Protein-DNA interactions of the RLE LINE R2Bm

By

SHALINI RACHAKONDA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Biology at
The University of Texas at Arlington
December 2020

Arlington, Texas

Supervising Committee:

Dr. Shawn M. Christensen, Supervising Professor
Dr. Matthew K. Fujita
Dr. Piya Ghose
Dr. Paul Chippindale

ABSTRACT

Protein-DNA interactions of the RLE LINE R2Bm

Shalini Rachakonda, Masters (Thesis)

University of Texas at Arlington, 2020

Supervising Professor: Shawn M Christensen

Long interspersed elements (LINEs) are a major group of non-long terminal repeat (non-LTR) retrotransposable elements that are ubiquitous in eukaryotic genomes. These selfish genetic elements influence the structure and function of the host genome. R2 elements are site-specific LINEs that insert at a specific target site in the 28S rRNA genes of the host. R2 elements encode a single open reading frame which makes a multifunctional protein containing reverse transcriptase, DNA endonuclease, and nucleic acid-binding domains. The R2 RNP reverse transcribes its mRNA to DNA at the site of insertion in order to integrate into the host genome. The first half of the integration reaction involves cleavage of the host DNA by the element encoded DNA endonuclease and priming of first strand cDNA by the free 3'-OH generated by the DNA endonuclease, a process called Target Primed Reverse Transcription (TPRT). The second half of the reaction, second strand cleavage and second strand DNA synthesis are accomplished with a second round of DNA cleavage / DNA polymerization events. The N terminal domain of the R2 encoded protein contains from 1-3 zinc fingers and a Myb domain. The first chapter of my thesis briefly reviews what is known about the R2 integration reaction. In chapter 2, I investigate the DNA binding potential of the zinc fingers. In the third chapter, I investigate the sequences of the DNA target required for binding to the R2 protein.

ACKNOWLEDMENTS

I would like to express my sincere thanks to my principle investigator, Dr. Shawn Christensen for entrusting me with his important research ideas and for giving me his mentorship, guidance and motivation throughout the tenure of my Masters. I would also like to thank Dr. Piya Ghose Dr. Matthew Fujita and Dr. Paul Chippindale for their valuable advises and immense support. I would also like to thank Dr. Kimberly Bowels for supporting me and guiding me not just as an advisor but also as a friend. I thank my fellow lab mates – past (Brijesh Khadgi) and present (Santosh Dhamala, Joydeep Chatterjee, Sun Tran and Thu Tran) who were understanding and cooperative throughout my project. Thank you to each one of you for helping me out throughout; you all shall always be my cherished lab family. I thank the Department of Biology at University of Texas at Arlington and all the staff members for helping and supporting me throughout my time here as a Masters student. Last, but not the least, I would like to thank my parents, my sister and my friends for always being there, believing in me, guiding me, encouraging me and most of all, loving and supporting me all these years.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Transposable elements (TEs), also known as transposons or jumping genes are selfish sequences of DNA that jump to different locations in the genome and replicate, making multiple copies and getting inherited along with the host chromosomes [1]. TEs can make up a sizable fraction of a genome [2]. About 45% of the human genome appears to be TEs [3–5]. TEs play an important role in the genome function and evolution as their replication can cause insertion, deletion and recombination events [6,7]. TEs are also an important source of novel genetic material resulting in new genes and regulatory sequences for the host [6,8,9]. There are two distinct mechanistic classes of TEs, Class 1 (Retrotransposons) and Class 2 (DNA transposons), based upon whether the integration intermediate is RNA or DNA [10]. Retrotransposons encode a reverse transcriptase and integrate through an RNA intermediate. Retrotransposons are further classified based upon the mode of integration and the presence of other major genes. LTR retrotransposons reverse transcribe the RNA intermediate into double-stranded DNA in a virus-like particle using a tRNA as an initial primer for cDNA synthesis. LTR retrotransposons then use an element encoded integrase or a recombinase, to integrate the double-stranded DNA into the genome [11,12].

Non-LTR retrotransposons encode a DNA endonuclease, which acts as a DNA nickase, cleaving one host chromosomal DNA strand at the insertion site [13–18]. The nick generates free 3’ hydroxyl which is used to prime the first strand cDNA synthesis by reverse transcription of the element RNA at the site of insertion: a process that has been dubbed target primed reverse transcription

(TPRT) [4,13,19–22]. As a result, non-LTR retrotransposons are sometimes referred to, or classified as, target-primed retrotransposons. A major group of target-primed retrotransposons are the Long-interspersed elements (LINEs). LINEs encode either a restriction like DNA endonuclease (RLE) or an apurinic-apyrimidinic DNA endonuclease (APE) (Figure 1) [23,24]. RLE LINEs are generally site-specific, inserting into specific sequences in the genome, while APE LINEs tend to insert non-specifically. RLE LINEs are considered to be the more ancient one of the two [25,26]. RLE LINEs have one single open reading frame (ORF)[3] and are about 3-4 kb in length (Figure 1). A single ORF of the RLE LINE often encodes one to three zinc fingers (ZF), a Myb motif at the amino terminal 5' end, a reverse transcriptase (RT) in the center, a linker region with a gag-like Zinc knuckle and a PD(D/E)xK-family DNA endonuclease on the 3' end [14,17,18,26–31]. The RLE and APE bearing LINEs are divided into 6 groups- R2, RandI, L1, RTE, I, Jockey- these are further subdivided into more than 28 clades [32,33]. APE bearing LINEs are composed of L1, RTE, I, Jockey and include all other clades except for RandI [34]. The mammalian L1 element of the L1 clade is one of the most well studied elements belonging to the APE LINEs group. The RLE LINEs are composed of R2, R4, NeSL, CRE, HERO, and Genie clades [23,31,35–37]. The R2 element of the R2 clade is one of the most well studied elements belonging to the RLE LINE group [13,25,31,38–41]. My work focuses on the R2 element. Below, I will summarize what is known about R2 elements.

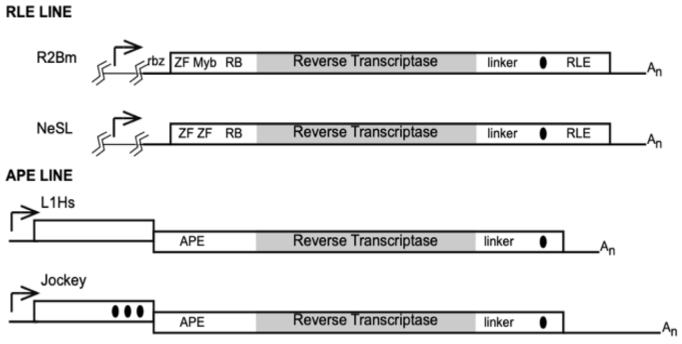

Figure 1. Representative RLE and APE LINE ORF structures with major motifs. Dissimilarity between the two LINEs represented using two examples of each. The arrows represent promoters, black ovals represent IAP/gag-like zinc knuckle motifs. Abbreviations: ZF- Zinc finger, RB- RNA binding domain, RLE- Restriction-like endonuclease, APE- apurinic-apyrimidinic endonuclease, rbz- ribozyme.

Discovery of R2 elements

Non- rDNA segments of about 5kb in length have been observed into interrupt a fraction of 28S ribosomal genes [42]. The insertions which were divided into two classes based on the nucleotide sequence of the junction regions- type I and type II [42]. Type I insertion elements were observed in Drosophila virilis [43], D.melanogaster [44,45], Calliphora erythrocephala [46], and Bombyx mori [47,48]. Type II insertion elements were found only in D. melanogaster [44,45] and B. mori [47,48]. This Type I and Type II sequences were later name R1(type I) and R2 (type II) elements (R refers to the ribosomal insertion site / order of discovery) [42]. R1(type I) in a site-specific APE LINE. R2 (type II) is site-specific RLE LINE. The R2 supergroup/superclade is named after the original finding of the R2 element inserted at the "R2" site. R2 elements have been found to target other sites in the ribosomal genes. The R8 site is located in the 18S rRNA gene [41,49]. The R9 site is located in the 28S rRNA gene upstream of the R2 site (Figure 2). Based on the reverse transcriptase sequences, the R2 supergroup consists of four clades - R2-A, R2-B. R2-C, R2-D [41,50,51]. Different clades have a different number of Zinc fingers in the N-terminal region along with a single Myb domain [35,50,51]. R2-A consists of three zinc-fingers (CCHH, CCHC, CCHH) upstream of the Myb domain (Figure 3A) [50]. R2-C has two zinc-fingers, both CCHH, upstream of the Myb domain (Figure 3A) [50]. R2-D has a single CCHH zinc-finger upstream of the Myb domain [50] (Figure 3A). The R2-D clade elements which include R2Bm uses the ZF and Myb DNA binding motifs to bind downstream of the target insertion site. In contrast, the R2-A clade elements which include R2Lp and R8 Hm bind to the sequences upstream of the insertion site through the same motifs (Figure

3B). NeSL is not a part of the R2 clade. R2-A clade is thought to represent the ancestral R2 clade. R2-A clade members have been found to target the R2, R8, and R9 sites[41,49]. R2-D clade members have only been found to target the R2 site [18,26,35,50]. R2s have been observed to be widely distributed among the animal phyla including Nematoda, Arthropoda, Echinodermata, Cnidaria, Chordata, and Platyhelminthes [37,41,50,52]. There were reports of them being found in cyclostomes, fish, reptiles, hagfishes, birds and coelacanth [50,51,53,53]. R2 families have also been recently discovered in Ctenophora, Mollusca, and Hemichordata [50,53].

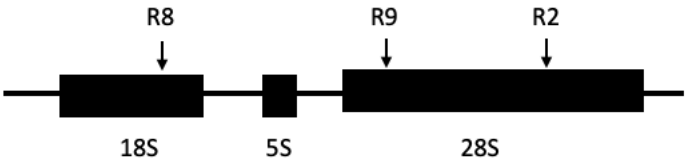

Figure 2. Ribosomal array unit. 18S, 5S and 28S depicted as 3 separate blocks. The insertion sites for R2, R8 and R9 have been indicated using arrows.

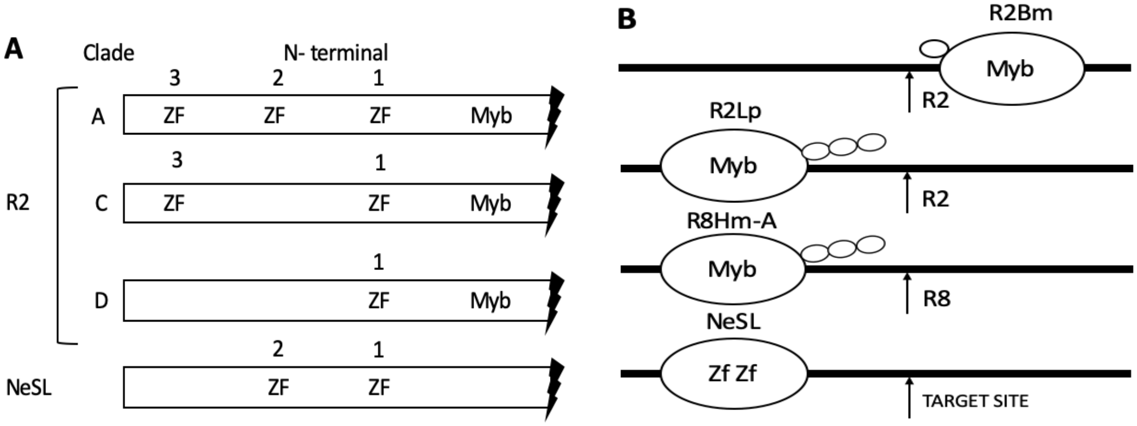

Figure 3. N- terminal domains of R2 and NeSL clades and their target site DNA binding. (A) R2-A clade (R2Lp, R8Hm-A) has 3 ZFs and one Myb domain. R2-C has 2 ZFs and one Myb domain and R2-D clades have one ZF and one Myb domain. The NeSL clade has two ZFs with no Myb. **(B)** The R2Bm (R2-D clade) Myb (large circle) binds to target DNA downstream to the insertion and the Zf are orientated towards the insertion site. R2Lp, R8Hm-A (R2-A clade) Myb

ploypeptides bind upstream of the insertion site with the ZFs (small circle) oriented towards the insertion site. The NeSL clade uses the ZFs (large circle) to bind upstream of the insertion site.

The lifecycle and integration mechanism of RLE and APE LINEs is functionally very similar [54]. The DNA is transcribed to make an element RNA using the element encoded promoter or a promoter located upstream of the inserted element. This RNA transcript is exported out to the cytoplasm where it is translated to a protein. The element protein binds to the element RNA it was translated from to form an integration-component ribonucleoprotein (RNP) complex. The RNP complex enters the nucleus and starts the integration at a new site.

Transcription of R2 elements

The R2 of Drosophila targets the R2 site in the 28S ribosomal gene, just like in many other species (Figure 2). Host factors make the individual ribosomal RNAs from the long transcript generated by the ribosomal promoter (Figure 4). Ribosomal units with an R2 element inserted into the 28S rDNA, are transcribed, generating a co-transcript that includes the R2 element [55–57] [39] [40] [41]. The HDV-like ribozyme at the 5’ end of the R2 element processes the element away from majority of the upstream ribosomal sequences [55,56,58,59]. It has been determined that the self-cleavage site in most R2 elements is within the 28S rRNA gene from 9 to 36 nucleotides upstream of the R2 5’ junction or the insertion site [55]. The capability of the HDV-like ribozyme to cause rapid and efficient self-cleavage of 28S/R2 co-transcript was discovered by in vitro studies of synthesized RNAs. *Drosophila simulans* has a precise self-cleavage site located at the 28S/R2 5’ junction [18,55]. R2 elements with ribozymes that are predicted to cleave within the 28S sequences tend to generate 5’ junctions with fewer small nucleotide additions and/or small target deletions upon insertion than do elements whose ribozymes cleave at the 28S/R2 site [55]. Sometimes, they also generate junctions with tandem duplications of upstream 28S sequence and the length is consistent with the location of the predicted ribozyme cleavage site [55]. Other RLE LINEs are observed to encode HDV-like ribozymes (e.g. R4) and might be typical for site-specific elements with no self-encoded promotors [58,60]. The processing of the 3' end of the R2 element is not known.

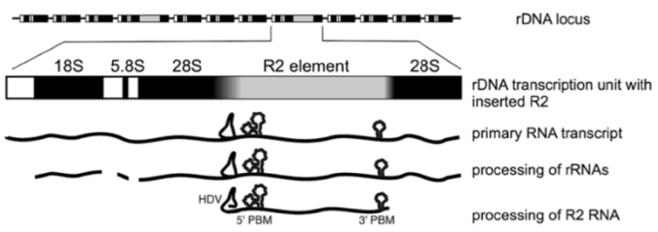

Figure 4. Transcription of an rDNA unit with an R2 element insertion. The R2 element inserted into the rDNA unit in the rDNA locus. Primary transcript that initiates at the promoter of the ribosomal unit is processed to form rRNAs and then to R2 RNA.

Translation of R2 elements

Translation initiation in R2 takes place with a cap independent mechanism because the R2 element lacks 5' methyl guanosine cap as the transcript is derived from a polI transcript and further processed by HDV-like ribozyme. 5' UTR of R2 is dominated by conserved RNA structure because of the constraints of the HDV-like ribozyme. Conservation of amino acids is dominant over RNA structure in the ORF. The ORF and the RNA structure appear to be linked in an area of overlap [55,56,61,62]. The pseudoknot structure of HDV has been hypothesized to function as an internal-ribosome-entry-site (IRES) similar to the ones found in viruses and a few cellular mRNAs [55,56,61,62].

Formation of the R2 RNP

The 5' PBM is a structured segment present in the 5' UTR of the R2Bm element RNA (Figure 5A). The R2Bm protein binds to this region of the element RNA. It has been hypothesized that this region could be an IRES location in R2Bm and other related moths [61]. The R2 5’ UTR found in Drosophila is very small when compared to the one found in R2Bm. This is because of the presence of the 5’ protein binding motif and an HDV-like ribozyme [56,61,62]. The 3’ PBM is a major component of the 3’ UTR region of the R2 transcript (Figure 5A). It has a conserved secondary sequence [13] [22,61,63]. RNP formation with the protein bound to the 3’ PBM is essential for the first half of the integration reaction (i.e. TPRT) [13,22,64]. The R2 protein from Bombyx mori (R2Bm) recognizes the 3’ RNA of *Drosophila melanogaster* (R2Dm) and many other distantly related R2 elements, even with no evident sequence similarity.

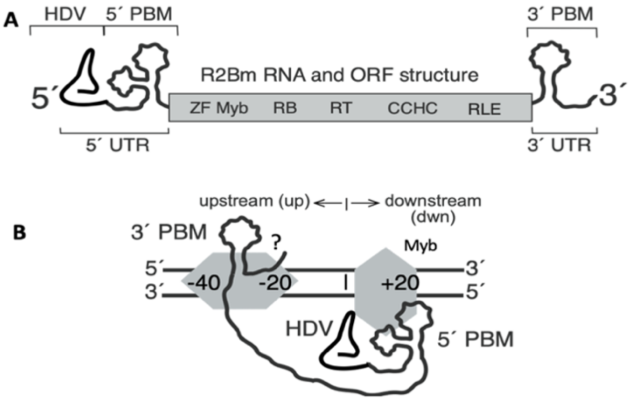

Figure 5: R2Bm RNA and ORF structure and binding to the target 28S rDNA.

(A) The 3’ UTR contains a 3’ Protein binding motif (PBM). The 5’ UTR contains a 5’ PBM and an HDV like ribozyme. **(B)** R2Bm RNP bound to the 28S rDNA (black parallel lines). R2 protein (gray hexagon) recognizes sequences upstream of the insertion site in the presence of a 3’ PBM RNA. This footprints to the -40 to -20 region of the target DNA. The R2 protein binds the sequences downstream of the insertion site when associated with the 5’ PBM RNA. The footprint of this association goes from about the insertion site to +20 bp. The R2 protein binds downstream with the help of Myb domain. The schematics of the upstream binding are unknown. Figure modified from reference 68.

Integration mechanism of R2 elements

The 5’ untranslated region (UTR) of the R2Bm element RNA contains a 5’ protein binding motif (PBM) and an HDV-like ribozyme [65]. The 3’ UTR of the R2Bm element has the 3’ PBM as a major component. The full length R2 protein behaves different when associated with different RNAs. In the presence of the 3’ RNA, the protein conformation is favorable to bind upstream and perform TPRT [66]. The DNA interactions in the upstream region are not known [64,66]. When the 5’ PBM is associated with the R2Bm protein, the conformational changes in the protein favor binding downstream through the nucleic acid binding Myb domain (Figure 5B) [33] [66,67]. In previous studies, it was determined that this association likely provides the endonuclease required for second strand DNA cleavage [33,66,67].

Integration of an R2 element occurs in two main half reactions, both of which involve a DNA cleavage step and a DNA synthesis step. The first half reaction involves first strand cleavage and TPRT (first strand cDNA synthesis) [13,68]. The element encoded DNA endonuclease nicks the chromosomal target site forming a free 3’ hydroxyl group [67,69]. This 3’ -OH is then used to prime cDNA synthesis with the help of the reverse transcriptase [29]. This is followed by a recombination event or a template jump from the 5′ end of the R2 RNA onto the top-strand (second strand) of the 28S rDNA towards the upstream side- step 3 in (Figure 6) [68]. This forms a 4-way junction [68]. This is followed by second strand (top strand) cleavage and second strand synthesis. The 4-way junction has not been observed *in vivo* but has been determined to be a key integration intermediate that has been explored *in vitro* [68]. A simple unified mechanism for the formation of the 5’ junction and complete integration emerged because of the junction formation.

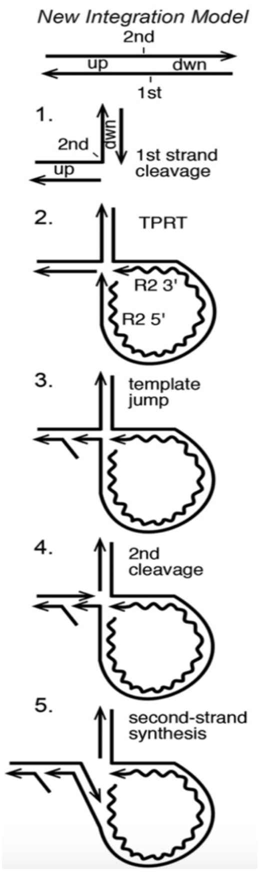

Figure 6. R2Bm integration model. The five steps of the integration: (1) First/ bottom strand cleavage; (2) TPRT (3) a template jump/recombination event that generates an open '4-way' DNA junction; (4) second-strand / top strand DNA cleavage; and (5) second-strand DNA synthesis. Abbreviations: up- target sequences upstream of the insertion site, dwn- target sequences downstream of the insertion site and TPRT is target primed reverse transcription. The RNAs are represented with a wavy line. Figure modified from reference 68.

REFERENCES

1. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**, 21-42 (2012).
2. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* **6**, 699-708 (2005).
3. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
4. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
5. Deininger, P. L., Moran, J. V., Batzer, M. A. & Kazazian, H. H. J. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**, 651-658 (2003).
6. Richardson, S. R. et al. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061 (2015).
7. Casola, C. & Betrán, E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses. *Genome Biol Evol* **9**, 1351-1373 (2017).
8. Craig, N. L. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 423-456 (ASM Press, Washington, DC, 2002).
9. Zingler, N., Weichenrieder, O. & Schumann, G. G. APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet Genome Res* **110**, 250-268 (2005).
10. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007).
11. Bourque, G. et al. Ten things you should know about transposable elements. *Genome Biol* **19**, 199 (2018).
12. Arkhipova, I. R. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* **8**, 19 (2017).
13. 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).
14. 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).
15. Feng, Q., Moran, J. V., Kazazian, H. H. J. & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916 (1996).
16. Guo, H., Zimmerly, S., Perlman, P. S. & Lambowitz, A. M. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J* **16**, 6835-6848 (1997).
17. Govindaraju, A., Cortez, J. D., Reveal, B. & Christensen, S. M. Endonuclease domain of non-LTR retrotransposons: loss-of-function mutants and modeling of the R2Bm endonuclease. *Nucleic Acids Res* **44**, 3276-3287 (2016).
18. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).

19. Anzai, T., Takahashi, H. & Fujiwara, H. Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol Cell Biol* **21**, 100-108 (2001).
20. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).
21. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910 (2002).
22. 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).
23. 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).
24. Weichenrieder, O., Repanas, K. & Perrakis, A. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986 (2004).
25. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3**, MDNA3-0001 (2015).
26. 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).
27. Ostertag, E. M. & Kazazian, H. H. J. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35**, 501-538 (2001).
28. Martin, S. L. & Bushman, F. D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**, 467-475 (2001).
29. Mahbub, M. M., Chowdhury, S. M. & Christensen, S. M. Globular domain structure and function of restriction-like-endonuclease LINEs: similarities to eukaryotic splicing factor Prp8. *Mob DNA* **8**, 16 (2017).
30. Eickbush, T. H. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 813-835 (ASM Press, Washington, DC, 2002).
31. 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).
32. 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).
33. 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).
34. 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).
35. 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).
36. 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).
37. 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).

38. Eickbush, T. H. & Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* (2008).
39. Kojima, K. K. & Jurka, J. Ancient Origin of the U2 Small Nuclear RNA Gene-Targeting Non-LTR Retrotransposons Utopia. *PLoS One* **10**, e0140084 (2015).
40. Kojima, K. K. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mob DNA* **9**, 2 (2018).
41. 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).
42. Burke, W. D., Calalang, C. C. & Eickbush, T. H. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol* **7**, 2221-2230 (1987).
43. Rae, P. M., Kohorn, B. D. & Wade, R. P. The 10 kb Drosophila virilis 28S rDNA intervening sequence is flanked by a direct repeat of 14 base pairs of coding sequence. *Nucleic Acids Res* **8**, 3491-3504 (1980).
44. Dawid, I. B. & Rebbert, M. L. Nucleotide sequences at the boundaries between gene and insertion regions in the rDNA of Drosophilia melanogaster. *Nucleic Acids Res* **9**, 5011-5020 (1981).
45. Roiha, H., Miller, J. R., Woods, L. C. & Glover, D. M. Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in D. melanogaster. *Nature* **290**, 749-753 (1981).
46. Smith, V. L. & Beckingham, K. The intron boundaries and flanking rRNA coding sequences of Calliphora erythrocephala rDNA. *Nucleic Acids Res* **12**, 1707-1724 (1984).
47. Eickbush, T. H. & Robins, B. Bombyx mori 28S ribosomal genes contain insertion elements similar to the Type I and II elements of Drosophila melanogaster. *EMBO J* **4**, 2281-2285 (1985).
48. Fujiwara, H. et al. Introns and their flanking sequences of Bombyx mori rDNA. *Nucleic Acids Res* **12**, 6861-6869 (1984).
49. 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).
50. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165 (2005).
51. Luchetti, A. & Mantovani, B. Non-LTR R2 element evolutionary patterns: phylogenetic incongruences, rapid radiation and the maintenance of multiple lineages. *PLoS One* **8**, e57076 (2013).
52. Jakubczak, J. L., Burke, W. D. & Eickbush, T. H. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci U S A* **88**, 3295-3299 (1991).
53. Kojima, K. K., Seto, Y. & Fujiwara, H. The Wide Distribution and Change of Target Specificity of R2 Non-LTR Retrotransposons in Animals. *PLoS One* **11**, e0163496 (2016).
54. Moran, J. V. & Gilbert, N. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 836-869 (ASM Press, Washington, DC, 2002).
55. Eickbush, D. G., Burke, W. D. & Eickbush, T. H. Evolution of the r2 retrotransposon ribozyme and its self-cleavage site. *PLoS One* **8**, e66441 (2013).
56. Eickbush, D. G. & Eickbush, T. H. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA co-transcript. *Mol Cell Biol* (2010).

57. 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).
58. Ruminski, D. J., Webb, C. H., Riccitelli, N. J. & Lupták, A. Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes. *J Biol Chem* **286**, 41286-41295 (2011).
59. Webb, C. H., Riccitelli, N. J., Ruminski, D. J. & Luptak, A. Widespread occurrence of self-cleaving ribozymes. *Science* **326**, 953 (2009).
60. Sanchez-Luque, F. J., Lopez, M. C., Macias, F., Alonso, C. & Thomas, M. C. Identification of an hepatitis delta virus-like ribozyme at the mRNA 5'-end of the L1Tc retrotransposon from Trypanosoma cruzi. *Nucleic Acids Res* (2011).
61. Moss, W. N., Eickbush, D. G., Lopez, M. J., Eickbush, T. H. & Turner, D. H. The R2 retrotransposon RNA families. *RNA Biol* **8**, (2011).
62. Kierzek, E. et al. Secondary structures for 5' regions of R2 retrotransposon RNAs reveal a novel conserved pseudoknot and regions that evolve under different constraints. *J Mol Biol* **390**, 428-442 (2009).
63. Ruschak, A. M. et al. Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* **10**, 978-987 (2004).
64. 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).
65. Chen, J. H. et al. A 1.9 A crystal structure of the HDV ribozyme precleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester cleavage. *Biochemistry* **49**, 6508-6518 (2010).
66. 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).
67. 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).
68. Khadgi, B. B., Govindaraju, A. & Christensen, S. M. Completion of LINE integration involves an open '4-way' branched DNA intermediate. *Nucleic Acids Res* **47**, 8708-8719 (2019).
69. Zimmerly, S., Guo, H., Perlman, P. S. & Lambowitz, A. M. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**, 545-554 (1995).

CHAPTER 2

ABSTRACT

Long Interspersed Elements (LINEs) are a major group of non- LTR retrotransposable elements which are ubiquitous selfish genetic elements found in the eukaryotic lineages. LINEs play a significant role in structure and expression of the host genome. R2 elements are a group of RLE LINEs that insert at a specific target site in the 28S rDNA locus of the host. The R2 elements encode a multifunctional protein with a reverse transcriptase, DNA endonuclease and nucleic-acid binding domains. The elements insert through a process called Target Primed Reverse Transcription (TPRT) where the element encoded DNA endonuclease generates a 3'-OH which is used to prime first strand synthesis using the RNA as template. The protein binds to the target DNA using the Myb domain. The Zinc fingers were observed to interact with the DNA sequences near the Myb interaction region oriented towards the insertion site. The DNA binding potential of the zinc fingers of R2-A clade and R2-D clade elements are explored. The zinc finger polypeptides by themselves do not seem to bind target DNA. The minimal number of amino acids required to retain the DNA binding ability of R2Bm Myb was also explored.

INTRODUCTION

Long-interspersed-elements (LINEs) are a major group of target-primed retrotransposons or non-LTR retrotransposons. LINEs encode either a restriction like DNA endonuclease (RLE) or an apurinic-apyrimidinic DNA endonuclease (APE) in addition to a shared reverse transcriptase domain [1,2]. RLE LINEs insert site specifically during integration while APE LINEs tend to retrotranspose non-specifically [1,2]. The R2 element, an RLE LINE, has been used as a model system for understanding the integration reaction of LINEs. R2 and RLE LINEs in general are also of interest as elements that might be amenable to protein engineering: in order to change the element's site specificity. If one understood how R2 recognizes its target DNA, one might be able to engineer a R2 element to be an adaptable gene-targeting/ gene replacement vehicle. R2 elements have a single open reading frame (ORF) [3] and are about 3-4 kb in length (Figure 1 of Chapter 1). The R2 ORF encodes for a large multifunctional protein that binds element RNA, binds target DNA, cleaves target DNA, and reverse transcribes the element RNA into DNA at the site of insertion, a process called target primed reverse transcription (TPRT) [3–8].

Based on the reverse transcriptase sequence analysis, the R2 group of elements can be divided up into four (sub)clades: R2-A, R2-B, R2-C, and R2-D clades [9–11]. R2 clades encode a variable number of ZFs along with a Myb domain located in the amino terminal region of the protein[25 23 24 12]. The two major R2 clades are the R2-A and the R2-D clades. The R2-A clade is considered to be the ancestral clade and contains three zinc fingers along with the Myb domain (ZF3, ZF2, ZF1, Myb). The three zinc fingers have consensus cysteine-histidine spacing of $CX_2CX_3FXT/SX_2GX_3HX_4H$, $CX_2CX_{12}HX_3C$, and $CX_2CX_{12}HX_4H$, corresponding to ZF3, ZF2, and ZF1 respectively [9,13,14]. R2-C has two zinc fingers and the Myb (ZF3, ZF1, Myb). R2-D has a single zinc finger and the Myb (ZF1, Myb) [9] (Figure 3A of Chapter 1). Elements belonging to RLE lines outside of the R2 clade also contain amino-terminal zinc fingers, but no Myb domains. HERO encode just one single amino-terminal zinc-finger while NeSL, CRE, Genie clades typically encode two N-terminal ZFs. Elements from R4 clade lack both ZFs and Myb domains.

R2D clade elements have only been found in the R2 site of the 28S rDNA [9,13,15,16] while R2-A clade elements have been found to target either the R2 site in the 28S rDNA, the R9 site in the 28S rDNA, or the R8 site in the 18S rDNA gene(s) [10,17]. That is to say, R2 elements have changed target specificity over evolutionary time.

Previous studies have shown that the Myb domain of R2 binds target DNA. R2Bm, a D clade element uses the Myb to bind to DNA sequences downstream (+10 to +20) of the R2 site. R2Lp, an A clade element that targets the R2 site, uses the Myb to bind to sequences upstream (-40 to -20) of the target site [12,18]. The Myb of an A clade element (R8Hm-A) that targets the R8 site, also binds upstream of the insertion site. In these previous studies, the ZFs were successively chopped off of a construct that expressed the ZF/Myb region of the amino terminal domain of various R2 proteins [61] [70] [67]. The zinc fingers seemed to interact with target DNA near the Myb interaction region and oriented toward the cleavage site [12,19]. However, expression of just the zinc fingers was not exhaustively examined. The idea for researching more into the idea of ZFs binding DNA comes from the DNA binding ability of the two ZFs in NeSL. In this chapter, I will further examine the DNA binding potential of the amino-terminal ZFs and Myb. I have identified the minimal number of amino acids that retain the binding function of the R2Bm Myb. I have also tested the DNA binding potential of the three ZFs of an A clade member in different combinations as well as that of the single zinc finger from a D clade element. The ZF polypeptides do not appear to be able to bind to DNA as polypeptides.

MATERIALS AND METHODS

The templates of the ZF and Myb regions of the R2Lp and R8Hm-A were ordered from IDT DNA. The R8 target site, primers for making DNA templates of R2Bm ZF plus Myb, Myb and different combinations of ZF of R2Bm, R2Lp and R8 were ordered from Sigma Aldrich. The R2Bm Zf, ZFMyb and Myb polypeptides were made using the R2Bm delta N codon optimized plasmid as the template DNA. PCR was performed with optimal annealing conditions to make different fragments of R2Bm, R2Lp and R8Hm-A. These fragments were further gel purified to acquire the template for making protein. The protein was made using PURExpress® In Vitro Protein Synthesis Kit by NEB by addition of 250 ng of DNA template to make a good quality protein. This protein was further used to perform band-shift studies shown in Figure 8. A native 5% polyacrylamide gel was used for all the fragments except for R2Bm ZF and R2Bm Myb 2 (Fig 1A) for which a 10% polyacrylamide gel was used. For R2Lp and R2Bm polypeptides, the 28S 120-mer ribosomal locus target sequence was used as a target DNA, whereas, for the R8 Hm polypeptides, the 18S ribosomal locus target sequence was used. Myb 1 was the most constricted Myb polypeptide made. Myb 2 had a few base pairs added to the 3’ and the 5’ ends. Myb 3 and Myb 4 have a few base pair additions on the 3’ end. Myb 4 protein was used as a positive control as it had previously demonstrated DNA binding ability [61] [12,19]. The ZFZFZFMyb protein of R2Lp has been expressed previously to shift 28S R2 target DNA. Various limits of ZF polypeptides of R2Lp and R8 Hm (Figure 1B) were made in an in vitro system (PURExpress® In Vitro Protein Synthesis Kit by

NEB) to study the DNA binding for each fragment. All proteins studied was accompanied with a no protein lane with the same amount of DNA for a comparative study.

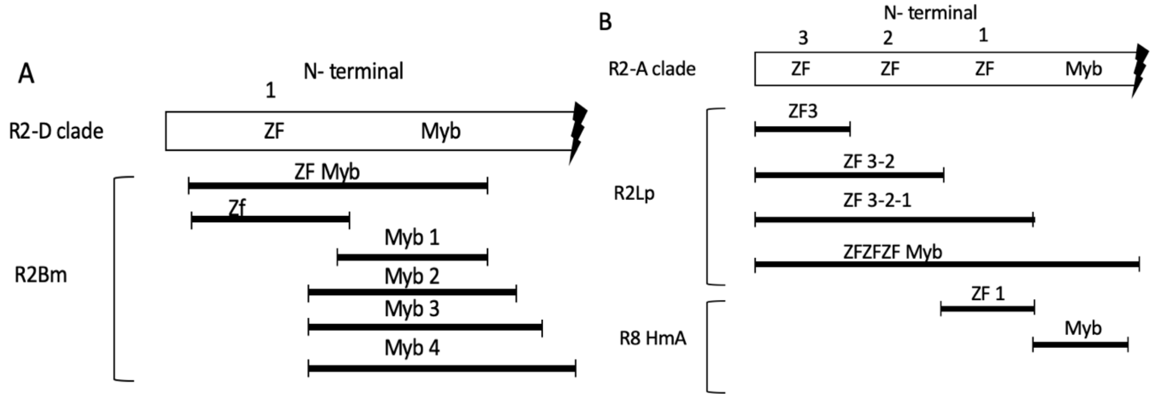

Figure 1 Amino terminal polypeptides used in DNA binding studies. (A) N-terminal domain of R2Bm (R2-D clade) and the limits of the various polypeptides used in DNA binding studies. Myb 1, Myb 2, Myb 3, Myb 4 are not the number of Myb domains but are rather different polypeptide lengths used to determine the smallest polypeptide that will form a functional Myb. **(B)** N-terminal domain of R2Lp and R8Hm-A (R2-A clade) and the limits of the various polypeptides used in DNA binding studies.

RESULTS

In order to study how the R2-D and R2-A clade members target the 28S rDNA, different constructs of ZF and Myb were made. The black blocks under the N-terminal domain in Figure 1 denote the fragments of R2Bm, R2Lp and R8 Hm that were expressed. To understand if the amino terminal Zinc finger motifs of the R2Bm, R2Lp and R8Hm-A are used to bind target DNA, the proteins made from the fragments were used in Electrophoretic Mobility Shift Assay (EMSA). The migration of naked DNA or free DNA and the protein-DNA complexes or bound complexes have been marked in Figure 2 as F (Free) and B (Bound).

When the band-shift assays for the different Zinc finger polypeptides were studied, it was observed that neither the single zinc finger polypeptides nor any combinations of the zinc finger polypeptides were able to bind and shift the target DNA (R2 120-mer target DNA for R2 and R8 target DNA for R8 polypeptides). Even a combination of all three zinc fingers of R2Lp has not shown any target DNA shifting. The ZF1 of R8Hm-A corresponds to the R2Bm ZF and so they would be expected to work in a similar fashion. Neither of the two ZF proteins was able to bind and shift their respective target DNAs.

The R8 Hm Myb construct did not form a functional polypeptide and hence, it does not have a protein-DNA complex band (Figure 2). In R2Bm, the smallest Myb polypeptide (Myb 1) was constricted on both the 5' and 3' ends (Figure 1A). This Myb polypeptide appears to have lost its DNA binding abilities. Myb 4 corresponds to the R2Bm Myb protein sequence that has been previously investigated to study the DNA binding and showed positive results and so was expected to shift the target DNA (Figure 2) [61 70]. Myb 3 and Myb 2 were different polypeptide lengths used to determine the smallest polypeptide that will form a functional Myb polypeptide that binds to the target DNA. Both the polypeptides formed a functional Myb. The Myb 2 protein-DNA complex band runs very close to the free DNA band.

Figure 2 shows the studies on the target DNA binding abilities of Zinc finger plus Myb proteins of R2Bm and R2Lp. It can be observed that both the proteins show a robust shifting of the target DNA as expected.

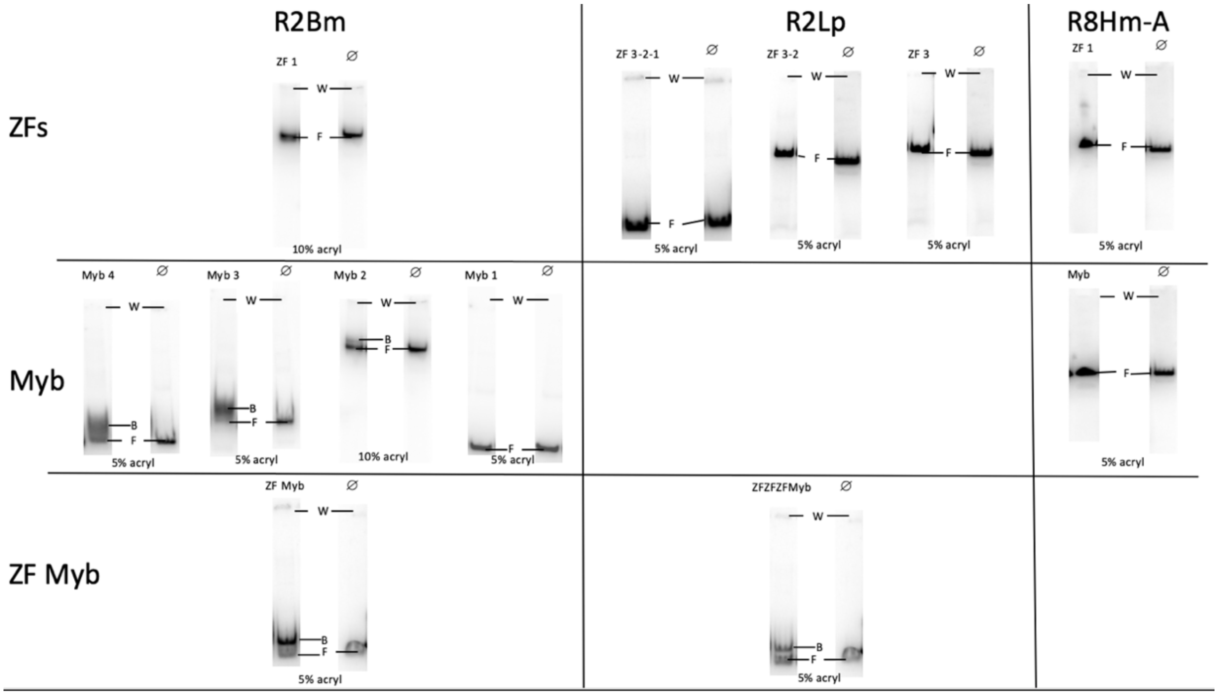

Figure 2 Electrophoretic mobility Shift Assay (EMSA) of different constructs of the N-terminals of R2Bm, R2Lp and R8Hm-A run on 5% and 10% acrylamide as noted. The constructs for the above gels are mentioned in **Figure 1**. Mobility shift of DNA was studied for R2Bm ZF, variable size constructs/polypeptides of R2Bm Myb (R2Bm Myb 1, Myb 2, Myb 3, and Myb 4) and R2Bm ZF Myb polypeptides. Various amino acid terminal polypeptides from the R2-A clade (R2Lp ZF 3-2-1, R2Lp ZF 3-2, R2Lp ZF3, R8Hm-A ZF1), R8Hm-A Myb and R2Lp ZF Myb were studied for DNA mobility shift. The DNA shifting can be observed by comparing the proteins to their corresponding no protein lane (represented as ϕ). Abbreviations: W- Well complex, F- Free DNA, B- Bound/ shifted DNA.

DISCUSSION

While it had been determined in earlier studies that the R2 Myb domain binds to DNA, those studies expressed a polypeptide with a generous amount of sequences to either side of the Myb (see Myb construct 4 in Figure 1) [20 70]. I had wanted to determine the smallest polypeptide capable of correctly folding into a functional R2 Myb so that domain swap studies involving the Myb could be done in the future. The first Myb construct I tested for both R2Bm and R8Hm-A included only the Myb domain as conservatively defined by homology alignments. The Myb, when expressed as a "minimalistic" polypeptide did not bind to target DNA. This was true for the R8Hm-A Myb as well as the R2Bm Myb. I then went back to the published R2Bm Myb construct ("Myb 4") [61] [19]and did a deletion series from the carboxyl-terminal end (Myb 3 and Myb 2). Myb 2, 3, and 4 polypeptides all bound target DNA. Thus, the smallest Myb polypeptide that was functional was the Myb 2 polypeptide. In future studies involving Myb domain swaps, the region corresponding to Myb 2 will be swapped with other Myb domains targeting other DNA sequences.

The earlier studies involving the R2 ZF and Myb polypeptides mainly expressed the ZFs in conjunction with the Myb [21]. While the ZFs appeared to bind to DNA in the context of being connected to the Myb domain, it remained unclear as to how relevant that association was, particularly as the earlier study did not observe the R2Lp ZFs (fingers 3-2-1) to bind to DNA in the absence of the Myb domain. The R2Bm single zinc finger was not tested individually as separate polypeptide. In my study, I tested the first zinc finger (Figure 2) from R2Bm and R8Hm-A. I saw no binding to DNA as assayed by electrophoretic mobility shift assays (Figure 2). I also tested polypeptides to R2Lp Zinc Finger(s) 3, 3-2, and 3-2-1. Again, I saw no binding to DNA. Thus, it appears that it is unlikely that all three zinc fingers bind to DNA in R2-A clade members, or we would have expected the three-zinc finger polypeptide to bind to DNA. It remains possible, however, that we were not generous enough with flanking amino acids to generate correctly folded polypeptide(s). It is also possible that the shift caused by the ZFs, if any, was small enough that the bound band may have all co-migrated with the free DNA. To explore this possibility, we ran some of the EMSA gels on native 10% polyacrylamide instead of our normal 5% polyacrylamide native gels. We still did not observe the ZFs binding to DNA. Further investigation into the ability of the Zinc finger polypeptides to bind to the element 3' PBM RNA and 5' PBM RNA will be done. This could provide more information about the nucleic acid interactions of the Zinc finger domains.

REFERENCES

1. 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).
2. Weichenrieder, O., Repanas, K. & Perrakis, A. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986 (2004).
3. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605 (1993).
4. Anzai, T., Takahashi, H. & Fujiwara, H. Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol Cell Biol* **21**, 100-108 (2001).
5. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927 (1996).
6. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
7. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**, 5899-5910 (2002).
8. 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).
9. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165 (2005).
10. 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).
11. Luchetti, A. & Mantovani, B. Non-LTR R2 element evolutionary patterns: phylogenetic incongruences, rapid radiation and the maintenance of multiple lineages. *PLoS One* **8**, e57076 (2013).
12. Christensen, S. M., Bibillo, A. & Eickbush, T. H. Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33**, 6461-6468 (2005).
13. 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).
14. Eickbush, T. H. in *Mobile DNA II* (eds Craig, NL, Craigie, R, Gellert, M & Lambowitz, A. M.) 813-835 (ASM Press, Washington, DC, 2002).
15. 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).
16. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).
17. 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).
18. 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).

19. Thompson, B. K. & Christensen, S. M. Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: Plasticity of integration mechanism. *Mob Genet Elements* **1**, 29-37 (2011).
20. Moss, W. N., Eickbush, D. G., Lopez, M. J., Eickbush, T. H. & Turner, D. H. The R2 retrotransposon RNA families. *RNA Biol* **8**, (2011).
21. 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).

CHAPTER 3

ABSTRACT

Non-LTRs are a diverse group of elements which include LINEs. LINEs are selfish genetic elements found in all eukaryotic genomes. LINEs play a significant role in the structure and function of the host genome. The R2Bm RLE LINEs integrate specifically in the R2 site in the 28S rRNA genes by a series of DNA binding, DNA cleavage, and DNA synthesis reactions. With the help of DNase I footprint studies, it has been understood that the R2 protein binds the target DNA both upstream (-40 to -20) and downstream (+10 to +20) to the insertion site. The protein uses the Myb domain for binding the sequences downstream of the insertion site but the part of the protein that binds the DNA in the upstream subunit is unknown. The target sequences required for the R2Bm protein to recognize and bind the DNA in the upstream and downstream regions is explored using a SELEX based approach.

INTRODUCTION

R2Bm is a site-specific RLE LINE that targets the R2 site in the 28S rDNA gene [1–4]. DNase I DNA footprint studies have determined that the R2 encoded protein binds to sequences both upstream (-40 to -20) and downstream of the insertion site (+10 to +20) [64 5]. R2 protein bound to the 3' UTR of the element RNA adopts a conformation that binds to target DNA upstream of the insertion site [6]. R2 protein associated with a segment of element RNA from the 5' end of the RNA forces the protein to adopt a conformation that forces the R2 protein to bind to sequences downstream of the insertion site [6,6 66 5]. The downstream subunit binds using the element encoded Myb domain [6,7]. It is unknown what part of the protein is used to bind a subunit upstream of the insertion site[7,8]. The sequence upstream and downstream are imperfect palindromes [5,6]. This chapter investigates the target DNA sequences required for interacting with the R2Bm encoded protein in the upstream and downstream regions.

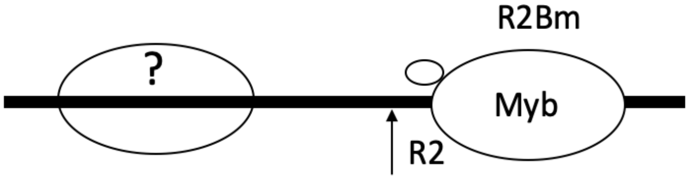

Figure 1. The upstream and downstream DNA- protein interactions in R2-D clade. It is known that the R2Bm, a D clade element, uses the Myb (large circle) to bind DNA sequences downstream to the R2 site (indicated by an upward arrow). The upstream protein DNA interactions are unknown.

The approach used was a SELEX based approach. Target DNA with windows of randomized sequence was allowed to bind to R2Bm RNPs containing either the 5' RNA or the 3' RNA. DNA molecule that could bind to the R2 RNP were separated from those that could not and analyzed by DNA sequencing (Figure 2). The pool of the DNA library becomes smaller with subsequent rounds of selection consisting of the sequences that are most tightly bound. Further sequencing of the extracted bound complexes gives the aptamers for the particular set of experiments.

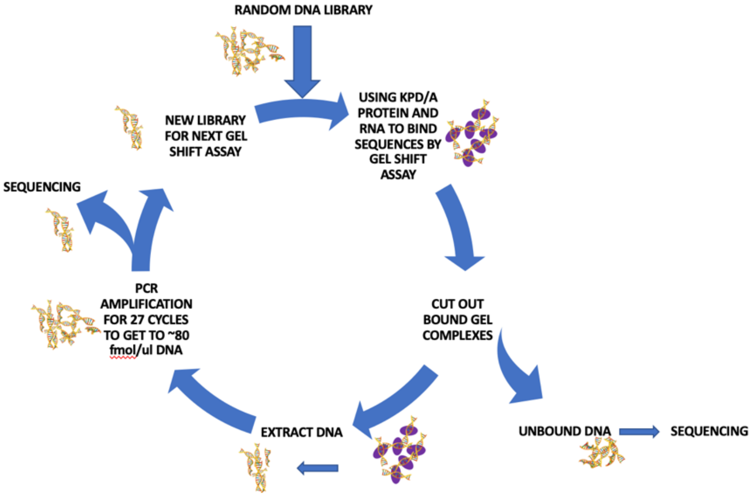
RANDOM DNA LIBRARY
USING KPD/A PROTEIN AND RNA TO BIND SEQUENCES BY GEL SHIFT ASSAY
CUT OUT BOUND GEL COMPLEXES
UNBOUND DNA
SEQUENCING
EXTRACT DNA
PCR AMPLIFICATION FOR 27 CYCLES TO GET TO ~80 fmol/ul DNA
SEQUENCING
NEW LIBRARY FOR NEXT GEL SHIFT ASSAY

Figure 2. SELEX based selection. A round of selection from the initial random DNA library is carried out using delta N KPD/A protein and specific RNA (5’ RNA or the 3’ RNA). The RNP (ribonucleoprotein) bound complexes are extracted and further amplified for subsequent rounds of selection and sequencing analysis.

MATERIALS AND METHODS

The 28S target 120-mer DNA (R2 target site) was modified by replacing the specific target site sequences with random sequences at equal lengths, forming 4 overlapping windows of 18 nucleotides each (Figure 3). The top strands of the four oligonucleotides (Figure 3) were ordered from Sigma Aldrich and a small amount was used to perform a primer extension reaction to make double stranded DNA library. Radioactive ^{32}P labelled primers were used for the primer extension and the PCR conditions were optimized to be feasible for the new primers. The primer extension and the subsequent amplification PCRs were performed at a 58°C annealing temperature. Oligo 1 and Oligo 4 have been studied so far. R2Bm Delta N KPD/A protein (endonuclease mutant) was used as the target ligand for selection. The protein expression and purification of the R2 endonuclease mutant protein (KPD/A) was performed. The 3’ and 5’ PBM RNA were made by invitro transcription reaction and further purified using a Qiagen column (Qiagen PCR Purification Kit).

.

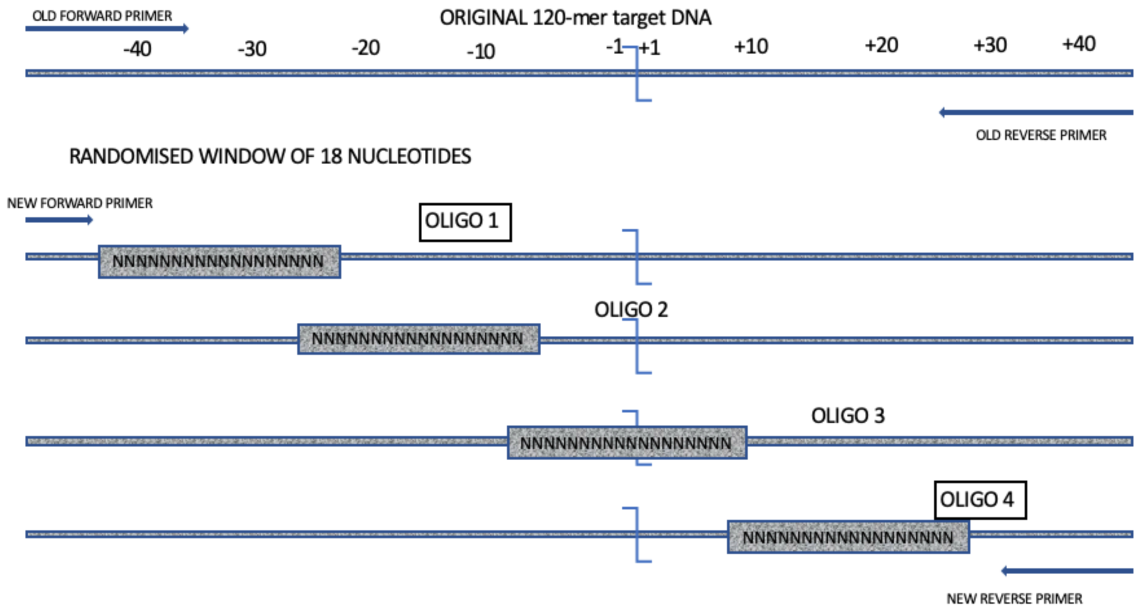

Figure 3. The 28S 120-mer target DNA and the SELEX oligonucleotides. The 120-mer target DNA is shown with the insertion site and the base pairs upstream and downstream. The random sequence window is represented as a grey block with 18 Ns. New forward and reverse primers were constructed as the old forward and reverse primers extend into the random windows. The Oligo 1 and Oligo 4 (black outline box) were studied.

The primer extension mix was used to perform an Electrophoretic Mobility Shift Assay (EMSA) using a delta N KPD/A (endonuclease mutant) protein and the protein binding motif RNA- 3' PBM RNA for Oligo 1 and 5' PBM RNA for Oligo 4. The selection of the DNA aptamers becomes more stringent as the amount of specific RNA increases. We used 4 times more RNA than a usual EMSA reaction for Oligo 4. The EMSA reactions were performed with different dilutions of protein and the lanes with about 60-70% bound complexes were chosen for a scale up EMSA reaction with

5X DNA, 5X RNA and 7.5X protein. The Scaled Up EMSA was run on a 5% polyacrylamide gel and the wet gel was exposed onto a radiograph film to view the bands. This film, after development, was aligned with the wet gel and the well complex, free DNA complex and the Bound RNP complexes were acquired and let sit in crush and soak buffer overnight. This was followed by chloroform extraction and ethanol precipitation. The concentration of DNA in each complex was estimated and was resuspended in 1X TE to make a stock of ~ 1fmol/µl. This was used as a template for amplifying the product. The bound complexes were amplified using ^{32}P radiolabeled primers for further rounds of selection. All extracted complexes were PCR amplified using cold primers (non- radiolabeled) for further sequencing.

5 rounds of selection were done for the Oligo 1 and two rounds for Oligo 4. The data has been presented below.

RESULTS

The selection from initial primer extension of Oligo 1 (Round 1 in Figure 4) was performed using delta N KPD/A protein and 3' RNA. This was followed by gel extraction and amplification of the bound complexes for a second round of selection. Successive rounds of selection were carried out for four rounds (Rounds 1, 3 and 4 are shown in Figure 4). DNA from the bound fraction was sequenced for round 4, but no selection of specific sequences was observed (data not shown). For Oligo 4, the selection from the initial primer extension was performed in the presence of 5' PBM RNA. The weak bound DNA band from round 1 was isolated and DNA was extracted for further rounds of selection. The bound complexes were more prominent for the second round (Round 2 in

Figure 4) of selection, the DNA from the second round was extracted and sequenced to acquire LoGos (Figure 5).

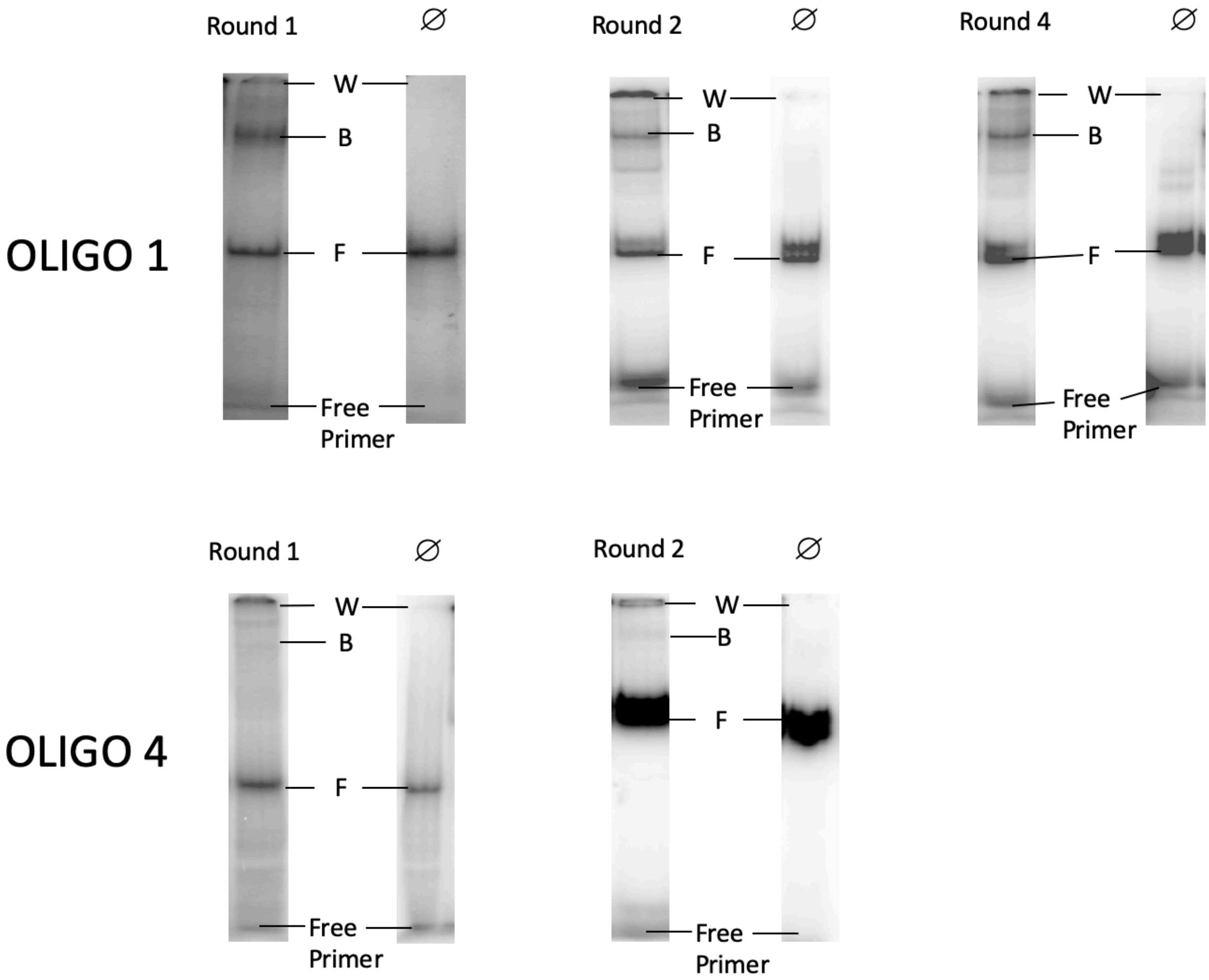

Figure 4. Band shift assays for Oligo 1 and Oligo 4. Tight binding R2 RNP complexes from the Oligo 1 were selected and amplified for further selection. This was repeated for 5 rounds and some band selections are shown. Two rounds of selection for Oligo 4 and the band selections are shown.

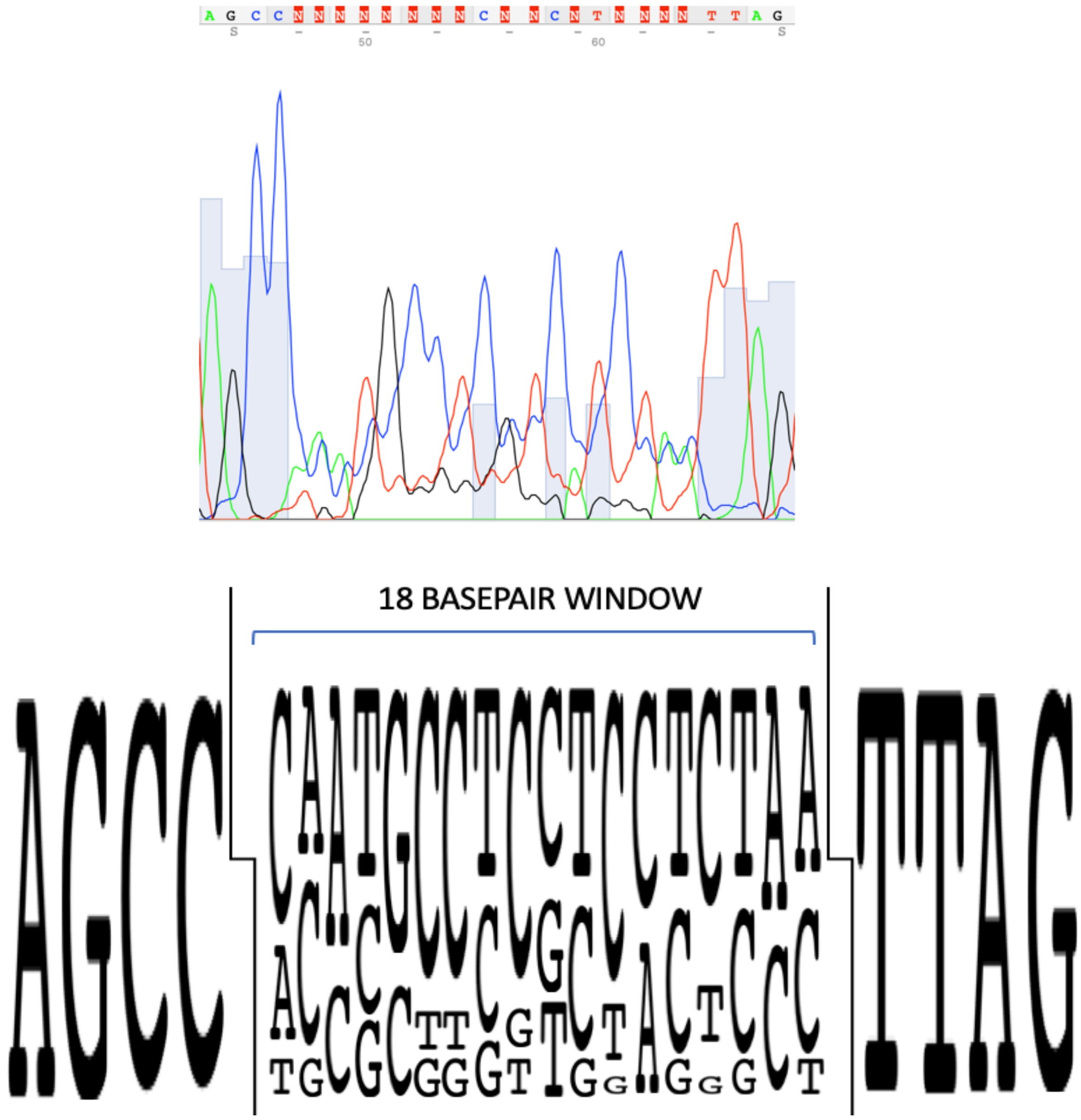

Figure 5. DNA Electropherogram and sequencing LoGos for round 2 selection of Oligo 4. Oligo 4 round 2 selection DNA electropherogram with the DNA base sequence and the corresponding LoGos has been depicted. The Blue peaks are Cytosine, the black peaks are Guanine, the red peaks are Thymine, and the green waves are Adenine. The LoGos was made based on the size of the peaks. The 4 bases to the left and right of the random window Logos correspond to the 28S target 120-mer sequence.

DISCUSSION

The random sequence window for target DNA 1 (Oligo 1) extends from -42 to -25 which roughly corresponds to the area of the target DNA that DNA footprint studies indicate that the R2 protein binds to in the presence of the 3’ PBM RNA, particularly prior to first-strand DNA cleavage [8]. Unfortunately, upon sequencing the final round of selection for target DNA 1, no identifiable LoGos was apparent upon (sequencing data not shown). Possible reasons for the lack of selection might be that a lower than expected concentration of 3’ RNA was used in the initial rounds of selection on target DNA 1 (Oligo 1). The RNA concentration was increased in the later rounds. In addition, the forward primer was much shorter than the reverse primer in the PCR reactions used to amplify back up the selected DNA after each round of selection. It turned out the short forward primer was inefficient. The longer primer was more efficient and made an excess of one strand. These single stranded DNAs can be seen running just above the free double stranded DNA in the EMSA gels (Figure 4). The single stranded DNA associated (non-specifically) with the R2 protein, increasing the noise during selection. The same situation (an efficient primer paired with an inefficient primer) also existed for the PCR primers used to amplify Oligo 4. Why the single stranded DNA did not impair the studies on target DNA 4 (Oligo 4; see above) as it appears to have done with target DNA 1 (Oligo 1) is not clear at this point. It may be that the R2 protein is less able to bind to the single stranded DNA in the presence of 5’ PBM RNA.

The random sequence window for target DNA 4 (Oligo 4) extends from +7 to +24 which roughly corresponds to the area of the target DNA that DNA footprint studies indicate that the R2 protein binds to in the presence of the 5’ PBM RNA [6]. Even though only two rounds of selection were performed due to time limitations, a nice LoGos was obtained upon sequencing the round 2 selected DNA. The number of residues that appeared to be under selection within the 18 bp randomized window was more extensive than we were expecting. The reason and/or implications for so many positions within the randomize window being selected by the R2 protein upon binding will require follow up and repeat experiments. That said, there was a large degree of agreement between the LoGoS data and published DNA footprint data (missing nucleoside and methylation interference footprint data) for the R2BM amino terminal polypeptide [6]. It is known that the Myb domain interacts with the sequences downstream to the target insertion site [5].

The part of the R2 protein used to bind to upstream target DNA sequences is largely unknown [7,8]. It was hoped that comparison of the sequence logos of Oligo 1 and Oligo 4 might help us understand the DNA interactions in the upstream target sequences. If the sequences selected by the protein are similar for both the oligonucleotides, it can be concluded that the Myb domain is responsible for the DNA interactions upstream and downstream. I plan to repeat the studies described in this chapter after modifying the procedure to fix some of the identified procedural problems identified above. I also plan to extend the study to include target DNAs with randomized windows from oligos 2 and 3.

REFERENCES

1. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**, 793-805 (1999).
2. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011 (2015).
3. Kojima, K. K. & Fujiwara, H. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* **22**, 2157-2165 (2005).
4. 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).
5. 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).
6. 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).
7. 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).
8. 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).