34 13]. Most restriction-like endonuclease (RLE) containing NLRs have zinc fingers located in their N-terminal region followed by the RT domain, a CCHC motif, and the RLE domain. Most of the biochemical information for the mechanism of replication of non LTR retrotransposon comes from the R2 element in *Bombyx mori* and will be discussed below.

1.2.4.3 R2: RLE bearing non LTR retrotransposon

R2 elements that insert into the 28S rDNA were originally found in *drosophila melanogaster* and *bombyx mori.* Since then, they have been found in numerous other species belonging to phylum arthropoda and deuterostome. R2 elements are vertically transmitted and their copy number is tightly restricted in the 28S rDNA.. The amino terminal end of the R2 ORF encodes for a variable number of DNA binding motifs (Zinc fingers and myb motifs) [45]. Based on the number of N-Terminal zinc finger motifs and phylogenetic analysis of the RT domain, R2 elements can be divided into four subclades- R2-A, R2-B, R2-C and R2-D. R2-A elements contain three zinc fingers, R2-C elements contain two, R2-B elements contain either 2 or 3 and R2-D elements contain one zinc finger motif [46]. Followed by a set of DNA binding domains on its amino-terminal end, R2s have a centrally located RT domain. On their carboxy-terminal end is located a CCHC domain and an endonuclease. R2 element found in *Bombyx mori* (R2Bm) was used as a model to understand the integration mechanism.

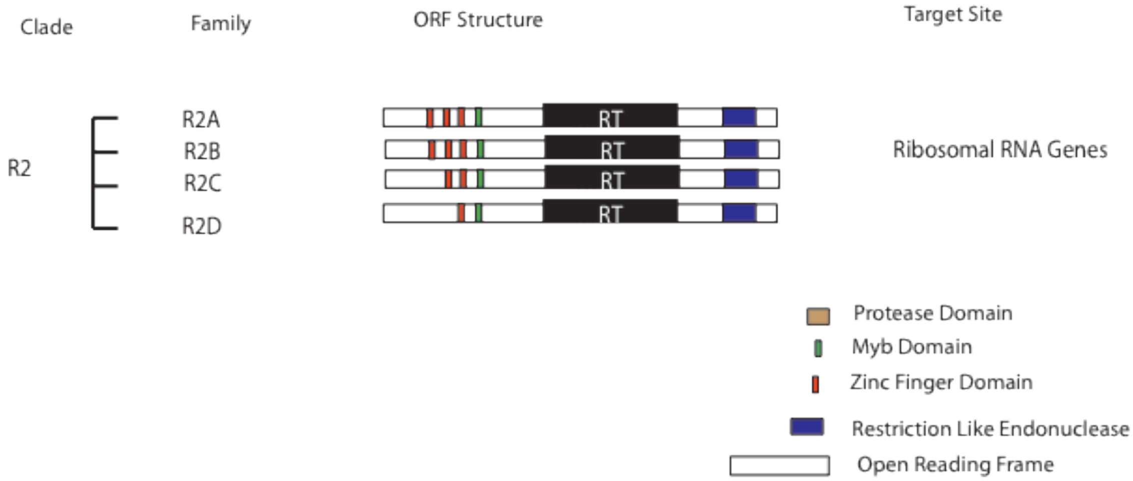

Figure 1.4: Classification of R2 elements. R2 elements can be subdivided into four families- R2A, R2B, R2C and R2D based on the number of zinc finger domains on the N-terminal end. Different families with their domain structures are shown.

1.2.5. Ribosomal RNA genes: Safe haven for R2 elements

Ribosomal genes are highly conserved and and are required in great abundance in order to generate the copious amounts of rRNA needed by cells. The rDNA genes are arranged as a cluster of hundreds of tandem units. Each unit consists of a gene that codes for the small ribosomal subunit (18S) and a gene that codes for the large ribosomal subunit (5.8S and 28S). Usually more rDNA units are encoded than are actually needed for survival [47] [48]. One interesting feature of the ribosomal locus is its ability to eliminate variation in the number of ribosomal units. This process of eliminating variation is called concerted evolution.

Variation in the number of ribosomal units among individuals can be attributed to unequal homologous recombinations between different rDNA units within the loci on different chromosomes [48]. Sister chromatid exchange (SCE) is the most common mode of recombination in rDNA locus. This kind of unequal crossing over results in one locus carrying more ribosomal units than the other. The locus with very few ribosomal units are selected against while the locus with a very large number of ribosomal units are subjected to loop deletion (Figure 1.3 B) [49].

Ribosomal units in many groups of animals contain insertions of TEs. Ribosomal units that are interrupted by TEs are rendered non-functional, suggesting that the ribosomal locus would eliminate TE insertions through negative selection. Despite this, ribosomal locus is populated by many TEs. R2 elements are found in the 28S ribosomal sequence of numerous arthropod and deuterostome species. Even though a number of 28SrDNA lose their activity due to TE insertions, the host is not affected since they make more number of ribosomal copies than are actually required. The number of insertions in a locus is a balance between the deletion rate and retrotransposition rate. Recombinations within the ribosomal genes tend to remove old dysfunctional copies of retrotransposons and at the same time providing new sites for insertions [49 50]. The R2 levels in the 28SrDNA thus continually fluctuate between low and high levels. [47]

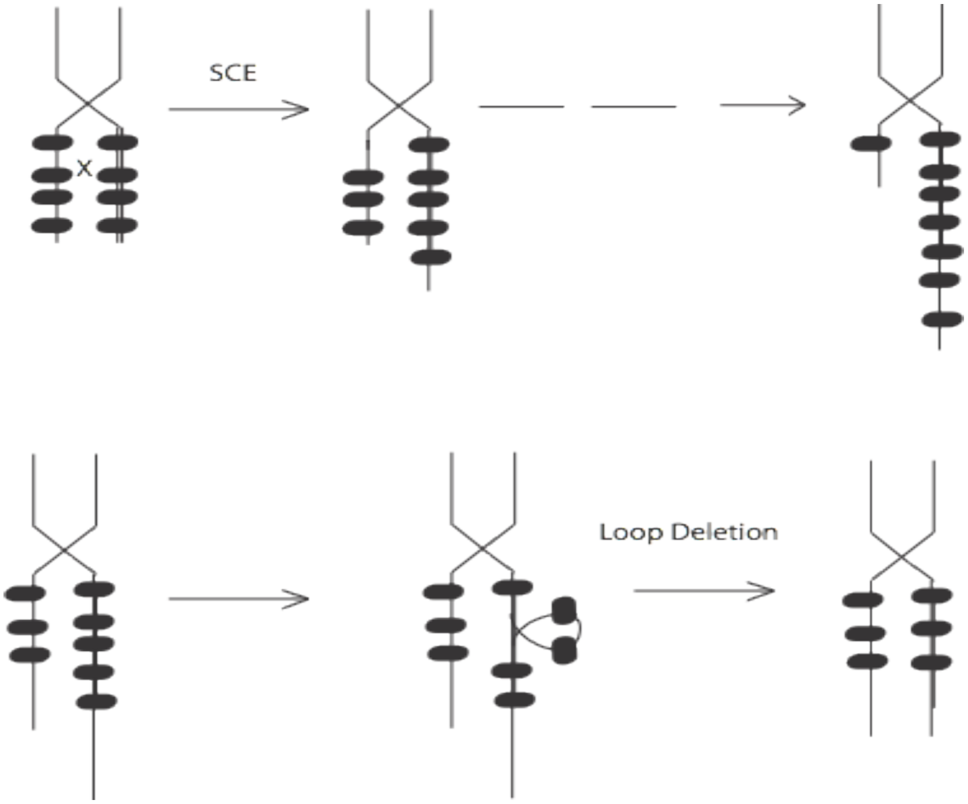

Figure 1.5: Diagram of the major types of recombinations that occur in the ribosomal locus. The chromosomes are represented by vertical lines and each black oval represents a ribosomal gene. (Top)Sister chromatid exchange (SCE) is a kind of recombination reaction where there is an unequal exchange of ribosomal units between sister chromatids. This may result in one locus larger in size than the other. (Bottom) The second type of recombination that are found in the ribosomal locus is the looping out of parts of the chromosome which results in deletion of few ribosomal units. Figure adapted from [49]

1.2.6. R2 integration mechanism: Target primed reverse transcription

Target-primed reverse transcription (TPRT) is the mechanism used by non-LTR retrotransposons where the 3‘ OH group liberated after cleavage by the endonuclease is used to prime reverse transcription of the element RNA. This reaction occurs via the formation of a ribonucleoprotein complex (RNP) where the encoded protein(s) associate with their encoding mRNAs [44]. The integration of R2 is believed to involve the formation of an RNP complex consisting two of protein subunits, of different conformations, bound to a single RNA. One protein subunit is bound to a conserved RNA structure of 5’ UTR called the 5' protein binding motif (PBM) and the other protein subunit is bound to a structure in the 3' UTR called the 3’ PBM. The RNP complex conducts element integration in a series of steps discussed in the following sections (esp., sections 1.2.6.3- 1.2.6.6). The model is based on the R2 retrotransposon from *bombyx mori* (R2Bm) with some data from additional arthropod species including *Drosophila*.

1.2.6.1 Co-transcription and processing of R2 from 28SrRNA

An important step in the retrotransposition cycle is to generate the RNA transcript. In case of LTR retrotransposons, the repeats carry the sequence required to start and stop transcription. Non-LTR retrotransposons, in most cases appear to carry internal promoters while in other cases, they rely on their co-transcription with an active transcription unit using RNA polymerase I. The 5‘UTR of R2 encodes a Hepatitis delta virus-like (HDV) ribozyme which enables the element to separate itself from rest of the 28SrDNA (processing). This keeps the element independent of all the events associated with ribosome assembly and regulation. Both R2 ribozyme and HDV ribozyme contain a double pseudoknot and five base-paired region and also share sequence similarity in the region surrounding the catalytic core. This pseudoknot is conserved across many R2 species and is located downstream of uncapped 5’ end and multiple stop codons. Most mRNAs that are processed from RNA polymerase I lack 5’ methyl cap that causes them to be unstable and lose their ability to react with the translation machinery. However, in case of R2, even though they lack the 5’ methyl cap, 5’ end of R2 generated by the ribozyme tends to form a stable structure that is resistant to exonuclease degradation [51] [52].

1.2.6.2 Translation of R2 transcript

The R2 ORF begins at the end of the ribozyme structure and this boundary region is characterized by the presence of multiple in-frame stop codons [52]. There is no conserved Met initiation codon downstream from the stop codons. Since the co-transcribed R2 RNA lacks the 5‘ methyl cap and also considering there is no amino acid conservation upstream of the presumptive ORF, translation initiation would need a non canonical 5‘ methyl cap independent translation mechanism [53] [52]. Most other systems like viral RNAs that lack the 5‘ methyl cap possess a secondary structure called internal ribosomal entry site (IRES) [51]. A common feature between these IRES containing systems and R2 RNA is the presence of a region of complementarity between 5‘ UTR and 18S subunit. This region of complementarity serves as the ribosome landing pad during translation initiation [53]. For the reasons stated above, an internal ribosome initiation mechanism has been proposed that likely involves one of the conserved RNA secondary structures observed in the 5' UTR. The 5' PBM is more likely to be involved in an internal ribosome initiation signal than the HDV-like ribozyme due to the presence of a pseudoknot structure in the 5' PBM's and also its proximity to the start of the presumptive ORF [52].

1.2.6.3 RNP formation

R2 RNA consists of conserved regions with in the 5' UTR (the HDV-like rybozyme and the 5' PBM), sequence coding for the ORF, and the 3' protein binding motif (3‘PBM) that lies with in the 3‘UTR. The 5‘PBM consists of four hairpin secondary structures and one pseudoknot while the RNA secondary structure of 3‘PBM region is highly variable and is clade specific [52]. As their name implies the 5' PBM and 3' PBM regions play important roles in RNP formation and ultimately in the integration mechanism. Integration competent RNPs are thought to contain two protein subunits and a single RNA molecule. One protein subunit binds to the 5' PBM [54] and one to the 3' PBM [44]. Upon binding 5' PBM and 3' PBM, each subunit adopts a distinct conformation and is destined to perform different roles in the integration reaction (see below). There does not appear to be any protein-protein interactions between the two subunits. Binding

RNA confers protein conformation and integration functions as well as timing functions that will be discussed in the following sections.

1.2.6.4 Step 1 of the integration reaction: Endonuclease from the upstream subunit cleaves the bottom strand

The protein subunit bound to the 3'PBM binds upstream of the insertion site using an unidentified protein motif to do so. It is hypothesized that this DNA motif might lie with in the carboxyl-terminal domain of R2, perhaps the three helix bundle, the CCHC, or even the endonuclease itself. The R2 protein makes a nick on the DNA, liberating the 3'OH which is used as a primer for reverse transcription of the R2 RNA by the reverse transcriptase [55].

1.2.6.5 Step 2 of the integration reaction: target primed reverse transcription

Once the endonuclease from the upstream subunit cleaves the bottom strand, the upstream R2 protein undergoes a conformational change such that the endonuclease moves away and places the reverse transcriptase close to the nick [44]. The reverse transcriptase uses the 3' OH from the cleaved DNA to prime the reverse transcription reaction which requires the presence of 3'PBM [56]. Interestingly, the cleavage reaction can take place in presence of any RNA while the reverse transcription reaction specifically requires the presence of 3'PBM [45 57].

1.2.6.6 Step 3 of the integration reaction: The endonuclease from the downstream subunit cleaves the top strand

The protein subunit complexed to the 5' PBM binds downstream of the insertion site using R2 protein's amino terminal ZF and myb motifs to do so [54]. This protein subunit cleaves the top DNA strand. In order to cleave the top DNA strand, the endonuclease domain of the subunit must be in the opposite orientation to that of the upstream subunit. For this reason, it is hypothesized that the upstream and downstream R2 subunits are bound to the DNA in opposite orientation to each other (see also section 1.2.6.6) [44]. Interestingly, the second strand cleavage reaction requires the 5' PBM to be removed from the downstream subunit [56]. It is hypothesized that the removal of the 5' PMB occurs during reverse transcription of that section of the RNA, pulled out by the reverse transcriptase [23].

1.2.6.7 Step 4 of the integration reaction: Synthesis of the second strand of DNA

The downstream subunit is hypothesized to perform second strand synthesis. This step has not yet been experimentally shown but evidences support the possibility of this reaction. Here the downstream subunit is bound in such a way that the RT faces away from the cleavage site (see section 1.2.6.5). The RT has both RNA-templated DNA polymerase activity (i.e., the TPRT required activity) and also DNA-templated DNA polymerase activity, the activity needed to perform second strand synthesis [23]. In addition, R2 has a demonstrated an efficient strand displacement activity that would be required during second strand synthesis to displace the RNA strand off the RNA:DNA heteroduplex generated during the initial reverse transcription step (section 1.2.6.3)

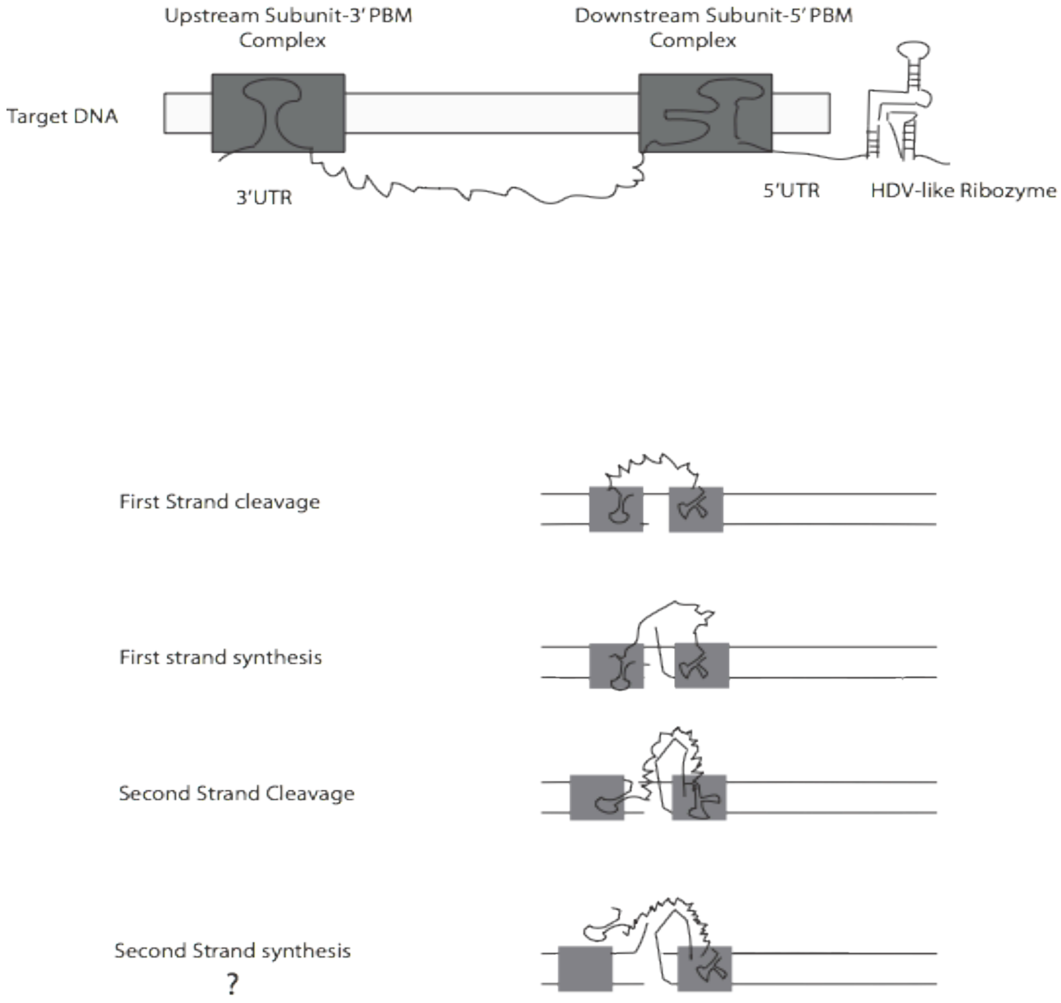

Figure 1.6: R2 Integration model. Here grey boxes represent R2 subunits and the two horizontal lines are the top and bottom strands of target DNA. Two R2 subunits are utilized in the integration reaction. When the R2 protein binds to the 3' PBM it adopts the upstream binding conformation such that the amino-terminal domains get sequestered and exposes the second DNA binding domains. R2 adopts the downstream binding conformation when it binds to the 5'PBM. In case of the downstream subunit, the carboxy terminal end is sequestered exposing the amino-terminal domains (Top). The integration reaction takes place in four steps. In the first step, the upstream binding subunit provides the endonuclease to generate the first strand nick which is then used by the reverse transcriptase from the upstream subunit to prime the first strand synthesis in second step. The third step involves the use of endonuclease from the downstream subunit to generate the second strand cleavage. Finally, it is hypothesized that R2 uses the reverse transcriptase from the downstream subunit to perform second strand synthesis (Bottom). Figure adapted from [44] and [51]

1.2.6.8 Role of Amino terminal DNA binding motifs in targeting insertion events

As mentioned earlier, the amino terminal end of R2 elements are highly variable with the number of zinc finger domains ranging from 1-3 and some containing a different DNA binding domain called myb. ZFDs are small functional folded domains that in its stable form occur in coordination with one or more zinc ions. These proteins usually function as transcription factors where they recognize and bind certain DNA sequences using side chain-base interactions. Phosphate backbone interactions between ZF and DNA adds to the binding affinity. Cysteine-histidine, CCHC zinc finger domains are the most abundant and these make contact with DNA using alpha helix found between the second cysteine and first histidine. [58]

Myb domains consist of ~ 50 amino acid long repeats which forms a three -helix bundle structure. The second and the third helix exists as a helix turn helix conformation. The third helix serves as the recognition helix that makes contact with the DNA in the major groove. [59] Most DNA binding proteins require a minimum of two ZF and two myb domains to make a tight contact with DNA by wrapping around the DNA structure. However, R2-D clade elements exemplified by R2Bm uses one myb and one ZF to bind region of DNA that are about 10 bp away. Instead of wrapping around DNA, the R2 protein gets access into the major groove crossing the minor groove. It is possible that the R2 protein tracks along the major groove making loose contacts with the top strand. The R2-A group binds the target site differently than the R2-D group. The R2-A group uses the amino-terminal ZFs and a Myb to bind to the upstream DNA sequences as opposed to downstream sequences. The different DNA binding modes indicate R2-A and R2-D are independent targeting (or re-targeting) events to the canonical R2 site. The distinct DNA binding modes also suggest that the R2-A and R2-D elements use a different linkage configuration between the various nucleic acid binding domains and catalytic activities involved in the integration reaction. [45 60]

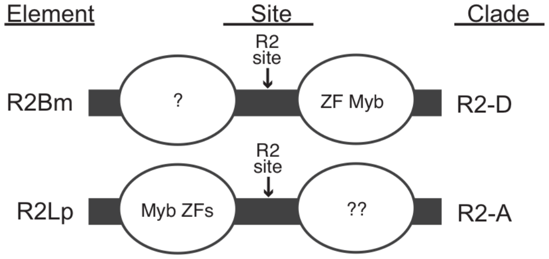

Figure 1.7: DNA targeting configurations of the DNA Binding domains. The target DNA is represented by the black rectangular box where the ovals overlaying on top are the retrotransposons. R2Bm protein targets downstream of the insertion sequence with the amino terminal DNA binding domains such that the zinc fingers binds closer to the insertion site. However, in case of R2Lp amino terminal domains bind upstream of the insertion site.

1.2.7 R9: An R2A clade element targeting a novel site

R9Av elements are NLRs belonging to R2-A clade found in *Adineta vaga*, a bdelloid rotifer. R9 elements insert into the 28s rRNA gene 1436 bp upstream of where other R2 elements insert. R9 has a similar structural organization to other R2-A clade elements but is characterized by two distinct features: site-specific insertion into a previously unreported target sequence within the 28S gene, and an unusually long target site duplication of 126 bp. It is formed of a single open reading frame coding for a 1102-aa polypeptide, which contains all of the domains expected for a typical R2 element. Four copies of R9 were found and each one of them had the same length of TSD, indicating a very high precision during the second strand cleavage reaction.

I am particularly interested in finding out how R9 being an R2 element targets a new site. We hypothesize that the amino-terminal DNA binding domains have changed their site specificities and now target the new site. One of the major goals of our lab is to understand the targeting mechanism of R2 elements. Once the targeting mechanism is understood, we hope to engineer R2 elements to be used as gene targeting vectors. This would enable us to target these transposons to desired sites and hence be used as gene-targeting vectors.

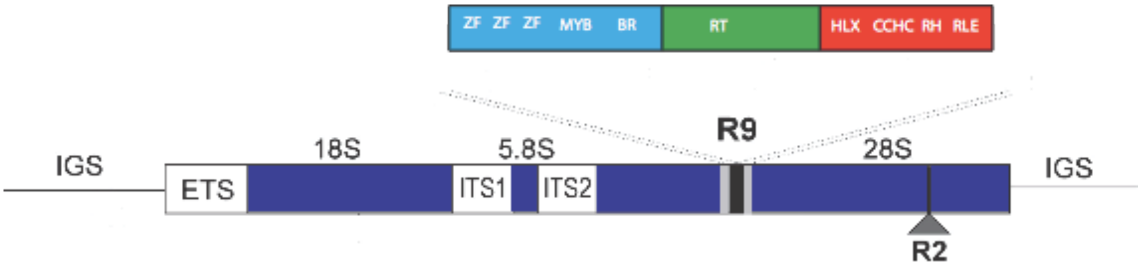

Figure 1.8: Ribosomal unit from A.vaga bearing the R9 insertion. The blue rectangular box is the A.vaga ribosomal unit divided into 18S and 28S units. ZF- Zinc finger domain, BR- basic region, RT- reverse transcriptase domain, RLE- Restriction like endonuclease, ETS- External transcribed spacer, ITS1 and ITS2- Internal transcribed spacers. The vertical black box within the 28S box represents the R9 insertion and the grey region on either side of it are the target site duplications. The canonical R2 insertion site is shown by a grey triangle. The domain structure of R9 element (Top) consists of amino terminal end (Light blue) bearing three ZF and one myb followed by basic region, centrally located RT domain (Green) and carboxy terminal end (Red) containing the endonuclease. Figure adapted from [61]

CHAPTER 2

TARGETING NOVEL SITES: THE AMINO-TERMINAL DNA BINDING DOMAIN OF NON-LTR RETROTRANSPOSONS IS AN ADAPTABLE MODULE THAT IS IMPLICATED IN CHANGING SITE SPECIFICITIES

2.1 Introduction

Non-long-terminal-repeat (non-LTR) retrotransposons are a major class of eukaryotic transposable elements. These elements are vertically inherited and can impact the evolution of their host's genome in many ways. [32 62 63 64 65] Non-LTR retrotransposons replicate through an ordered series of DNA cleavage and polymerization events using encoded nucleic acid binding, endonuclease, and polymerase functions.[44] The element encoded protein(s), once translated, form a ribonucleoprotein (RNP) particle with the transcript from which they were translated—a process called cis-preference. The RNP binds to the target DNA, cuts one of the DNA strands, and uses the target site's exposed 3'-OH to prime reverse transcription of the element RNA into complementary DNA (cDNA)—a process called target primed reverse transcription (TPRT) [55]. The opposing target DNA strand is then cleaved [44]. The cDNA is turned into double stranded DNA, completing the integration event in a process that is not yet well understood [44].

Non-LTR retrotransposons can be grouped into those elements that harbor a restriction-like endonuclease (RLE) to initiate TPRT and those that harbor an apurinic-apyrimidinic endonuclease (APE) to initiate TPRT [44 66 67 68]. The RLE-bearing elements tend to be site specific (i.e., inserting into a given sequence within a genome) while the APE bearing elements tend to be nonspecific—although examples of both site-specific and nonspecific targeting can be found within each group [32 66 69 70]. Phylogenetic analysis of the reverse transcriptase domain from RLE and APE bearing elements indicates that the RLE-bearing elements are likely the earlier branching group [16].

The site specificity of RLE-bearing non-LTR retrotransposons make them attractive systems with which to study the TPRT integration reaction. In addition, once it is understood how RLE-bearing elements target DNA, it may be possible to engineer these elements for use as site specific gene targeting vehicles. We are conducting a systematic study of the DNA targeting functions of RLE-bearing non-LTR retrotransposons [44 60 45 56]. DNA recognition is the first step in the integration reaction and is therefore the step that must undergo modification when an element targets a novel site. Target site recognition is believed to be achieved primarily through distinct DNA binding motifs located within the coding region of the element, as opposed to any inherent specificity of the RLE. The endonuclease is believed to be largely nonspecific while the DNA binding domains target insertion events to specific sites in the genome [44 67 70 54 71].

RLE-bearing non-LTR retrotransposons encode a single multifunctional protein with RNA binding, DNA binding, DNA endonuclease, and reverse transcriptase activities. These retrotransposons have been phylogenetically classified into at least five clades based upon sequence comparisons of the reverse transcriptase domain: R2, R4, Genie, CRE, and NeSL (Fig. 1A) [67 72 73 17 74]. All of these clades share a similar basic ORF structure with a central reverse transcriptase domain (RT), a carboxyl-terminal cysteine-histidine motif (cchc), and a carboxyl-terminal restriction-like endonuclease. The major differences between the clades resides within the amino-terminal domain. Elements belonging to the Genie, CRE, and NeSL clades typically contain two amino-terminal zinc fingers (ZFs), while R2 clade elements contain a Myb motif and a variable number of ZFs. The R4 clade appears to lack both ZFs and Myb motifs. Within each clade amino-terminal structural variants exist (e.g., a variable number of ZFs). In addition, the namesake of the NeSL clade, NeSL-1, contains a cysteine protease domain (PRO) of unknown function [72].

R2 is the most well studied of the five clades. The R2 designation is actually a double criteria designation—a phylogenetic grouping in conjunction with an insertion site designation. R2 elements insert into a particular sequence within the 28S rDNA gene (see Figs 2.1A and 2.2) [32] [67] [46]. R2 elements are further subdivided into the R2-A, R2-B, R2-C, and R2-D groups based on reverse transcriptase phylogeny [46]. Each R2 group has a different configuration of ZFs and Myb motifs in the amino-terminal domain of the encoded R2 protein. The two major subdivisions are the R2-A and the R2-D groups (Fig 2.1). The R2-D group, exemplified by the well studied *Bombyx mori* R2 element (R2Bm), has a single amino-terminal ZF and a Myb motif. R2Bm uses two protein subunits to integrate [44]. These two subunits take on different conformations and roles in the integration reaction (Fig 2.1B) [44] [75]. One protein subunit is bound upstream of the insertion site, and the other one is bound downstream of the insertion site. The upstream subunit binds to the DNA using an undetermined DNA binding motif, cleaves the target DNA, and performs TPRT [44,54]. The downstream subunit uses the amino-terminal ZF and Myb motifs to bind to the DNA [54]. The downstream subunit cleaves the second DNA strand and is hypothesized to perform second strand synthesis. The 3' untranslated region (UTR) of the R2 RNA is bound to the upstream subunit, and the 5' UTR is bound to the downstream subunit [75]. The R2-A group binds the target site differently than the R2-D group. The R2-A group elements have three amino-terminal zinc fingers as well as a Myb motif and are thought to be the more ancestral R2 group. The R2-A group element from *Limulus polyphemus* (Lp), R2Lp, uses the amino-terminal ZFs and a Myb to bind to the upstream DNA sequences as opposed to downstream sequences [60]. The different DNA binding modes indicate R2-A and R2-D are independent targeting (or retargeting) events to the canonical R2 site. The distinct DNA binding modes also suggest that the R2-A and R2-D elements use a different linkage configuration between the various nucleic acid binding domains and catalytic activities involved in the integration reaction [60].

Recently elements that phylogenetically group with the R2-A group have been discovered that do not target the canonical R2 site. R9 from Adineta vaga (R9Av) targets a site within the 28S about 1436 bp upstream of the R2 site, and R8 from Hydra magnipapillata (R8Hm) targets a sequence in the 18S rDNA (Fig 2.2) [76] [61]. The R2 site is thought to be the more ancestral site as most species that have retained R2 clade elements have done so at the R2 site. Given this interpretation, the R8 and R9 sites are instances of R2 clade elements acquiring novel site specificity [76]. In this paper we investigate the role that the amino-terminal ZF and Myb motifs have in targeting an R2 clade element to a novel genomic site. We show that the amino-terminal domain of R9Av has been modified so as to target the R9 site and not the R2 site. Interestingly, the orientation of the binding of the ZFs and Myb motif differs from the R2-A and the R2-D elements that target the R2 site. We also extended our DNA targeting studies to elements beyond the R2 clade. Of the five clades—R2, R4, Genie, CRE, and NeSL—only R2 has a Myb motif in the amino-terminal region (Fig 2.1A). That Myb motif is a major contributor to the specificity observed in the R2 clade elements, with the ZFs providing fewer DNA contacts as a whole than the Myb motif [60] [45]. Genie, CRE, and NeSL clades only have ZFs—typically two ZFs—and no Myb motifs. In order to ascertain how Genie, CRE, and NeSL clade elements target DNA, we examined the DNA binding potential a NeSL clade element. We show that the NeSL-1 element uses its two amino-terminal ZFs to target DNA. Along with our previously reported study, our results indicated that the amino-terminal ZFs (and Myb motif if present) may represent a universal targeting module for all site specific RLE-bearing non-LTR retrotransposons that contain these motifs. The Myb and ZFs can undergo modification, allowing novel sites to be targeted. During modification, individual ZF and Myb motifs can be acquired or lost. In addition, the physical/temporal linkage configurations between the various nucleic acid binding activities (5' UTR RNA binding, 3' UTR RNA binding, upstream DNA binding, and downstream DNA binding) and catalytic activities (first strand cleavage, TPRT, second strand cleavage, and second strand synthesis) may get reconfigured as elements transition to target new sites in the genome.

2.2 Materials and methods

2.2.1. Generating expression constructs.

Constructs containing R9 and NeSL-1 derived amino-terminal putative DNA binding motifs were generated and named as follows: R9Av ZF3-Myb corresponds to codons 54 -295 of the extended ORF (stop codon to stop codon) from R9 *Adineta vaga* transposon (R9Av) genbank GQ398057.1, R9Av ZF1-Myb corresponds to codons 154 - 295 of R9Av, and NeSL-1 ZFs corresponds to codons 110 - 261 of the *Caenorhabditis elegans* NeSL-1transposon extended ORF (stop codon to stop codon). The polymerase chain reaction (PCR) primers used to amplify the above regions of interest from *A. vaga* and *C. elegans* genomic DNA are listed in Table 1. *A. vaga* genomic DNA was a gift from Irina Arkhipova (Josephine Bay Paul Center for Comparative Molecular Biology and Evolution). *C. elegans* genomic DNA was a gift from Andre Pires d Silva (University of Texas Arlington). The R9 fragments were cloned into the Gateway® donor vector pENTR/TEV/D-TOPO (Invitrogen K2535-20) using the manufacturer's protocol and then recombined into the Gateway®-compatible bacterial-expression destination vector pDESTTAP [60]. The NeSL-1 ZFs were cloned into the NdeI and BamHI sites of the bacterial expression vector pET28a (Novagen 69864-3). Initial ligation and recombination reactions were transformed into electroporation competent XL-1 Blue *Escherichia coli* (Agilent 200259) or chemically competent Oneshot Top10 E. coli (Invitrogen C4040-10) for screening purposes. Resulting colonies were screened by PCR and sequenced (Big Dye, Applied Biosystems 4337455). The expression constructs were maintained in Arctic Express RIL DE3 E. coli cells (Stratagene 230193) for expression.

2.2.2. Protein expression and purification.

Cells were grown in 200 mL of Luria Bertani medium to an A_{600} of 0.6-0.7 at 37ºC in an incubator shaker. The culture was then cooled to 12ºC. Isopropyl-beta- D-Galactoside (IPTG) was added to the cooled culture at a final concentration of 1 mM for the R9Av clones and 0.1 mM for the NeSL-1 clone. After the addition of IPTG, the cultures were further incubated for 24 hours at 12ºC in an incubator shaker. The cultures were centrifuged at 4,000x g for 20 minutes at 4°C. The pellets were washed with cold 10mM Tris-HCl pH 7.5 and were either used directly or stored at -80ºC.

The R9Av pellets were resuspended in 2.5 mL of Solution A (50% glycerol, 100 mM HEPES, 5 mM beta-mercaptoethanol, and 2 mg/mL of lysozyme (Amresco 0663)) and incubated on ice for 15 minutes and at room temperature for 15 minutes. The resuspended cells were lysed by adding 13.2 ml of solution B (100 mM HEPES, 1 M NaCl, 5 mM beta-mercaptoethanol, and 0.1% Triton X-100) and incubating on ice for 30 minutes. The resuspended pellet was then centrifuged for 20 hours at 69,888x g at 2ºC. The supernatant was mixed with the Talon resin (Clontech 635501) that had been prewashed with 10 ml of Talon column buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 0.2% triton X-100, 5 mM imidazole pH 7.5). The resin bound protein was washed with 20 mL of column buffer containing 10 mM imidazole and eluted with 300 ul of column buffer containing 300 mM NaCl and 150 mM imidazole. An equal volume 100% glycerol was added to the eluate for storage at -20ºC.

The NeSL-1 pellets were resuspended in 8 mL of lysis buffer (100 mM Hepes pH 7.5, 1 M NaCl, 1 mM beta-mercaptoethanol and 0.2% triton X-100, 10 units of DNase I) and passed through a French press two times. The cell lysates were centrifuged in an Eppendorf centrifuge at 12,000 RPM for 10 minutes at 4ºC. The supernatant was adjusted to 10% glycerol and passed over the Talon resin columns by gravity flow, as previously described[60]. Dillon Cawley did the work on NeSL.

Protein concentrations were determined by samples run on a SDS 6% polyacrylamide gel electrophoresis (PAGE) along with a bovine serum albumin standard. SDS PAGE were stained with Sypro Orange (Biorad 170-3120) or Comassie Blue R-250 (Amresco 0472-10G) and the band intensities were measured using imageJ 1.38X software [77]. The apparent purities were 80% or greater.

2.2.3. Electrophoretic mobility shift assays and DNA footprints.

The 5′ ^{32}P end labeled DNA substrates were generated (Table 1) and purified as previously described [44]. The binding reactions in Figure 3 were performed in 13 uL reactions: 10 uL of a solution containing 7.5 mM Tris-HCl (pH7.5), 50 mM NaCl, 2.75 mM $MgCl_2$, 0.5 mM $CaCl_2$, 0.8 mM ditriothritol, 9 ng of target DNA, 50 ng of poly dI-dC (Sigma Aldrich P4929-5UN); and 3 uL of protein diluted to an appropriate concentration in protein storage buffer (see above). The binding reactions were incubated at 25ºC for 20 minutes and then loaded on a 1X TBE (89mM Tris base, 89mM boric acid, 2 mM EDTA) native 5% polyacrylamide gel. The gels were run at 230 V for 30 minutes. Gels were dried and visualized on a phosphorimager screen.

The binding reactions for DNase I footprints were similar but lacked poly dI-dC and were performed under conditions that gave approximately 40%-60% bound species. 0.012 units of DNase I were used. Binding reactions treated with DNase were fractionated on native polyacrylamide to isolate the bound, free, and reference fractions. The bound and reference fractions were analyzed on a denaturing 6% polyacrylamide gel. The reference DNA fraction was from reactions that did not contain transposon protein. Missing nucleoside footprints were as previously described except for the use of the binding conditions noted above for DNase I footprints [45 78]

2.3 Results and discussion

2.3.1 DNA binding activity of an R2 element directed to a non-canonical target site: Mapping the DNA binding activity of R9Av

R2-D group elements that target the R2 site (e.g., R2Bm) use the amino-terminal ZF and Myb motif to secure an R2 protein subunit downstream of the insertion site [44] [45]. R2-A group elements that target the R2 site (e.g., R2Lp) use their amino-terminal ZFs and Myb motif to secure an R2 protein subunit upstream of the insertion site [60]. Although both R2-D and R2-A group elements target the same site, they do so differently. To better understand R2-clade mechanistic plasticity and site specificity, R9Av was examined. R9Av is an R2-A group element that does not target the canonical insertion site (see Fig 2.2). R9Av is also interesting as it generates a 126 bp target site duplication (TSD) upon insertion as opposed to the blunt or small 1-10 bp deletion observed for most R2 insertions [79].

Full-length non-LTR retrotransposon proteins are very difficult to express and purify in a soluble and active form. For this reason, amino-terminal derived polypeptides containing the ZFs and Myb motifs expected to be involved in targeting were expressed and purified in order to ascertain if the R9Av ZF and Myb motifs have modified so as to direct the R2 clade element to the R9 site instead of the R2. Initially, four target DNAs were used in electrophoretic mobility shift assays (EMSA) to identify if and where the R9Av amino-terminal ZF and Myb bind (Fig 2.3). The R9Av polypeptide spanned from ZF3 through the Myb motif and was given the name R9Av ZF3-Myb (see Fig 2.2). A diagram of the four target site DNAs used in the EMSA reactions are shown in Figure 2.3A. Target 1 consisted of the segment of the 28S that becomes duplicated upon R9 insertion (126 bp TSD region) along with 112 bp of upstream flanking sequence (net 238 bp). Target 2 was the 126 bp TSD region along with 101 bp of downstream flanking sequence (net 227bp). Target 3 was the 112 bp of upstream flanking sequence. Target 4 was the 101 bp of downstream flanking sequence. Each target DNA was end-labeled with

^{32}P and put into binding reactions with varying amounts of R9Av ZF3-Myb polypeptide. The polypeptide concentrations ranged from 240 nM to 9 nM in three-fold increments. The binding reactions contained 9 ng of DNA, which translates to around 9.6 nM for Targets 1 and 2, and around 21 nM for Targets 3 and 4.

The R9Av ZF3-Myb polypeptide bound best to Target 1, with observable binding occurring in the presence of 9 nM of polypeptide—roughly a 1:1 molar ratio of protein to DNA (Fig 2.3B, Lane 5). Greater than 50% of Target 1 was bound in the presence of 27 nM polypeptide (Fig 2.3B, Lane 4), and 100% of target bound at 80 nM—a 8.3:1 protein to DNA molar ratio (Fig 2.3B, Lane 3). At the higher protein to DNA ratios, additional protein-DNA complexes occur (Fig 2.3B, Lanes 2-4). Presumably the fastest migrating protein-DNA complex represents a single polypeptide bound to DNA (hereafter called a monomer) as it appears in the lower protein concentration samples. The slower migrating protein-DNA complexes, then, may represent higher order protein-DNA complexes (e.g., dimer, trimer, etc). Target 3 was also bound efficiently by the R9Av ZF3-Myb polypeptide in that protein-DNA complexes were observed at low protein concentrations (e.g., Lane 5 of Fig 2.3D, a 1:2.3 molar ratio of protein to DNA). Target 2 was bound much less efficiently (at least 9X less efficient) than Target 1. Observable R9Av ZF3-Myb polypeptide binding to Target 2 did not occur until the 80 nM of polypeptide level—a 8.3:1 protein to DNA molar ratio (Fig 2.3C, Lane 3). The R9Av ZF3-Myb polypeptide did not bind appreciably to Target 4 (Fig 2.3E).

The EMSA reactions involving the R9Av ZF3-Myb polypeptide indicate that this region of the R9 protein is involved in DNA targeting. Because Target 1 and Target 3, which have the upstream DNA in common, were bound most efficiently by the R9Av ZF3-Myb polypeptide (respectively), it appears that the protein motifs contained in the polypeptide are responsible for securing an R9 protein subunit to DNA sequences upstream of the TSD region. However, as Target 2, which contains the TSD region, bound the R9Av ZF3-Myb polypeptide better than Target 4, it is possible that some base-specific interactions may occur within the TSD region in addition to the upstream sequence. Any presumed association of the polypeptide with the TSD region could not be due to the action of either ZF3 or ZF2 as the binding profile of the R9Av ZF1-Myb polypeptide mirrored the binding of the R9Av ZF3-Myb polypeptide on the four targets used in Figure 3 (data not shown). Binding to Target 1 was the most robust, followed by Target 3, then Target 2, and lastly Target 4. It is possible that the greater association of protein with Target 2 relative to Target 4 is related to local DNA structure (e.g., bent DNA for positioning a nucleosome) causing increased association of the polypeptide.

To better map the site of interaction between the R9Av ZF3-Myb polypeptide and the target DNA, a DNase I protection based DNA footprint analysis was performed. Assuming the binding of the R9Av ZF3-Myb polypeptide to target DNA is specific to the upstream sequence and perhaps the TSD region, as indicated by the EMSA results, it should be possible to precisely map the site (or sites) of interaction by DNase footprinting. A footprint signal is indicative of specific binding. So as not to unduly skew the mapping results, the footprint analysis was done using Target 1.

The R9Av ZF3-Myb polypeptide was bound to target DNA that had been end-labeled on either the top strand or the bottom strand, respectively. The R9Av ZF3-Myb polypeptide:DNA ratios were adjusted so as to form primarily monomers or monomers and dimers similar to that seen in Figure 2.3B, Lanes 4 and 5. The binding reactions were subsequently subjected to DNase I treatment under conditions that yield one cleavage event per DNA target. The DNase I treated protein-DNA complexes were fractionated into bound and free/reference fractions by EMSA (data not shown) prior to analysis on a denaturing polyacrylamide gel (Fig 2.4A). Segments of DNA that show protection from DNase I cleavage when compared to reference DNA indicate areas of DNA bound by the R9Av ZF3-Myb polypeptide. Of the bound fractions, Lanes 3 are from the dimer and Lanes 4 are from the monomer protein complex. The areas of DNase I protection denoting R9Av ZF3-Myb binding have been marked with long thick black lines next to the denaturing gel. Short thin black lines denote DNase I hypersensitive sites that were induced by the binding of the R9Av ZF3-Myb polypeptide.

The footprinted region for a single polypeptide monomer is restricted to sequences upstream of the TSD region, specifically from base pair position -47 to -27 on the top strand and from base pair position -46 to -27 on the bottom strand (Fig 2.4A, Lanes 4). The base pair numbering is relative to the presumptive site of bottom strand cleavage and TPRT. The TPRT site is inferred from the orientation of inserted R9Av elements as well as the DNA cleavage sites required to generate the observed TSD. Negative numbers represent base pair positions upstream of the site of TPRT, and positive numbers represent base pairs positions downstream of the TPRT site. No obvious additional footprint signal was observed in any higher order complexes (e.g., the dimer complex, Lanes 3) beyond that defined for the monomer (Lanes 4). There is a DNase I hypersensitive site on the bottom strand at base pair position -46 induced by binding of a single polypeptide unit. Additional DNase I hypersensitive sites are observed in the presence of additional polypeptide units being bound (see top strand Lane 3, positions +16 and +74).

In order to ascertain the binding orientation of the R9 Myb relative to the ZFs and to gain a higher resolution footprint, a missing nucleoside footprint analysis was performed on the R9Av ZF3-Myb polypeptide [78]. A second missing nucleoside footprint was performed on the truncated polypeptide, R9Av ZF1-Myb, which is missing ZF3 and ZF2. DNA with random abasic sites (i.e., missing nucleoside DNA) was exposed to protein in a binding reaction followed by EMSA fractionation. The resulting DNA fractions were analyzed on a denaturing polyacrylamide gel. Bands corresponding to a DNA base that, when missing, interfered with protein binding yielded a footprint signal. The missing nucleoside data for the R9Av ZF3-Myb polypeptide bound to DNA are presented in Figure 2.4B. Only the fastest migrating protein-DNA complex (i.e., the monomer) was analyzed. The DNA bases that were found to interact with the R9Av ZF3-Myb polypeptide are located from -43 to -31 on the top strand and from -46 to -31 on the bottom strand, in good agreement with DNase I footprint data. The missing nucleoside data for the R9Av ZF1-Myb are presented in Figure 2.4C.

The missing nucleoside data have been summarized for both polypeptides, along with the DNase I data, in Figure 2.4D. Both DNase I and missing nucleoside footprint data confirm that the R9Av amino-terminal DNA binding region, ZF3-Myb, binds to DNA sequences upstream of the 126 bp TSD region. If a specific interaction exists between the ZF3-Myb polypeptide and the TSD region, the interaction is not stable enough to footprint in our reactions. In addition, no additional footprint signal was detected in any of the higher order complexes, indicating a single specific binding site for the polypeptide on the target DNA is likely. The shorter R9 polypeptide, R9 ZF1-Myb, made contacts on both strands in the region of -40 to -31, but just on the top strand in the region spanning -43 to -41. The lack of footprinting on the bottom strand in the region of -46 to -40 with the shorter polypeptide indicated that either, or both, ZF3 and ZF2 associate with the bottom strand in this region. The region from -40 to -31, which footprinted on both strands with both polypeptides, is likely where the Myb motif binds. ZF1 binding may then account for the top strand signal from -43 to -41. This interpretation is consistent with other R2 elements where the Myb motif contacted both DNA strands over roughly a 10 bp region and ZF1 contacted at least one strand over a 3-5 bp region [45,60]. Additional studies would have to be done to confirm the ZF assignments in R9Av; however, it is clear that the Myb motif binds closest to the insertion site and the ZFs farther away.

R9Av and R2Lp are both R2-A group elements that use their amino-terminal ZFs and Myb to target different sites in the genome. R9Av targets the R9 site, and R2Lp targets the R2 site. Both elements use their respective amino-terminal DNA binding module to bind a protein subunit upstream of the insertion site [60]. Interestingly, R9Av and R2Lp bind in opposite orientation to each other with respect to the binding order of the Myb motif and ZFs relative to the insertion site. In R2Lp, the ZFs are closest to the insertion site, while in R9Av the Myb motif binds closest to the insertion site. In the R2-D clade element R2Bm, the Myb motif and ZF bind downstream of the insertion site, with the ZF being closest to the insertion site [45].

2.3.2 DNA targeting by a non-R2 RLE-bearing non-LTR retrotransposon

For both R2-A and R2-D group elements, the Myb domain appears to account for the largest continuous swath of base specific contacts for the subunit using the amino-terminal motifs to bind to DNA [45,60]. Indeed, we have been unsuccessful in getting the ZFs of R2-A and R2-D group transposons to bind tightly enough to be footprinted in the absence of the Myb motif [45,60]. Of the known RLE-bearing non-LTR retrotransposons, only R2 clade elements contain a Myb motif. In this aspect at least, R2 is not representative of other RLE-bearing non-LTR retrotransposons. Amino-terminal ZFs (typically two ZFs) have been identified in elements belonging to the NeSL, CRE, and Genie clades (Fig 2.1A). In order to extend our studies of target site recognition beyond R2, the NeSL clade element NeSL-1Ce was examined. NeSL-1Ce contains two amino-terminal ZFs and targets the spliced leader-1 gene of *Caenorhabditis elegans* (Fig 2.5A). In order to test if the NeSL-1Ce ZFs function in target recognition similar to R2's Myb plus ZF(s) pairing, a polypeptide containing the NeSL-1Ce amino-terminal ZFs (NeSL-1 ZFs) was cloned, expressed, and purified. The purified polypeptide was assayed for DNA binding activity against a 125 bp target DNA containing the NeSL-1 insertion site. EMSA analysis showed a slower migrating complex consistent with the NeSL-1 ZFs polypeptide binding to target DNA (Fig 2.5B). DNase I footprint analysis (Fig 2.5C) was used to determine if the binding was specific. Regions of DNA protected from DNase I degradation in the presence of bound polypeptide were localized to two closely spaced regions: (1) top strand base pair positions -21 to -19, bottom strand -20 to -17; (2) top strand -9 to -7, bottom strand -7 to -5. Base pair positions are relative to the NeSL insertion site (i.e., TPRT site), with negative integers representing base pair positions upstream of the TPRT site and positive integers representing base pair positions downstream of the TPRT site. The DNase I footprint has been overlaid on the insertion site sequence in Figure 2.5D. The NeSL-1 ZF polypeptide was found to bind to DNA sequences upstream of the insertion site (i.e., within the spliced leader exon). The two zones of DNase I protection observed likely correspond to the binding of the two ZFs, respectively. The binding orientation of the two ZFs is unknown. Dillon Cawley did the work on NeSL.

2.3.3 Summary of RLE-bearing non-LTR retrotransposon DNA binding modes and implications for the integration model

The insertion model posited for R2 elements, and by analogy all RLE-bearing non-LTR retrotransposons, requires two subunits of protein to affect element insertion, one subunit bound to each side of the insertion site. The catalytic domains of the two subunits must be in opposite orientation to each other in order to carry out the two half reactions required for insertion (see Fig 2.1B for more information derived from R2Bm). Amino acid alignments of R2 elements would appear to argue that there is tight integration between the RT and the carboxyl-terminal domain. There is not much room for flexible linkers between the highly conserved domains in the carboxyl terminal domain, unlike the amino-terminal domain where there is variability in the number and makeup of the conserved motifs as well as variable spacing between some of the conserved regions [67]. Examinations of 5' junctions of native R2 insertions in various Drosophila species indicate that the processes of second strand cleavage and second strand synthesis are rapidly evolving [79]. In R2Bm, where most of the biochemistry has been done, the subunit that performs second strand cleavage (and presumably second strand synthesis) interacts with the 5′ RNA and binds to the target DNA downstream of the insertion site using the amino-terminal ZF and Myb (Fig 2.2B, Fig 2.6). The upstream subunit binds through an (as yet) unidentified protein domain (indicated by the single question mark in Fig 2.6). In R2Bm, the subunit that uses the unidentified DNA binding domain binds the 3′ RNA and performs first strand cleavage and TPRT. The unidentified DNA binding domain has been hypothesized to be located in the carboxyl-terminal domain [44 67].

Collectively, our amino-terminal DNA binding data on R2 and NeSL elements indicate that most site-specific RLE-bearing non-LTR retrotransposons likely use their amino-terminally located DNA binding motifs to load a transposon subunit onto the target site. There appears to be variability in how the ZFs and Myb motifs bind target DNA (Fig 2.6). In some cases the amino-terminal motifs are used to secure the upstream subunit and in other cases the downstream subunit. In some cases the ZFs are closest to the insertion site, and in other cases the Myb is closest to the insertion site. This variability may indicate plasticity in how the binding and catalytic domain functions are wired into the overall insertion mechanism. The orientation changes may relate to the absolute orientation of the RLE and RT catalytic domains relative to the insertion site, although flexible linkers could either decouple binding orientation from catalytic orientation or even re-wire the linkage to be opposite of what is known for R2Bm.

In R2Lp the upstream and downstream subunits appear to be swapped relative to R2Bm [60]. However, orientation of the ZFs and Myb motif relative to the insertion site are identical to R2Bm (i.e., the ZFs are nearest the insertion site) [60]. The R2Lp downstream subunit is hypothetical and is based upon the R2Bm model of insertion. The downstream R2Lp subunit is marked with two question marks in Figure 2.6, one question mark to signify the presence of the subunit being hypothetical and one question mark to signify that, if present, the subunit would be expected to bind to DNA using the same unidentified (carboxyl-terminal?) protein domain that secures the R2Bm upstream subunit to target DNA. In the case of R9Av, the amino-terminal ZFs and Myb bind upstream of the insertion site as in R2Lp, but the binding orientation of the ZFs and Myb appear flipped compared to R2Lp (Fig 2.6). The two R9Av subunits would be expected to be near each other in space, assuming the 146 bp TSD region was wrapped around a nucleosome and the downstream subunit bound just downstream of the TSD [61]. As in R2Lp, the R9Av downstream subunit would be targeted via the unidentified DNA binding domain. In the case of NeSL-1, the upstream subunit is again bound using the amino-terminal ZFs (orientation unknown), and the hypothetical downstream subunit would be bound to target DNA using the unidentified DNA binding domain. In each case, it is tempting to speculate that the subunit that binds the 3′ UTR RNA binds to DNA using the hypothetical carboxyl-terminal DNA binding domain (or the RLE) and performs TPRT as the RT and conserved carboxyl-terminal motifs appear to be more tightly linked. The variable amino-terminal domain is attached to the RT through a variable length spacer [67]. The subunit that performs the rapidly evolving second strand cleavage and second strand synthesis reactions would be bound to target DNA using the amino-terminal ZFs (and Myb). If this model continues to hold, it may be possible to engineer an RLE-bearing element (e.g., R2Bm) to target elsewhere in the genome by swapping out the DNA binding motifs. For RLE-bearing retrotransposons to be used as site-specific gene targeting vehicles, however, the unidentified DNA binding domain will need to be identified. In addition, a greater knowledge of the 3D and globular domain structure is of great interest.

Table 2.1 - Primers used for cloning and for generating target DNAs.

Name	Sequence (5′ to 3′)
R9 ZF3-Myb, fwd	CACCAGTGACAAGCAGGATAATATTAACATAGTTAATGTTAAGGCG
R9 ZF3-Myb, rvs	TTAATTGTTATTACTAGAAGATATGTCACTGTCACTGTCTTCGC
R9 ZF1 Myb, fwd	CACCAATACTAATCAAGTAATATCAAGAAATCCACTTCAGTGCGT
R9 ZF1 Myb, rvs	TTAATTGTTATTACTAGAAGATATGTCACTGTCACTGTCTTCGC
NeSL ZFs, fwd	GGGAATTCCATATGAAGCGTCGTGTCGAGCTGG
NeSL ZFs, rvs	CGCGGATCCTTACCTGGTTTTCGGGTCATCATCC
Target 1, fwd	CACCCGTGTGTAGAATGAAGCCAGGGGAAA
Target 1, rvs	ATTTAAAGTTTGTCAATAGGTTGAGCGTATTTTACGCCC
Target 2, fwd	CACCGTAGCTGGTTCCGTCCGAAGTTTCC
Target 2, rvs	TGTGCATCCAACAGCGCCAG
Target 3, fwd	CACCCGTGTGTAGAATGAAGCCAGGGGAAA
Target 3, rvs	CTAGATGGTTCGATTAGTCTTTCGCCCC
Target 4, fwd	GGGTATGAAGTTCATTTTTCTACTTCATTGAGAATGAACAG
Target 4, rvs	TGTGCATCCAACAGCGCCAG
NeSL Target, fwd	TGCTGTAGGTTGGTGTTAGGCGC
NeSL Target, rvs	CGTTCCAAAATTTATAGCTAACGCC

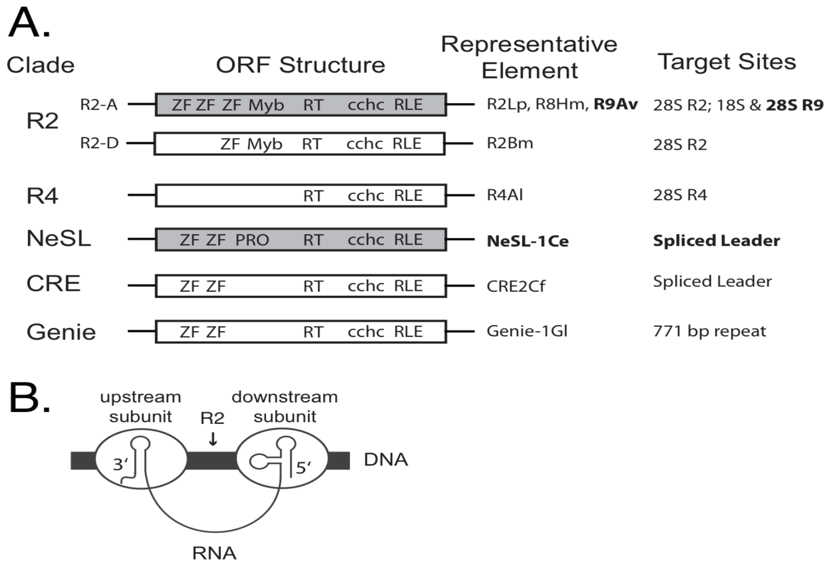

Figure 2.1: Restriction-like endonuclease bearing non-LTR retrotransposon structure and insertion mechanism. (A) A generalized Open Reading Frame (ORF) structure diagram of each of the major recognized RLE-bearing non-LTR retrotransposons clades is depicted along with the target site(s) and the name of a representative element is given. In the structure diagrams the ORF is depicted by a rectangle. The lines flanking ORF rectangle are the 5′ and 3′ untranslated regions, respectively. The two major structural variants of R2 are given. The element names in bold text and gray filled structures are the subject of the paper. Abbreviations: R9 element from *Adineta vaga* (R9Av), R8 from *Hydra magnipapillata* (R8Hm), R2 from *Limulus polyphemus* (R2Lp), R2 from *Bombyx mori* (R2Bm), R4 element from *Ascaris lumbricoides* (R4Al), NeSL from *Caenorhabditis elegans* (NeSL-1Ce), CRE2 from *Crithidia fasciculata* (CRE2Cf), Genie-1 from *Giardia lamblia* (Genie-1Gl), Zinc Finger (ZF), Myb motif (Myb), Reverse Transcriptase (RT), cysteine-histidine rich motif (cchc), and restriction-like endonuclease (RLE). The lines flanking the ORFs are the 5′ and 3′ untranslated regions, respectively. Elements and conserved domains are not drawn to scale. (B) Model of R2 insertion as shown for the R2Bm element [75]. The R2Bm RNA contains higher order RNA structures that function as protein binding motifs. Two subunits of protein are bound to single RNA, forming a pseudo-dimer of protein linked through the RNA. The R2Bm protein bound to the 3′ UTR RNA binds adopts a protein conformation that binds upstream of the insertion site (insertion site = arrow) through a unidentified protein R2 protein domain. Protein bound to the 5′ RNA adopts a protein conformation that binds downstream of the insertion site through the amino-terminal ZF and Myb motifs. Insertion is proposed to be catalyzed by the two protein subunits in four steps. Step 1: the RLE from the upstream subunit is responsible for first-strand cleavage. Step 2: the RT of the upstream subunit catalyzes first-strand TPRT using the cleaved DNA as a primer. Step 3: the downstream subunit cleaves the second DNA strand. Step 4: the downstream subunit provides the polymerase to perform second-strand synthesis using the cleaved DNA as a primer. Step 4 has not yet been shown to occur in vitro.

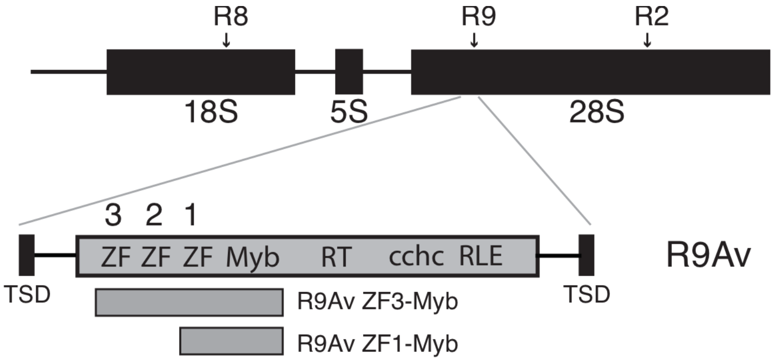

Figure 2.2: R2-A group structure and target sites. A ribosomal array unit is depicted with the 18S, 5S, and 28S ribosomal genes indicated by individual black boxes separated by intervening sequences (lines). The relative positions of R9, R8, and R2 insertion sites are marked by arrows. The domain structure of R9Av (gray) is depicted. Abbreviations and symbols are as in Figure 1. R9Av is flanked by a 126 bp target site duplication (TSD) (black rectangles). The portions of R9Av cloned, expressed, and tested for DNA binding activity are indicated. The R9Av subclones are named based on which ZF and Myb domains are included in the clone (see ZF numbering above the R9Av structure).

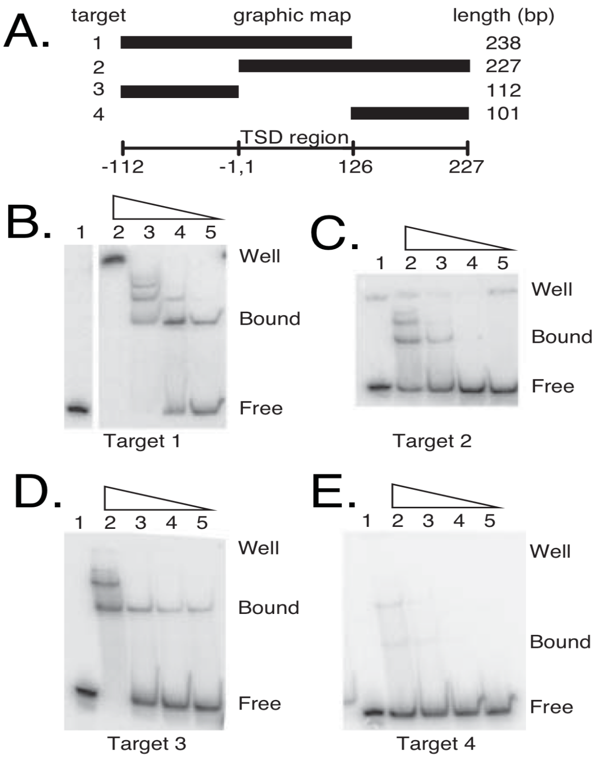

Figure 2.3: Electrophoretic mobility shift assay (EMSA) of the R9Av ZF3-Myb polypeptide bound to potential target DNAs.

(A) Diagram of the target site DNAs used in the EMSA reactions shown in panels B-E. Target 1 consisted of the segment of the 28S that becomes duplicated upon R9 insertion (126 bp TSD region) along with 112 bp of upstream flanking sequence (net 238 bp). Target 2 was the 126 bp TSD region along with 101 bp of downstream flanking sequence (net 227bp). Target 3 was the 112 bp of upstream flanking sequence. Target 4 was the 101 bp of downstream flanking sequence. Target 1 was used for the footprint assays in Figure 4. Below the targets is a ruler. The numbers correspond to base pair positions relative to the presumptive R9 bottom strand cleavage site (i.e., the presumptive site of TPRT), with negative numbers corresponding to target sequences upstream of the TPRT site and positive numbers downstream. Top strand cleavage site is expected to occur at bp 126, thus generating the TSD upon insertion [61].(B) R9Av ZF3-Myb polypeptide EMSA on Target 1. All lanes represent 13 ul binding reactions containing 9 ng (9.4 nM) target DNA end-labeled with ^{32}P in the presence of 90 ng of cold calf thymus DNA as a competitor. Lane 1 is the DNA reference lane (no protein). Lane 2 contains 240 nM of R9Av ZF3-Myb polypeptide. Lane 3 contains 80 nM of R9Av ZF3-Myb polypeptide. Lane 4 contains 27 nM of R9Av ZF3-Myb polypeptide. Lane 5 contains 9 nM of R9Av ZF3-Myb polypeptide. The triangle above Lanes 2-5 indicates the R9Av ZF3-Myb polypeptide titration series (240 nM - 9 nM). (C) R9Av ZF3-Myb polypeptide EMSA on Target 2 (9.8 nM). Lanes and symbols are as in panel B. (D) R9Av ZF3-Myb polypeptide EMSA on Target 3. All lanes are as in panel B except that 20 nM of end-labeled target DNA was used. (E) R9Av ZF3-Myb polypeptide EMSA on Target 4. All lanes are as in panel B except that 22 nM of end-labeled target DNA was used.

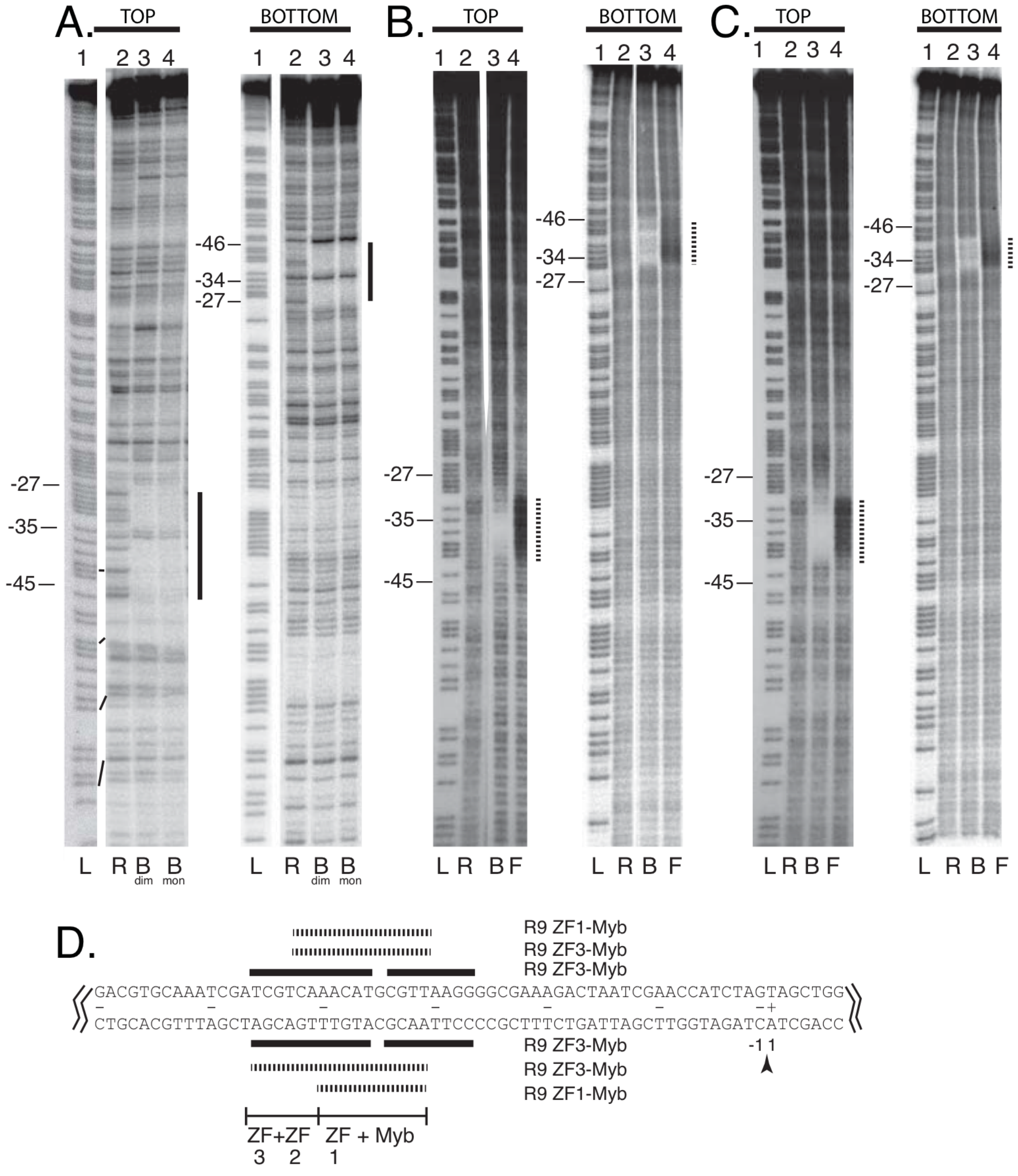

Figure 2.4: DNA footprints of the R9Av ZF3-Myb and polypeptides.
(A) DNase I footprint on Target 1. The 238 bp Target 1 DNA was 5′ end labeled on either the top (left panel) or bottom strand (right panel). Lanes 1, adenine-plus-guanosine ladders (L). Lanes 2, DNase I pattern of naked DNA (R). Lanes 3, two polypeptides (i.e., dimer) bound DNA (B_{dim}). Lanes 4, single polypeptide (i.e., monomer) bound DNA (B_{mon}) The numbers to the left of the footprint correspond to base pair positions relative to the presumptive R9 bottom strand

cleavage site (i.e., the presumptive site of TPRT) as in Figure 3. Regions of DNA that are protected from DNase I degradation by the presence of the R9Av ZF3-Myb polypeptide are marked with thick black lines. Short thin black lines mark polypeptide binding induced DNase I hypersensitive sites. (B) Missing nucleoside footprint of R9Av ZF3-Myb polypeptide bound to 5′-end-labeled hydroxy-radical-treated 238 bp Target 1 DNA. Missing nucleoside footprinting is a binding interference based assay. Hydroxyradical treatment of target DNA generates abasic sites and cleaves the DNA backbone at the abasic site. Protein is then added to the treated DNA in a binding reaction. Abasic sites that interfere with protein binding are under-represented in the bound fraction and over represented in the free fraction. Binding reactions are fractionated into the component bound and free fractions by EMSA prior to being analyzed on denaturing polyacrylamide gels. Lanes 1, adenine-plus-guanosine ladders. Lanes 2, missing nucleoside reference pattern. Lanes 3, bound DNA fraction. Lanes 4, Free DNA fraction. The numbers to the left are as in Figure 3. Nucleosides that, when missing, interfere with binding are marked with dashed lines. (C) Missing nucleoside footprint of R9Av ZF1-Myb polypeptide bound to 5′-end-labeled hydroxy-radical-treated 238 bp Target 1 DNA. Lanes are as in panel B. (D) Summary of the R9Av footprints. The presumptive bottom strand cleavage/TPRT site is indicated by the arrowhead. Base pair positions are numbered as in Figure 3 (i.e., the relative to the site of TPRT). Regions of DNA that are protected from DNase I degradation by the presence of the R9Av ZF3-Myb polypeptide are marked with thick black lines. Short thin black lines mark polypeptide binding induced DNase I hypersensitive sites. Nucleosides that, when missing, interfere with R9Av ZF3-Myb binding and R9Av ZF1-Myb binding, respectively, are marked with dashed lines. Jagged lines indicate that only the relevant portion of Target 1 is being shown.

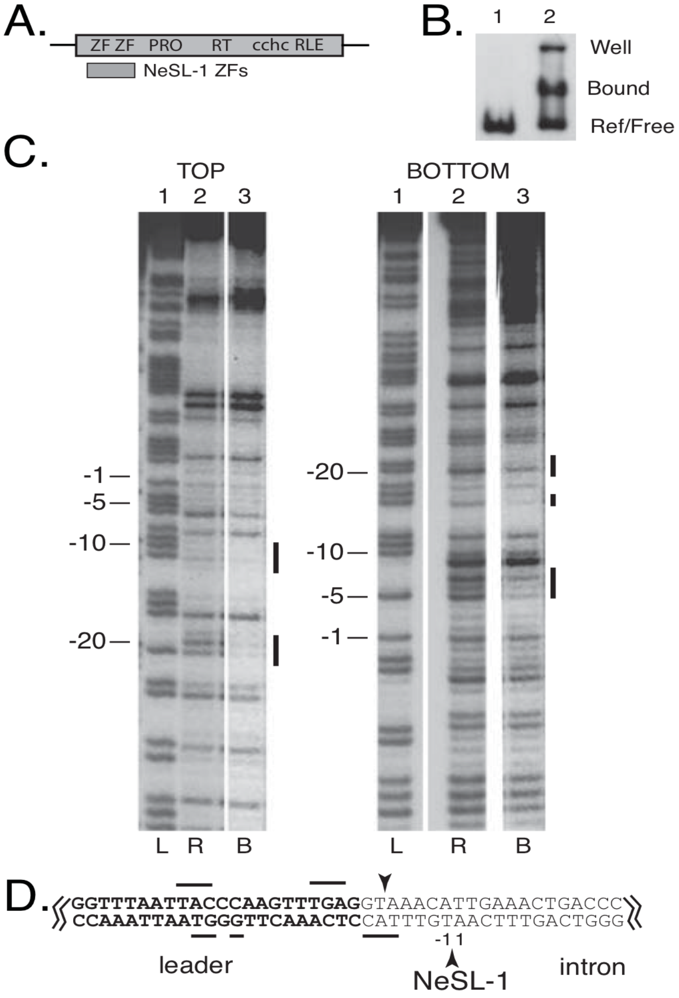

Figure 2.5: Target recognition by NeSL-1, a non-R2 clade element.
(A) Structure of NeSL-1. Symbols and abbreviations are as in Figure 2. Additional abbreviation: protease domain (pro). The region of NeSL-1 that was cloned and analyzed for DNA binding activity is indicated. (B) EMSA. EMSA using a 125 bp DNA fragment encompassing the spliced leader exon along with flanking sequences. Lane 1: reference DNA lane containing no protein. Lane 2: protein plus DNA lane. (C) DNase I footprint. Abbreviations and symbols are as in Figure 4. Base pair numbering scheme is relative to the site of TPRT, with negative numbers corresponding to target sequences upstream of the TPRT site and positive numbers downstream. (D) Summary of footprint on target sequence. Symbols are as in Figure 4. Jagged lines indicate that only the relevant section of the 125 bp target sequence that was used in the footprint analysis is shown. The leader sequence is in bold text. The intron sequence is in normal text. The presumptive cleavage sites are marked with arrowheads. The expected site of TPRT is at the bottom strand cleavage site. The base pair numbering is centered around this cleavage site. Sequences protected from DNase I degradation are marked with horizontal lines. This work was done by Dillon Cawley.

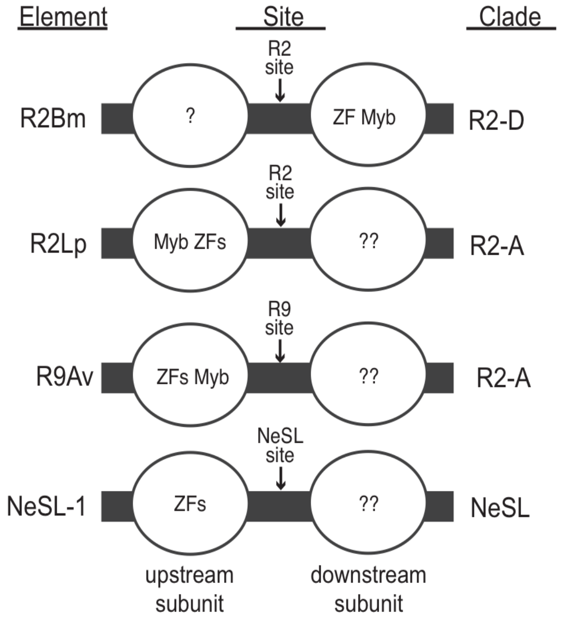

Figure 2.6: Summary of RLE-bearing non-LTR retrotransposon DNA binding modes and implications for the integration model. Abbreviations and symbols are as in Figure 1. Myb and ZF indicates that a subunit of the listed non-LTR retrotransposon protein binds to target DNA at the indicated position via the Myb and/or ZF motifs. The order of the listing of Myb and ZF indicates the determined binding order along the DNA. A single question mark indicates that the position, existence, and role of the protein subunit has been determined, but the protein motif used to secure the protein to the DNA has not been determined. This subunit binds 3′ RNA, cleaves the bottom strand DNA, and performs TPRT. Two question marks indicate that the existence and position of the subunit on the DNA is hypothetical; however, it would be expected to bind to DNA using the aforementioned undetermined DNA binding domain used in R2Bm. The subunit with two question marks is further speculated to bind 3′ RNA, cleave the bottom strand, and perform TPRT (as does its cognate in R2Bm). The subunit binding via the variable amino-terminal Myb and/or ZFs is known (R2Bm) or speculated to (R2Lp, R9Av, and NeSL-1) to bind 5′ RNA, cleave the top DNA strand, and (hypothetically) perform second strand synthesis—processes known to be highly variable between lineages. Alternatively, there is high degree of plasticity in how the ordered series nucleic acid binding (RNA and DNA), DNA cleavage, and polymerization functions are carried out by the two subunits.

CHAPTER 3

SUPPLEMENTARY INFORMATION

3.1 Troubleshooting

3.1.1. Generating Expression constructs and recombinant protein expression:

DNA segments containing different combinations of ZF and myb motif encoding sequences were PCR amplified from the *Adineta vaga* genome (Figure 3.1). These products were cloned into a destination vector using Gateway technology (Invitrogen). Initially, all the expression constructs were generated in the destination vector pDest17. pDest17 is a vector which uses bacteriophage T7 promoter for the expression of the gene of interest. This vector also contains a 6x His tag on its N-terminal which is useful during the purification of recombinant proteins (Discussed later). These recombinant constructs were maintained in arctic express DE3 RIL chemically competent cells. These cells are derived from BL21 *E.coli* cells and encode T7 polymerase. The expression of T7 polymerase which is under the control of *lacUV5* promoter is induced by Isopropyl β-D-1-thiogalactopyranoside (IPTG). The additional features of these cells are that they have enhanced efficiency for protein folding and solubility and they also overcome the problem of codon bias.

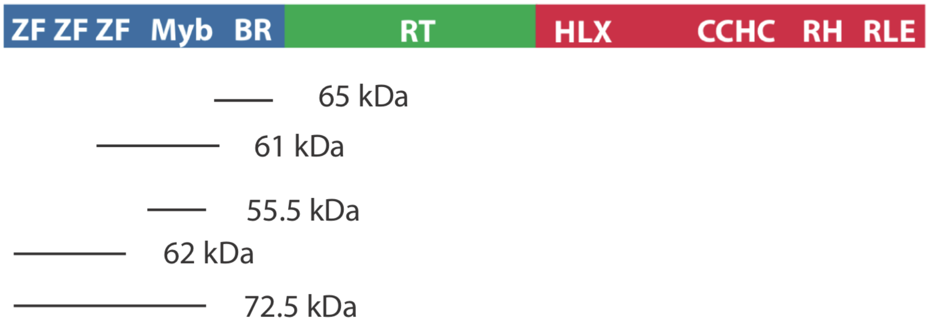

Figure 3.1: R9 domain structure with the location of different expressed segments. R9 consists of amino-terminal DNA binding domains (Blue), Central reverse transcriptase domain (Green) and Carboxy terminal domain bearing the endonuclease (Red). Zinc finger (ZF), BR (Basic region), Reverse transcriptase (RT), Restriction-like endonuclease (RLE). Below he domain structure is given a map of various constructs generated that spans different regions of the amino-terminal end. The molecular size of the polypeptide is also provided.

These expression constructs resulted in only insoluble protein, which could have occurred due to incorrect protein folding. In order to fix this protein solubility issue, a new destination vector called pDestTAP was used [60]. This expression vector, along with the features of pDest17, had an MBP fusion tag on its N-terminal end. All the recombinant DNA segments were re-transformed into pDestTAP and expressed. This time, soluble proteins for all the constructs except for construct BR (Figure 3.2), were obtained in concentrations enough to do further experiments.

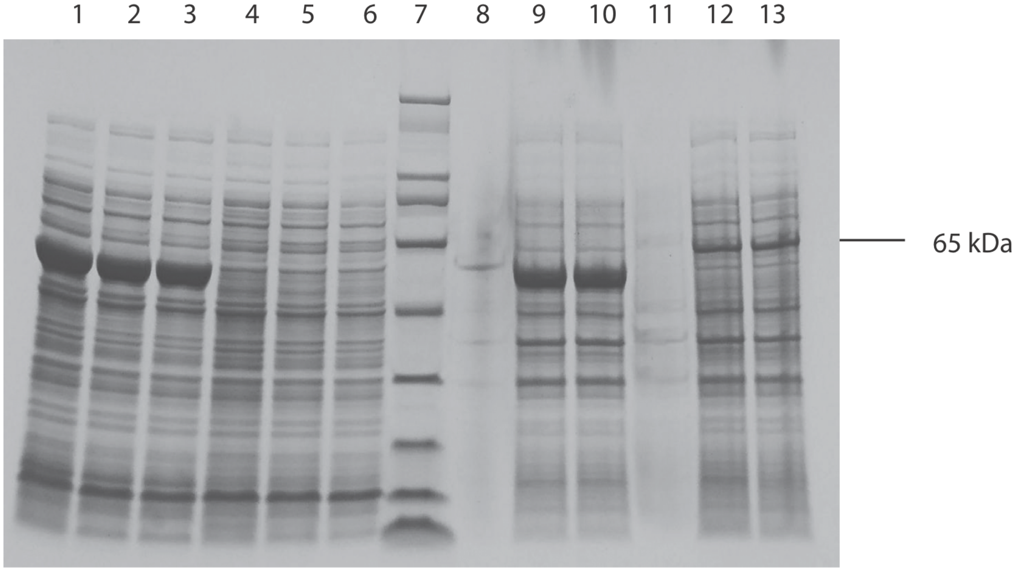

Figure 3.2: Profile expression of R9 myb and R9 BR construct. Sypro stained 7% SDS PAGE gel. Lanes 1-3 are the samples from the soluble fraction and lanes 8-10 are the insoluble fraction of the R9 myb protein expressed at 0.1, 0.3 and 1mM IPTG. Lanes 4-6 are the soluble fraction and lanes 11-13 are the insoluble fraction of R9 BR protein at 0.1, 0.3 and 1mM IPTG. Lane 7 is the size marker lane.

3.1.2 Protein purification

Protein purification is a very critical step for functional assays. All the steps of protein purification were done at 4^0C to prevent the proteins from denaturing and to protect them from protease activity. The cells were frozen and then lysed by detergent (Triton X100). The lysed cells were centrifuged for 20 hours in high salt to separate DNA from proteins. The supernatant obtained after the spin were purified using affinity chromatography. This technique takes advantage of the proteins ability to bind specifically to certain molecules. In order to facilitate this process, some proteins are modified by adding a few amino acids, called 'tag', to one of its two ends. In case of pDestTAP vector, 6X histidine tag is added to the N-terminal end.

The above-mentioned recombinant proteins were purified by affinity chromatography using TALON Metal affinity resin (Clonetech). TALON resin that contains cobalt has a very high affinity for adjacently placed histidine residues. Once the histidine-tagged proteins are bound to the column, they can be eluted off by adding imidazole. Since imidazole has a similar structure to histidine side chain, they outcompete histidines for resin binding (TALON instruction manual).

All the recombinant proteins were purified using the modified version of TALON purification method (Stringent TALON purification). These modifications were to make the purification conditions more similar to the traditional R2Bm purification buffers [60]. Unfortunately, this involved a significant amount of loss of protein yield during the stringent washes. This observation indicated to us that there was only a weak binding interaction between the protein and the resin or there was something in the wash buffer which resulted in the protein eluting off at lower imidazole concentrations (Figure 3.3 A).

We first attributed this problem to high concentration of reducing agents like beta-mercaptoethanol (BME) and dithiothreitol (DTT) which could have led to a detrimental reduction of cobalt, leading to loss of protein binding. Reducing agents are used in the purification process to help preserve the reduced sulfhydryl (-SH) groups in the protein. These reduced -SH groups are important for biological activity in certain proteins (TALON instruction manual). DTT was used during the lysis procedure which we thought could have got carried over. In order to test for the effect of these reducing agents, the supernatant was diluted by 50% with water and then passed through the talon resin. As seen in figure 3.2A, the pattern of protein loss at 15mM, 30mM and 45mM imidazole washes for the diluted supernatant was the same as in case of undiluted or direct supernatant. This convinced us to rule-out the possibility of reducing agents interfering with binding.

Since the problem continued to persist, our next suspect was the possibility of 6x histidine tag being hidden or blocked, resulting in loss of binding interaction between the resin and the protein. In order to circumvent a low efficacy ‘his’ tag, we took advantage of the presence of the maltose binding tag in the expression constructs. The proteins this time were purified by affinity chromatography using the amylose resin (Figure 3.2 B). Amylose resin uses the same principle as used by TALON resin. The differences lie in the ‘tag’ used for binding with the resin and the reagent used to elute out the bound protein. In this case, the amylose resin has a high affinity for maltose binding protein (MBP) tag and maltose is used to elute off the protein. We used n-octyl-β-D-glucoside (OG), since the amylose resin is very sensitive to triton X100 and also because it is easier to remove OG from the final protein extract. The recombinant proteins were purified by the standard column chromatography procedure. In order to dilute down the concentration of detergent in the protein, we doubled the quantity of the protein supernatant with water, before passing through the column. Stringent washes with increasing concentrations of maltose were done on the protein-bound amylose resin and the final elution step was done at 50mM maltose concentration. As seen in figure 3.2B, most of the protein eluted off at 2mM maltose wash, indicating that the protein did not bind the resin.

However, when supernatant was passed through talon first and then through amylose, the protein stopped eluting out during the washes, although the yield at the end of the procedure was quite low(Figure 3.2 C). As a last resort, we used the batch purification method with talon resin described in section 2.2.2, where the yield was high but the purity was comparatively low (Figure 3.2 D). Though the problem was not solved completely, sufficient protein was recovered for the next experiments.

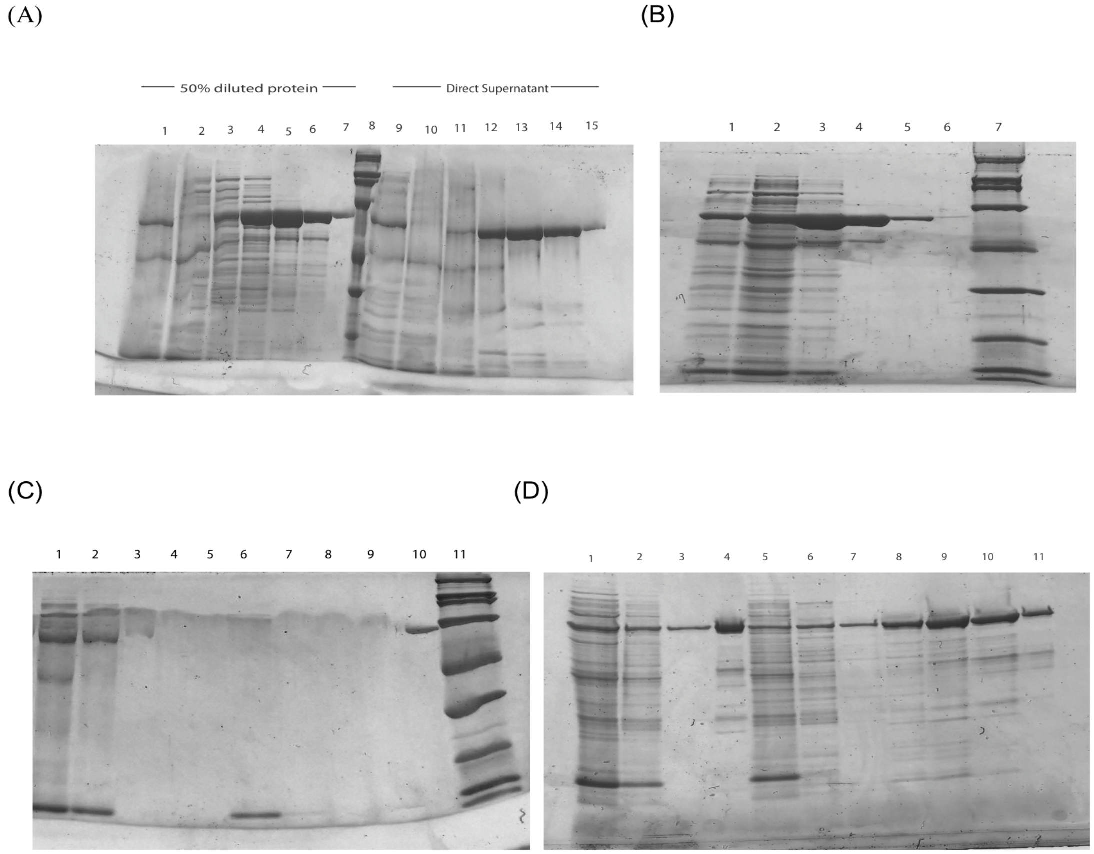

Figure 3.3: Troubleshooting Protein purification. Sypro stained 7% SDS PAGE gel (A) Test for the effect of BME. Lanes1-7 contain samples from 50% diluted supernatant while lanes 9-15 has samples from undiluted supernatant. Lane 8 is the size marker. Lanes 2-6 and 10-14 contain samples from the talon washes with increasing imidazole amounts (5mM,10mM,15mM,30mM and 45mM respectively). Lane 1 and lane 9 contains the sample from the supernatant prior to purification. Lane 6 and lane 15 is the eluant at 150mM imidazole. (B) Purification with maltose binding protein affinity tag. Lane 1 contains the sample from the supernatant before purification, Lane 2 is the flowhrough and lanes 3-6 contain the washes with increasing amounts of maltose (2mM, 5mM, 10mM and 50mM). (C) Tandem purification with talon and amylose resin. Supernatant was first passed through amylose resin and then the final wash and elute was passed through talon resin. Lanes 2-5 are the amylose washes with increasing maltose concentrations(1mM, 5mM, 10mM and 50mM). Lanes 7-10 are the talon washes with increasing concentrations of imidazole (0mM, 5mM, 10mM and 15mM). Lane 1 and lane 6 are flowthrough samples from amylose and talon respectively. (D) Batch purification method on talon resin. Lanes 1-4 are loaded with samples from bulk purification method and lanes 5-11 are loaded with samples from stringent was method. Lane 1 is the sample from the supernatant. Lanes 2-3 are samples from 10mM imidazole washes. Lane 4 is the protein eluate at 150mM imidazole. Lanes 5-7 are the 5mM imidazole talon washes, lanes 8-10 contain higher imidazole concentration washes (10mM, 20mM and 40mM respectively) and lane 11 is the eluant at 150mM imidazole.

3.1.3. Electrophoretic mobility shift assays

Following successful purification, electrophoretic mobility shift assays were done on the peptides with different targets (Figure 3.6A) using the reaction conditions previously described under section 2.2.3. This assay is used to detect DNA-protein interactions and is based on the observation that protein-DNA complexes formed from specific protein-DNA interactions migrate slower than the DNA. These complexes can have different mobilities depending on how many proteins form the complex which can be seen as discrete bands on a gel. For all these experiments, the gel set up was kept cold with the help of ice slurry. When the gels were run at higher temperatures, a smear was observed from the complex to the free DNA, indicating that proteins were losing their binding strength and thus falling off. The wells in the gel were rinsed with buffer before loading the sample to get rid of any unpolymerized acrylamide. Running gels without washing the wells always resulted in the bands looking wavy.

Since initial attempts did not show specific binding, a few parameters were changed and the experiments were repeated. We first tried the binding assay for myb peptide in presence of different non-specific competitor DNAs like poly dIdC and fish sperm DNA(Figure 3.6 B). Presence of such non-specific competitor DNAs prevents secondary binding of other DNA binding proteins from the sample, to the labeled target [80]. The binding assays for myb were also tested under various salt conditions and at different concentrations of target DNA (Figure 3.6 C). We then tried binding myb to the target DNA at a high NaCl concentration (150mM) since the myb polypeptide from R2Bm was previously shown to bind at that concentration [60]. As seen from figure 3.6, no DNA-protein complexes were observed under any of the tested binding conditions.

In the case of peptides made of zinc finger domains in the absence of myb, binding assays were done with different targets under the reaction conditions described previously (Section 2.2.3). Consistent with the ZF binding data for R2Lp (R2 element from *Limulus polyphemus)*, R9 ZFs also could not be shown to bind.

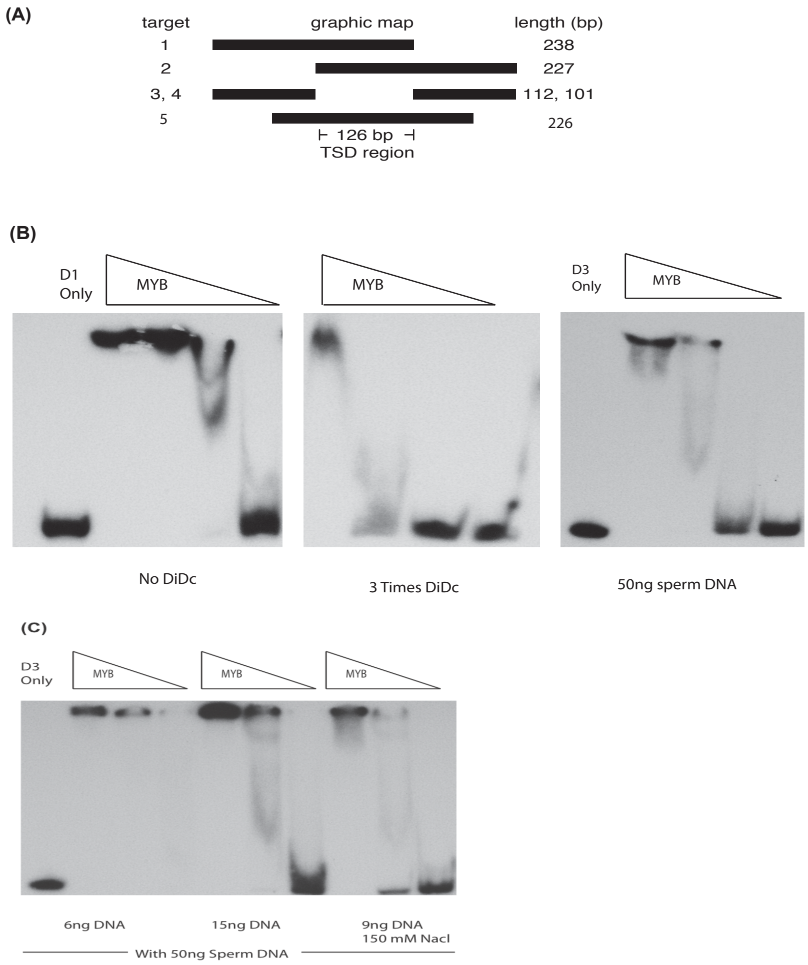

Figure 3.4: Troubleshooting enzyme mobile shift assay for myb polypeptide. The triangles represents decreasing concentrations of myb polypeptide from left to right (1103, 367, 122, 41 fmol) . The first lane contains the sample from the binding reaction in the absence of protein. (A) A map of different target DNAs with their respective length are shown. Horizontal black bars represent different DNA targets used in the binding reaction. (B) All binding reactions were done in presence of 9ng of DNA. The first snapshot (Left) is the binding assay of myb with target 1 in the absence of any competitor DNA. The second snapshot (Center) is the binding assay of myb with target 1 in presence of Poly dIdC with a concentration three times higher than target DNA. The third snapshot (Right) is the binding assay of myb to target 3 in presence of fish sperm DNA at a concentration 5 times that of target DNA. (C) This snapshot contains three sets of binding reactions. The first lane is a binding reaction in absence of myb. Lanes 2-4 are reactions in presence of lower concentration (6ng) of target DNA and 50mM Nacl. Lanes 4-7 are binding reactions in presence of higher concentration of target DNA (15ng) and 50mM Nacl. Lanes 8-10 are binding reactions at a higher concentration of Nacl (150mM).

3.2 Additional Experiments

3.2.1. Electrophoretic mobility shift assays for ZF1-myb

As discussed in section 2.3.1, binding assays were also done on the shorter polypeptide containing a zinc finger and myb domain. The binding affinity to various targets showed a very similar pattern to that of ZF3-myb polypeptide. This observation indicated that, the two zinc fingers present in ZF3-myb polypeptide did not add any specificity to binding specificity.

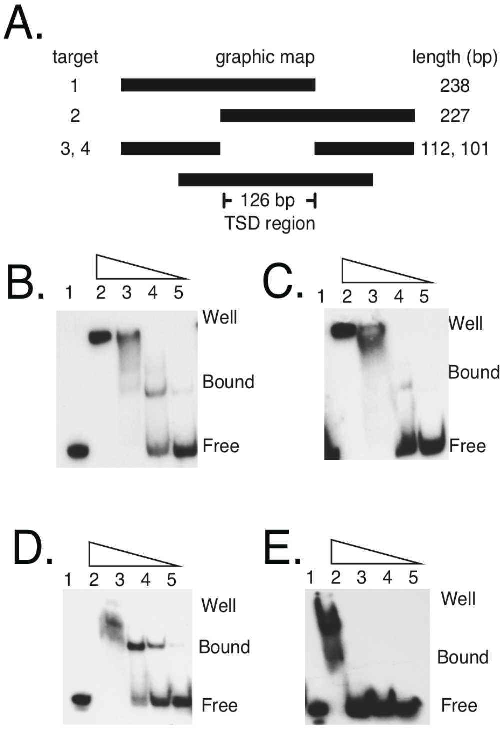

Figure 3.5: Electrophoretic mobility shift assay (EMSA) of the R9Av ZF1-Myb polypeptide bound to potential target DNAs. (A) Diagram of the target site DNAs used in the EMSA reactions shown in panels B-E. Target 1 consisted of the segment of the 28S that becomes duplicated upon R9 insertion (126 bp TSD region) along with 112 bp of upstream flanking sequence (net 238 bp). Target 2 was the 126 bp TSD region along with 101 bp of downstream flanking sequence (net 227bp). Target 3 was the112 bp of upstream flanking sequence. Target 4 was the 101 bp of downstream flanking sequence. Target 1 was used for the footprint assays in Figure 4. Below the targets is a ruler. The numbers correspond to base pair positions relative to the presumptive R9 bottom strand cleavage site (i.e., the presumptive site of TPRT), with negative numbers corresponding to target sequences upstream of the TPRT site and positive numbers downstream. Top strand cleavage site is expected to occur at bp 126, thus generating the TSD upon insertion.

3.2.2. Electrophoretic mobility shift assay and missing nucleoside footprint of ZF3-myb to R2Bm target

In order to confirm our hypothesis that R9 has changed its site specificity from the ancestral site to a new site, we performed a binding assay of R9 ZF3-1 polypeptide to the conserved ancestral target site from R2Bm. We observed that R9 bound the ancestral site in the same manner it bound to the new R9 target site. To test for sequence specific interaction we performed a missing nucleoside experiment using the procedure described in section 2.2.3. Missing nucleoside footprinting is a binding interference based assay. This method is used to find the base contacts between the DNA and protein by chemically removing bases from the target DNA. Hydroxyradical treatment of target DNA generates an abasic site and cleaves the DNA backbone at the abasic site. Protein is then added to the treated DNA in a binding reaction. Abasic sites that interfere with protein binding are under-represented in the bound fraction and over represented in the free fraction. Binding reactions are fractionated into the component bound and free fractions by EMSA prior to being analyzed on denaturing polyacrylamide gels. This experiment, however, showed no sign of important protein-DNA sequence interaction suggesting that R9 interacted non-specifically with the ancestral site.

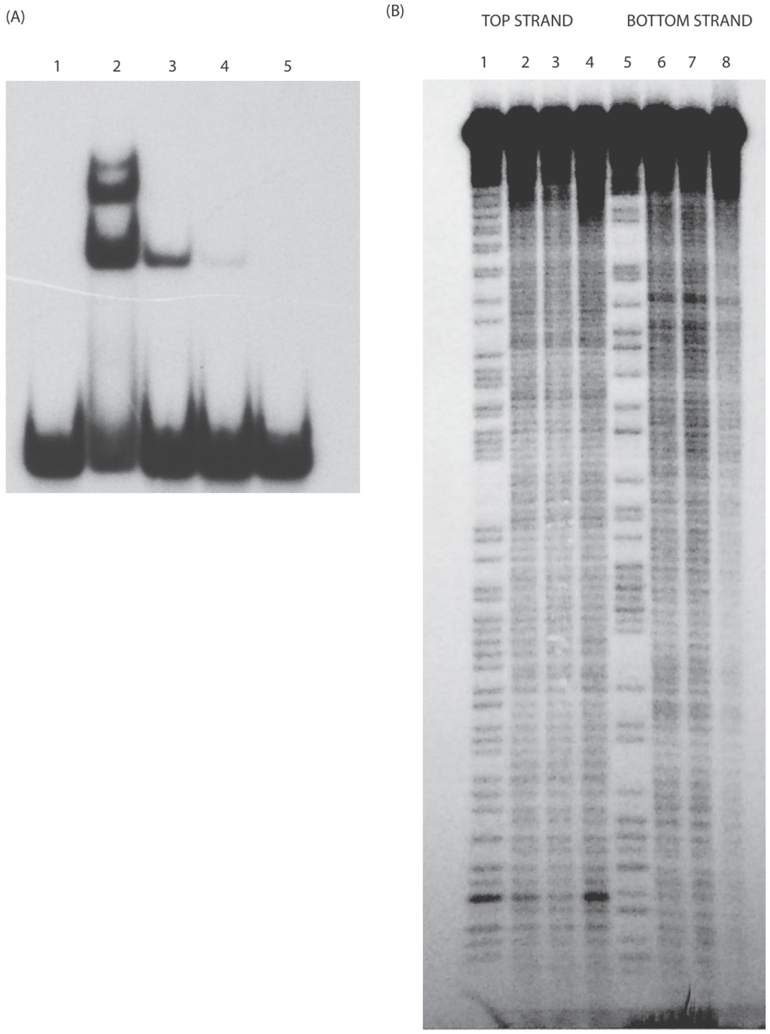

Figure 3.6: Binding assay of R9 ZF3-myb to R2Bm site. (A) This is a snapshot of a 5% Native polyacrylamide EMSA gel (19:1) of R9 ZF3-myb polypeptide binding to the R2Bm target site. Lane 1 is the sample from binding reaction in the absence of protein. Lanes 2-5 are samples from binding reactions containing one-third dilutions of the protein(3062 fmol- 37.8 fmol) in presence of 9ng of target DNA. (B) Missing nucleoside footprint of R9Av ZF3-Myb polypeptide bound to 5′-end-labeled hydroxy-radical treated 120 Bp R2Bm target DNA. Lanes 1-4 represent top strand while lanes 5-8 represent bottom strand. Lanes 1 and 5, adenine-plus-guanosine ladders. Lanes 2 and 6, missing nucleoside reference pattern. Lanes 3 and 7, bound DNA fraction. Lanes 4 and 8, Free DNA fraction.

3.3 Diagnostic footprint experiments

3.3.1 Diagnostic footprint to determine the optimum DNase concentration

To perform DNase footprint assays, the amount of DNase per reaction should be such that only one cut is made per DNA molecule. In order to determine this concentration of DNase, mock footprint reactions containing DNA in the absence of proteins were prepared. Each reaction containing equal amounts of DNA were cleaved by different dilutions of DNase. Based on the result from this diagnostic experiment, 0.01 units of DNase were used for all DNase footprint assays.

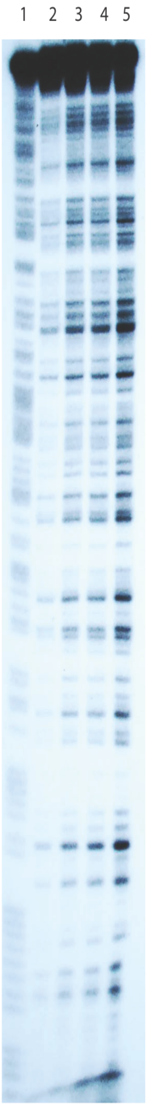

Figure 3.7: DNase diagnostic gel. 6% denaturing urea gel. Lane 1 contains the adenine-plus-guanosine ladder, Lanes 2-5 represent different concentrations of DNase- 0.004, 0.005, 0.006, 0.01 units respectively.

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. Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20**, 211-221 (2010).
4. Kazazian, H. H. J. Mobile elements: drivers of genome evolution. *Science* **303**, 1626-1632 (2004).
5. Mikkelsen, T. S. et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-177 (2007).
6. Gogvadze, E. & Buzdin, A. Retroelements and their impact on genome evolution and functioning. *Cell Mol Life Sci* **66**, 3727-3742 (2009).
7. van de Lagemaat, L. N., Landry, J. R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**, 530-536 (2003).
8. Lisch, D. & Bennetzen, J. L. Transposable element origins of epigenetic gene regulation. *Curr Opin Plant Biol* **14**, 156-161 (2011).
9. Sijen, T. & Plasterk, R. H. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* **426**, 310-314 (2003).
10. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* (2011).
11. Britten, R. J. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**, 177-182 (1997).
12. Yang, N. & Kazazian, H. H. J. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol* **13**, 763-771 (2006).
13. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007).
14. Xing, J. et al. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res* (2009).
15. Curcio, M. J. & Derbyshire, K. M. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**, 865-877 (2003).
16. Eickbush, T. H. & Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* (2008).
17. 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).
18. Kim, F. J., Battini, J. L., Manel, N. & Sitbon, M. Emergence of vertebrate retroviruses and envelope capture. *Virology* **318**, 183-191 (2004).
19. Pelisson, A., Mejlumian, L., Robert, V., Terzian, C. & Bucheton, A. Drosophila germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect Biochem Mol Biol* **32**, 1249-1256 (2002).
20. Baldrich, E. et al. Genomic distribution of the retrovirus-like element ZAM in Drosophila. *Genetica* **100**, 131-140 (1997).

21. Nelson, P. N. et al. Human endogenous retroviruses: transposable elements with potential? *Clin Exp Immunol* **138**, 1-9 (2004).
22. Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR retrotransposons. *Genome Biol* **5**, 225 (2004).
23. 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).
24. Gao, X., Havecker, E. R., Baranov, P. V., Atkins, J. F. & Voytas, D. F. Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* **9**, 1422-1430 (2003).
25. Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR retrotransposons. *Genome Biol* **5**, 225 (2004).
26. Llorens, C., Munoz-Pomer, A., Bernad, L., Botella, H. & Moya, A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol Direct* **4**, 41 (2009).
27. de la Chaux, N. & Wagner, A. BEL/Pao retrotransposons in metazoan genomes. *BMC Evol Biol* **11**, 154 (2011).
28. Hirano, N., Muroi, T., Takahashi, H. & Haruki, M. Site-specific recombinases as tools for heterologous gene integration. *Appl Microbiol Biotechnol* **92**, 227-239 (2011).
29. Goodwin, T. J. & Poulter, R. T. A new group of tyrosine recombinase-encoding retrotransposons. *Mol Biol Evol* **21**, 746-759 (2004).
30. Evgen'ev, M. B. & Arkhipova, I. R. Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res* **110**, 510-521 (2005).

31. Pyatkov, K. I. et al. Penelope retroelements from Drosophila virilis are active after transformation of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **99**, 16150-16155 (2002).
32. 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).
33. Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annu Rev Genet* **38**, 1-35 (2004).
34. 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).
35. Kojima, K. K. & Fujiwara, H. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15**, 1106-1117 (2005).
36. 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).
37. Martin, S. L. The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. *J Biomed Biotechnol* **2006**, 45621 (2006).
38. Martin, S. L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. *RNA Biol* **7**, 706-711 (2010).
39. 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).
40. Chaboissier, M. C., Busseau, I., Prosser, J., Finnegan, D. J. & Bucheton, A. Identification of a potential RNA intermediate for transposition of the LINE-like element I factor in Drosophila melanogaster. *EMBO J* **9**, 3557-3563 (1990).
41. Busseau, I., Chaboissier, M. C., Pelisson, A. & Bucheton, A. I factors in Drosophila melanogaster: transposition under control. *Genetica* **93**, 101-116 (1994).

42. Doucet, A. J. et al. Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet* **6**, (2010).
43. Khazina, E. et al. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat Struct Mol Biol* **18**, 1006-1014 (2011).
44. 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).
45. 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).
46. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165 (2005).
47. Zhou, J. & Eickbush, T. H. The pattern of R2 retrotransposon activity in natural populations of Drosophila simulans reflects the dynamic nature of the rDNA locus. *PLoS Genet* **5**, e1000386 (2009).
48. Eickbush, T. H. & Eickbush, D. G. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**, 477-485 (2007).
49. Zhang, X., Eickbush, M. T. & Eickbush, T. H. Role of Recombination in the Long-term Retention of Transposable Elements in rRNA Gene Loci. *Genetics* (2008).
50. Eickbush, T. H. & Eickbush, D. G. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**, 477-485 (2007).
51. Eickbush, D. G. & Eickbush, T. H. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol* (2010).
52. Moss, W. N., Eickbush, D. G., Lopez, M. J., Eickbush, T. H. & Turner, D. H. The R2 retrotransposon RNA families. *RNA Biol* **8**, (2011).
53. 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* **8**, 3-10 (1999).
54. 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).
55. 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).
56. 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).
57. Luan, D. D. & Eickbush, T. H. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* **15**, 3882-3891 (1995).
58. Mandell, J. G. & Barbas, C. F. r. Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res* **34**, W516-23 (2006).
59. Wei, S. Y. et al. Structure of the Trichomonas vaginalis Myb3 DNA-binding domain bound to a promoter sequence reveals a unique C-terminal {beta}-hairpin conformation. *Nucleic Acids Res* (2011).
60. Thompson, B. K. & Christensen, S. M. Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: plasticity of integration mechanism. *Mobile Genetic Elements* **1**, 29-37 (2011).
61. 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).
62. Han, J. S. & Boeke, J. D. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *Bioessays* **27**, 775-784 (2005).

63. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
64. Ye, J. & Eickbush, T. H. Chromatin structure and transcription of the R1- and R2-inserted rRNA genes of Drosophila melanogaster. *Mol Cell Biol* **26**, 8781-8790 (2006).
65. Beck, C. R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-1170 (2010).
66. 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).
67. 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**, 502-511 (1999).
68. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).
69. 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).
70. Mandal, P. K., Bagchi, A., Bhattacharya, A. & Bhattacharya, S. An Entamoeba histolytica LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease. *Eukaryot Cell* **3**, 170-179 (2004).
71. Volff, J. N., Korting, C., Froschauer, A., Sweeney, K. & Schartl, M. Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. *J Mol Evol* **52**, 351-360 (2001).
72. 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).
73. 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).
74. 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* **21**, 207-217 (2004).
75. 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).
76. Kojima, K. K., Kuma, K., Toh, H. & Fujiwara, H. Identification of rDNA-specific non-LTR retrotransposons in Cnidaria. *Mol Biol Evol* **23**, 1984-1993 (2006).
77. MD, A., PJ, M. & SJ, R. Image Processing with ImageJ. *Biophotonics International* **11**, 36-42 (2004).
78. 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).
79. Stage, D. & Eickbush, T. 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).
80. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861 (2007).

BIOGRAPHICAL INFORMATION

Haridha Shivram was born and brought up in India. He got his bachelor of technology degree from Sathyabama university, Chennai, India. Haridha then went on to get his master's in science degree in biology from University of Texas at Arlington where he worked in the lab of Dr Shawn Christensen. He plans to continue working in the same lab before joining a school for Ph.D.