A BIOCHEMICAL APPROACH TO DETERMINE THE TARGET SITE RECOGNITION MECHANISM OF THE R2 RETROTRANSPOSABLE ELEMENTS

by

BLAINE K. THOMPSON

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

May 2011

ACKNOWLEDGEMENTS

I would like to formally thank the members of my masters thesis committee for their intellectual, physical, and fiscal contributions to the work performed during the completion of my masters degree. Dr. Shawn Christensen co-principal investigator, Dr. Cedric Feschotte co-principal investigator, Dr. Thomas Chrzanowski, Dr. Ellen Pritham, and Dr. Subhrangsu Mandal. This work was funded by The University of Texas at Arlington and the National Science Foundation.

I would also like to acknowledge Dr. Ramond Jones and the genome core facilities of the UTA Biology department and Clement Gilbert, a post doctoral associate of the Feschotte lab, for their intellectual contributions to this work.

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

April 18, 2011

ABSTRACT

A BIOCHEMICAL APPROACH TO DETERMINE THE TARGET SITE

RECOGNITION MECHANISM OF THE R2

RETROTRANSPOSABLE

ELEMENTS

Blaine Kealyn Thompson, M.S.

The University of Texas at Arlington, 2011

Supervising Professors: Dr. Shawn Christensen and Dr. Cedric Feschotte

Non-LTR Retrotransposons (NLRs) are selfish mobile genetic elements which parasitize the genomes of many organisms including humans. These elements transpose through an RNA intermediate and integrate directly into their genomic site using reverse transcription, a process called Target Primed Reverse Transcription (TPRT). Members of this family of transposons are abundant in the genomes of many eukaryotes including mammals, reptiles, fish, and insects. Several NLR family members such as R2 and NeSL-1 avoid causing deleterious mutations by specifically targeting repetitive genomic loci such as the 28S ribosomal DNA gene and the Spliced leader-1 exon respectively. This literature review focuses on the integration mechanism and evolution of NLRs, specifically R2.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

CHAPTER 1

LITERTURE REVIEW

1.1 Abstract

Non-LTR Retrotransposons (NLRs) are selfish mobile genetic elements which parasitize the genomes of many organisms including humans. These elements transpose through an RNA intermediate and integrate directly into their genomic site using reverse transcription, a process called Target Primed Reverse Transcription (TPRT). Members of this family of transposons are abundant in the genomes of many eukaryotes including mammals, reptiles, fish, and insects. Several NLR family members such as R2 and NeSL-1 avoid causing deleterious mutations by specifically targeting repetitive genomic loci such as the 28S ribosomal DNA gene and the Spliced leader-1 exon respectively. This literature review focuses on the integration mechanism and evolution of NLRs, specifically R2.

1.2 Genome Variation

All organisms, excluding viruses, are evolutionarily divided into the three domains of life: eubacteria, archaea, and eukarya. Eubacteria and archaea (prokaryotes) are among the most ancient or basal level organisms still in existence. It is believed that these organisms are derived from the last universal common ancestor (LUCA) which had emerged from the previous less complex forms of life. The eukaryote lineage is hypothesized to have emerged from the fusion of a eubacteria and an archaea genome, leaving remnants of genes from both lineages within the nuclear and plastid genomes of present day eukaryotes.

Genomes of organisms from the three domains of life and viruses are organized very differently. Viruses typically have very compact genomes with a significant amount of coding capacity and many overlapping open reading frames in both sense and antisense directions.

Bacterial genomes tend to also have compact genomes with a high coding capacity and contain no spliceosomal introns. Eukaryotic genomes are among the most diverse and gene poor genomes of all living organisms. Some eukaryotic genomes contain as little as 1% total protein coding sequence in their entire genome. Eukaryotic genes are usually interrupted by many introns that must be spliced and removed after transcription. Eukaryotic genomes also vary dramatically in size in relation to viruses and bacteria. Figure 1.1 illustrates the dramatic variation in genome size and protein coding capacity among these four groups of organisms.

Among the simplest and most compact genomes are those of viruses. Viral genomes may be composed of DNA or RNA, and may have one or many segments (*1)*. The average viral genome is composed of approximately 20,000 nucleotides (nt) encoding 10-40 genes. Viruses can be extremely variable in their mechanisms of replication and genome characteristics. Extreme viral genomes may encode as little as 1500 nt and 2 open reading frames, while the mimivirus genome contains upwards of 1.2Mb and 1262 genes (*2)*. There is nearly a one thousand fold variation in viral genomes discovered to date.

Bacteria are single celled prokaryotes with small gene dense chromosomes composed of dsDNA. Most bacteria have a single chromosome of approximately 4Mb, but there are a few species which have a second linear chromosome. Bacteria may also utilize very small extrachromosomal circular DNA molecules called plasmids. Multigene pathways are organized into operons, polycistronic units that are regulated and transcribed together. Within these highly evolved genomes there is room for variability. Obligate parasites have minimal genomes of ~150kb and ~200 genes while the largest sequenced bacterial genome contains over 13Mb and encodes ~9000 genes. Bacteria are less variable than viruses with only a 100 fold range in genome size.

Archaea are small prokaryotes that use histone like proteins to package their DNA. The current status of knowledge on the domain archaea is lacking; less than 100 species have been sequenced. Of the genomes that have been sequenced, archaea are the least variable in terms

of size and coding capacity. Archaea genomes range from 500kb to over 6Mb and may encode 500-5000 genes. These genomes are slightly less gene dense than bacteria and viruses with 7-27% of the genome encoding for no protein.

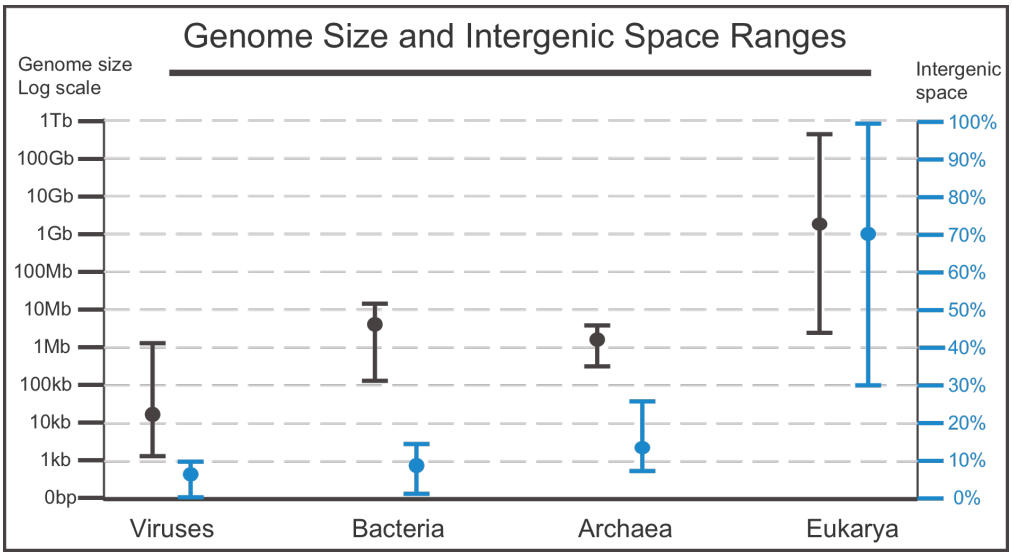


Figure 1.1: **Genome Size and Intergenic Space Variation**. Shown above is a plot of genome size and intergenic space. Left axis is a log scale starting at 1kbp to 1Tbp, genome size data is plotted in black lines. Right axis represents the fraction of intergenic space in percent, this data is plotted in blue. X-axis represents viruses, bacteria, archaea, and eukarya. Large dots represent the approximate means, bars represent the range of data. All data collected from ensemble genomes site.

Eukaryotes are regarded as having the most complex genomes among living organisms. Unlike prokaryotes, eukaryotes have multiple linear chromosomes. The haploid, or single copy, genome size may range dramatically between seemingly similar organisms. The average eukaryotic genome contains approximately 2Gb of DNA encoding some 20,000 genes. Even though gene number in eukaryotes is high, their coding capacity averages about 30%. Eukaryotic gene regulation is more complex than that of bacteria and viruses. Genes are spread out like islands over an ocean of non-coding DNA. In some extreme eukaryote genomes only 1% of the total DNA encodes for protein. Not all eukaryotic genomes are giant gene desserts. Obligate parasitic eukaryotes often have extremely small genomes, similar in size to that of

bacteria. Eukaryote genomes are the most diverse with nearly a 200,000 fold size difference among organisms sequenced to date.

Genome composition is complex and dynamic, constantly undergoing change and adaptation. These selective pressures are present across the tree of life, so what is it that makes the genomes of eukaryotic organisms so variable?

1.3 Transposable Elements and Eukaryotic Genomes

Transposable elements (TEs) are one of the major forces driving eukaryotic genome variation. TEs have been found in most of the sequenced genomes, but they are especially prevalent in eukaryotes. TE repeats may compose a significant portion of the total genomic DNA with approximately 52% of the human genome being recognizably derived from TEs. One family of elements has reached over 1 million copies in Humans (*3-6)*. TEs account for some of the major variations discovered within eukaryotic genomes (*7-9)*.

Transposons are similar to viruses in that they parasitize a host in order to copy themselves with no immediate/direct benefit to their host. Where TEs differ from viruses is that they are incapable of exiting their host cell, therefore, TEs may only persist through genomic integrations. TEs are composed of many families and groups which are divided by their mechanism of mobilization and integration into genomic DNA. There are two primary classes of transposons, those which mobilize through an RNA intermediate and those which mobilize through a DNA intermediate (historically class I and class II mobile genetic elements, but here are referred to as retrotransposons or DNA transposons respectively).

When transposons land within protein coding sequences, developmentally important regions, or regulatory networks they often lead to disease causing mutations. One family of retrotransposons is responsible for ~1:1000 mutations in humans (*10, 11)*. In addition to disrupting coding genes or regulatory sequences during transposition, TEs also create homologous sites throughout the genome. Unequal crossing over or non-homologous

recombination could lead to the deletion or duplication of large sections of host chromosomes. Chromosomal aberrations and other mutations make TEs a constant source of genome variability.

Over the hundreds of millions of years that eukaryotic genomes have co-existed with TEs, they have evolved mechanisms to defend themselves. Cellular defense enzymes and RNA interference play significant roles in TE suppression (*12-22)*. Chromatinization and epigenetic regulation of TEs is also a common mechanism to limit transposition. It is not advantageous for the TE to destroy its host, instead well adapted TEs have found a way to be minimally invasive on their host genome allowing them to remain largely undetected.

Genomes not only defend themselves against TEs, in some cases they exapt TEs to perform important cellular functions (*23)*. Old TEs may serve as templates for RNAi defense and TE silencing (*21, 24)* (*25)*. TE derived proteins perform a multitude of cellular functions including histone modification, telomere maintenance, and DNA binding (*25)* (*26, 27)* (*28-30)* (*31-33)*. Even the jawed vertebrate adaptive immune system is aided by TE derived proteins (*34, 35)*. A few TEs such as Piggybac and Sleeping Beauty are even being manipulated in the laboratory and medical settings to serve as potential gene therapy vectors (*36-38)*.

Transposable element evolution has become an area of viable interest to transposon and evolutionary biologists. A single genome may contain a large number of differing transposons. Some families or groups of transposons can acquire large copy numbers in the genome. Copy number will depend on how long the element is active and how long it takes the genome to suppress the element activity. Several groups of transposons show bursts or periods that the elements were active and mobile, others just date to a single period of time. TEs have been maintained in host genomes through the ages either by vertical inheritance or through horizontal transference (*39-42)*. Vertically transmitted elements must stay active to remain within a genome and may only be passed from a host to its progeny (*43-45)*. TEs may be

horizontally transferred from one host to another by some type of vector (46). This allows the TE to escape genomic conflict by constantly parasitizing a new host genome (*40, 46, 47)*.

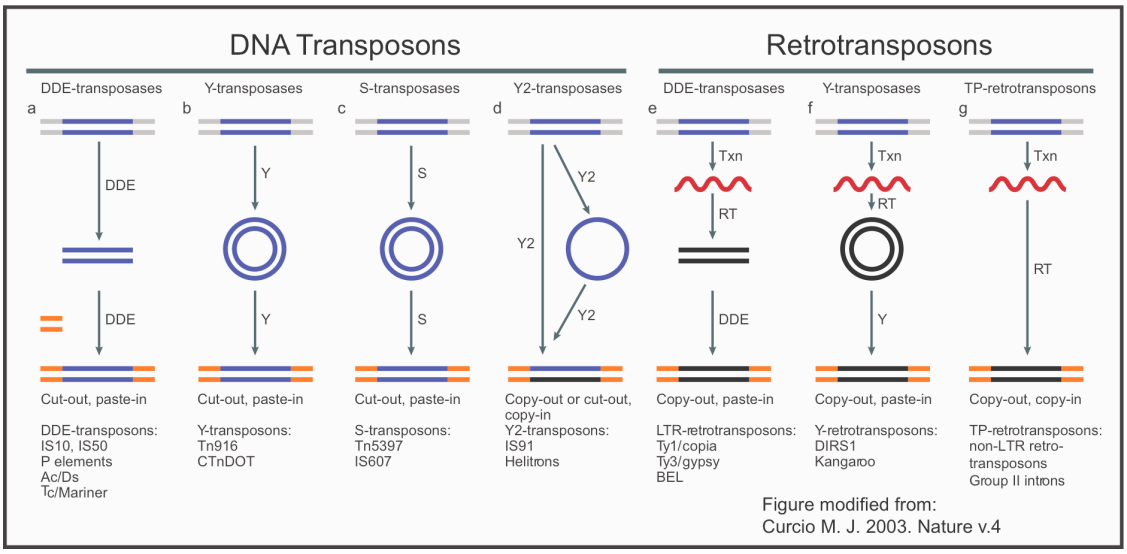


Figure 1.2: Transposable Element Mechanisms. Shown above are the general transposition mechanism of DNA transposons (a-d) and Retrotransposons (e-f). Blue horizontal bars represent TE DNA, gray bars represent flanking target DNA, red zig-zag lines represent RNA intermediate, blue circles represent double (two) or single (one) stranded DNA intermediate, Orang horizontal bars represent flanking genomic target site, and Black horizontal lines represent newly synthesized TE DNA.

Transposons mobilize themselves in a multitude of ways, using many different enzymes, mechanisms, and intermediates illustrated in figure 1.2. DNA transposons mobilize through a DNA intermediate only where the RNA serves only as the template for protein translation. DDE transposases perform trans-esterification reactions which “cut-out” or “paste-in” DNA. Transposase protein binds the element, excises the sequence, and pastes it into its new genomic site (1.2.a). P-elements and members of the Tc/Mariner superfamily utilize this type of transposase. Tyrosine (Y) recombinase may also cut and paste DNA from one site to another using a 3‘-phosphotyrosine linkage. Y-transposons employ a tyrosine recombinase to excise the transposon sequence, a consequence of this excision is a circular dsDNA molecule. Reversal of these catalytic steps results in transposon insertion (1.2.b). S-transposons encode a serine recombinase which functions to cleave DNA using a 5’-phosphoserine intermediate. The

catalytic steps of the S-transposons are identical to those of the Y-transposons (1.2.c). Y2-transposons are often referred to as “rolling circle” transposons. These elements are either cut or copied out of place to form a circular ssDNA intermediate. The circular DNA template is hypothesized to be copied into its target site as it rolls, creating head to tail tandem insertions (1.2.d). Helitrons and IS91 are thought to mobilize through this mechanism.

Retrotransposons mobilize through an RNA intermediate that must be reverse transcribed and integrated into DNA. The reverse transcriptase may catalyze RNA templated DNA synthesis and is the key component of all retrotransposons. Long terminal repeat retrotransposons are reverse transcribed in a concerted manner using their direct long terminal repeats (LTR) to create a dsDNA intermediate. A DDE integrase then integrates the DNA intermediate into the genome (1.2.e). Y-retrotransposons reverse transcribe their RNA into a circular dsDNA intermediate which is then inserted into the genomic site using a tyrosine recombinase enzyme (1.2.f). Lastly, target primed (TP) retrotransposons integrate directly into genomic DNA using a free 3’-OH. Element RNA is used as template to perform cDNA synthesis, a process called target primed reverse transcription (TPRT) (1.2.g). Non-LTR retrotransposons (NLR) and group II mobile introns use this TPRT integration mechanism (*48)*.

1.4 Reverse Transcriptase

Due to its distribution and diversity across the tree of life, it is believed that a reverse transcriptase (RT) was present in the last universal common ancestor (LUCA) and may have played a pivotal role in the emergence of the DNA world from the RNA world. There are 10 phylogenetically distinct clades of RT containing elements to date (excluding RNA dependent RNA polymerases): telomerase (TERT), group II introns (GII), diversity generating retroelements (DGR), non-LTR retrotransposons (NLR), long terminal repeat retrotransposons (LTR), retrons, mitochondrial retroplasmids (MtP), penelope elements (PLE), hepadnaviruses (HEP), and retroviruses (*45, 49)*. There are some RT containing elements found in bacteria

which do not fit into these groups, however the function and mechanism of these elements are currently unclear and have been excluded from this representation (*50)*. Figure 1.3 shows the relationship between RT containing groups (retroviruses are shown in blue and were added to this figure for clarity). Each of these major groups represents a set of elements which use the reverse transcriptase enzyme to perform different functions. The evolutionary ancestry of the reverse transcriptase gene is difficult to resolve and is often a topic of conflict between biologists (*51)*.

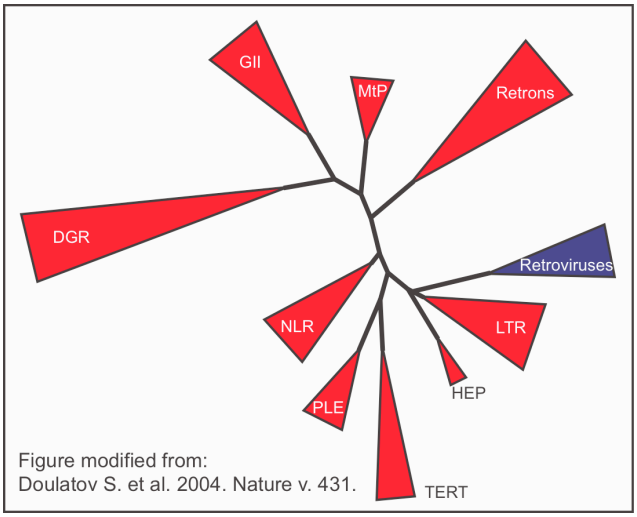


Figure 1.3: Reverse Transcriptase Tree of Life. An unrooted phylogram of reverse transcriptase containing sequences. TERT - telomerase, GII - group II introns, DGR - diversity generating retroelements, NLR - non-LTR retrotransposons, LTR - long terminal repeat retrotransposons, MtP - mitochondrial retroplasmids, PLE - penelope elements, HEP - hepadnaviruses, retroviruses, retrons. This tree excludes RNA dependent RNA polymerase, and newly discovered and ungrouped bacterial reverse transcriptases.

Telomerase is a cellular reverse transcriptase enzyme found in diverse eukaryotes, it is responsible for replicating the ends of linear chromosomes in many species (*53, 54)*. Without this essential enzyme linear chromosomes would shorten until protein coding genes may potentially be deleted during replication. A short Telomerase RNA template is transcribed and binds to the telomerase enzyme, the free 3'-OH group from the exposed end of the

chromosomes primes TPRT (*51, 54)*. Telomeres maintained by telomerase are typically composed of short tandem repeats of the RNA template sequence.

Mobile introns may have been amongst the first genomic invaders. Found within bacterial genomes and eukaryotic organelles, group II introns catalyze their own excision from primary transcripts. These self splicing ribozymes encode a functional reverse transcriptase gene and integrate via TPRT (*55, 56)*. As a primary transcript containing a group II intron is produced, the self-splicing intron RNA forms secondary and tertiary structures. These structures are catalytic and facilitate a trans-esterification reaction between a 2'-OH group on the 3' end of the intron and the 5' splice site excising the intron from the transcript. This reaction is completed by ligation of the exons which releases the lariat folded intron. The intron may then reverse splice into a new acceptor site either using the catalytic RNA or associated maturase proteins. Newly inserted introns must be reversed transcribed by the self encoded RT to complete a transposition event (*52)*.

Diversity generating retroelements (DGRs) are diversity generating cassettes which mutagenize the reverse transcription of a template repeat into its target, the variable repeat, within a protein coding gene in order to generate diversity (*49, 57)*. Template mutagenesis is key to this diversity generation and is performed during TPRT. The *Bordetella bacteriophage* makes use of its DGR to change tropism with its host, *Bordetella* species, as it transforms from free living to pathogenic during mammalian respiratory tract invasion (*57)*. This amazing discovery is not unique to the *Bordetella bacteriophage*, DGRs have been identified in over 30 genomes including bacteriophage and bacteria (*57)*.

Penelope elements (PLE) are a group of retrotransposons which encode an RT very divergent from its ancestral copy. These elements are highly variable and may encode direct or indirect LTRs and are thought to mobilize through a TPRT mechanism (*58)*. PLE elements have been found to mobilize introns and maintain telomeres (*26, 59)*. Many PLEs encode a Uri endonuclease with sequence similarity to GIY-YIG endonuclease of bacterial group I introns

and bacterial DNA repair enzymes (*60)*. Although RT and endonuclease activity has been shown for some PLEs, TPRT has not yet been demonstrated (*61)*.

Non-LTR retotransposons integrate through a TPRT mechanism and are the focus of the remainder of this document. Integration mechanisms will be discussed in greater detail below.

1.5 Non-LTR Retrotransposons

Non-LTR retrotransposons (NLR) are among the most successful groups of TEs in insects and mammals. NLRs were first identified as active in humans when an element transposed causing a mutation which led to the disease hemophilia. Analysis of NLRs has revealed several distinguishing features. NLRs lack the long terminal repeats which flank LTR elements. The reverse transcriptase enzyme of NLRs forms its own clade on the RT evolutionary tree. There are two primary domain structures of NLRs. First, those which encode a single open reading frame with a central reverse transcriptase and carboxyl terminal restriction like endonuclease (RLE) similar to *FokI* (*41, 62, 63)*. Second, NLRs which encode two open reading frames with an Apurinic/Apyrimidinic endonuclease (APE) and reverse transcriptase in ORF2 (*45)*. All NLRs share a cysteine histidine rich motif (CCHC) in their carboxyl terminus. The RLE bearing elements are phylogenetically early branching compared to APE bearing elements and are usually site specific. APE bearing elements branch later in the NLR tree and typically insert in non-specific sites.

NLRs transpose using the target primed reverse transcription reaction (figure 1.4). In this mechanism NLR protein binds a copy of its transcript and enters the nucleus. Once the protein-RNA complex has bound to its target site, a self encoded endonuclease generates a nick in the DNA. This 3'-OH primes TPRT (*45, 64-66)*. Although the initial steps of TPRT are well characterized, the final steps of integration, mainly second strand synthesis and 5' integration, remain largely unresolved. It is hypothesized that two sub-units of protein catalyze

the reaction (*65)*. The endonuclease domain from the second sub-unit could cleave the second strand of DNA at the target site. Second strand synthesis could also be performed by this second sub-unit, catalyzing DNA synthesis using the cDNA from TPRT as template and displacing the RNA from the RNA/DNA hybrid during polymerization. The gaps created by the staggered endonuclease cleavage site are hypothesized to be repaired by host machinery (*65)*.

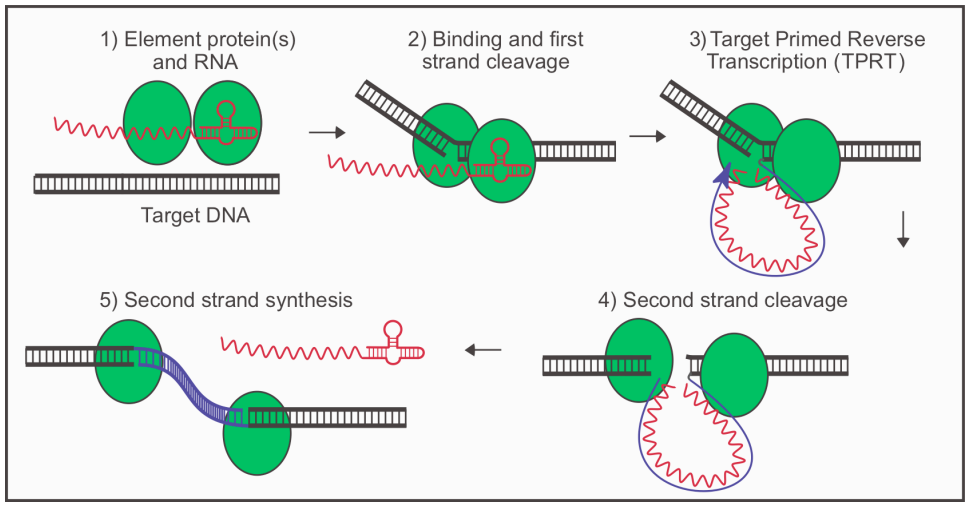


Figure 1.4: Target Primed Reverse Transcription (TPRT). A diagram of Target Primed Reverse Transcription mechanism. Non-LTR retrotransposons transpose through a target primed reverse transcription (TPRT). During TPRT an RNA template is bound to the dimer of self encoded protien (1). Upon binding RNA, the downstream subunit nicks the bottom strand of target site DNA (2). Once the 3' OH group is exposed the RT domain makes a cDNA from the RNA template (Target primed reverse transcription, 3). The second strand is cleaved (4) and synthesized (5) through an unkown mechanism.

Whether NLRs have high or low copy numbers in a genome, only a small fraction of insertions are full length. Many elements are 5' truncated, one of the hallmarks of the TPRT integration mechanism, which could occur during TPRT or during 5' integration. TPRT is speculated to stall or terminate prematurely before the reaction reaches the 5' end of the element sequence. Target site duplications or deletions are also associated with TPRT integrated elements. Many NLRs have poly A stretches at their 3' ends and often mobilize sequences not of their own origin. Short interspersed nucleotide elements (SINE) often steal NLR machinery in order to selfishly copy themselves. Cellular mRNAs may be captured by NLR

proteins and retroposed, potentially creating new retrogenes and psuedogenes. NLRs may also transduce sequences or genes that previously flanked the parental NLR element. NLRs are thought to contain internal promoters.

NLRs are divided into groups based on phylogeny, domain structure, and target site. Phylogenies of NLRs are constructed using their conserved reverse transcriptase gene. The NLR reverse transcriptase gene has 10 highly conserved amino acid motifs named domain 0-9 (domains 0 and 9 are unique to NLRs) (*41, 67)*. Figure 1.5.A shows the phylogenetic relationship of the major groups of NLRs discovered to date. Figure 1.5.B illustrates the domain structure of the best characterized element of each group, and figure 1.5.C lists the names of the represented elements and their preferred target site. The RLE bearing elements are depicted with thick black lines in panel A, members include Genie, CRE, R2, R4, and NeSL. R5 has two open reading frames but still employs a type II RLE and specifically targets the 28S gene. These elements tend to insert specifically into genes found in multiple copies in the genome to reduce fitness cost to the host. Elements which utilize the APE in this tree are depicted with thin black lines in panel A. These elements are mostly composed of 2 ORFs, they contain an APE endonuclease and many have a carboxyl terminal RNAse H domain. Many of these elements are not site specific and include the families L1, RTE, LOA, Tad1, Jockey, and I factor. R1 elements have a similar domain structure to all these elements except they are site specific for the 28S gene. The dichotomy of this tree reveals three interesting facts. Younger elements tend to be non-specific in nature, they employ an AP endonuclease, and they are encoded by two ORFs. Older elements tend to utilize a single ORF, an RL endonuclease, and tend to insert into specific repetitive loci. Found in the center of this tree is an element called DualEN, so named for its two endonucleases. This element encodes a single ORF containing an APE in its amino terminus and an RLE in its carboxyl terminus, much like other RLE bearing elements. It is possible that DualEN represents a transitional element from the RLE to APE bearing groups (*27, 30, 41, 45, 58, 63, 67-84)*.

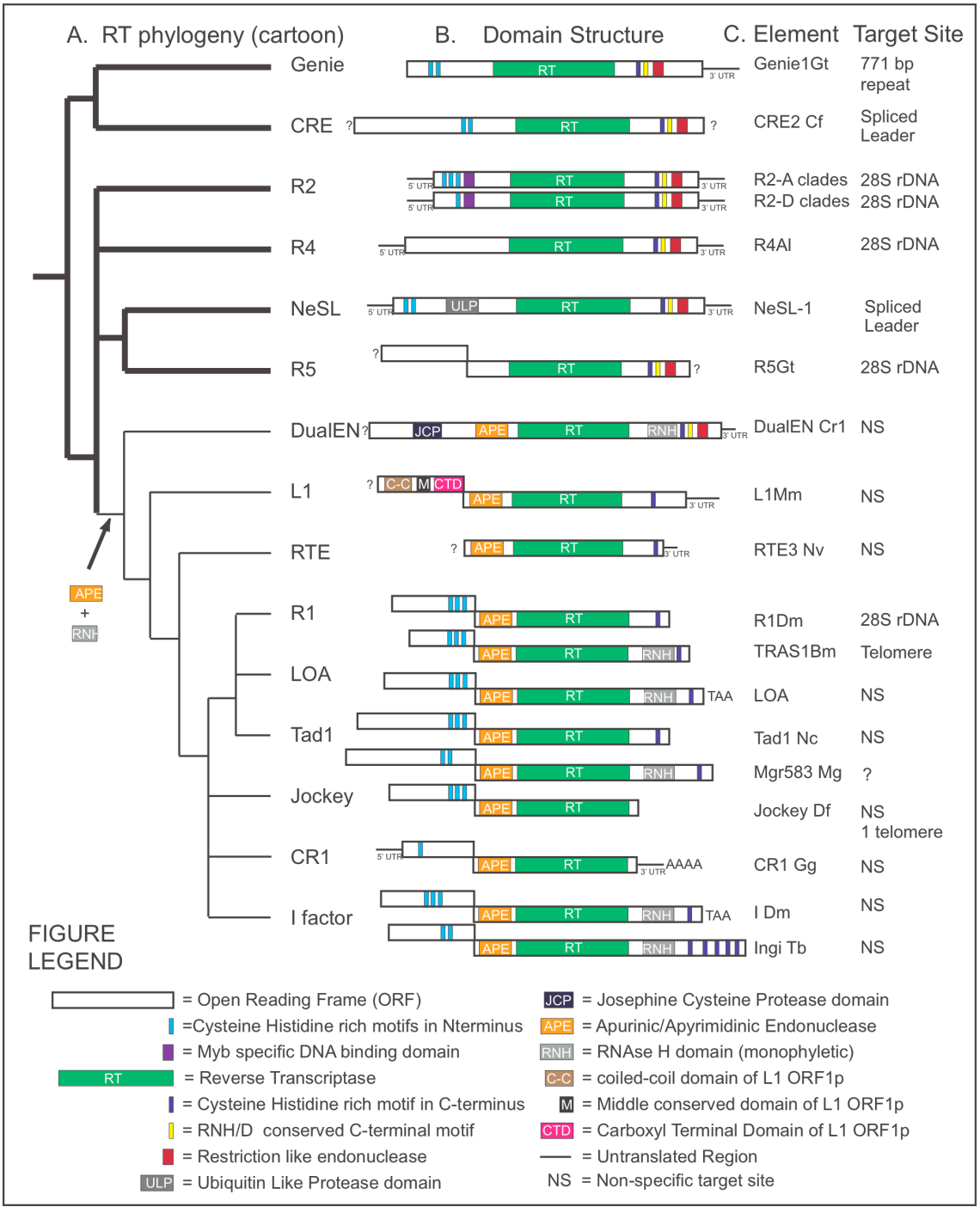


Figure 1.5: Phylogeny and Structure of Non-LTR Retrotransposons. Panel A shows a phylogeny of NLR elements, panel B shows element structures, and panelC shows element names and target sites. A figure legend of domain structures is shown below. *(147-169)*

1.6 APE Bearing Non-LTR Retrotransposons

The model element of the APE bearing group is the long interspersed nucleotide element 1 (LINE1 or L1). Note: the term LINE is often mistakenly used as a synonym for NLR or TP retrotransposon, here LINE will only be used to refer to the mammalian L1 element. The Line family of elements is one of the most successful members of NLRs. Human LINE1 (L1Hs) is the only autonomous TE known to still be active in the human genome. L1 mediated transposition events can cause disease and are responsible for ~1:1000 mutations in humans (*10, 11)*. L1 elements are abundant in mammalian genomes but have been found in fish, reptiles, plants, slime molds, and algae (*41)*.

Evolutionary analysis shows that LINEs are vertically inherited, although there is some evidence of infrequent horizontal transfer (*41, 85)*. L1Hs elements tend to be inserted into A+T rich regions allowing for nearly random insertion in the genome (*86)*. Coupling of random insertions with genomic duplications and deletions by recombination makes L1 such a potent source of mutation within the human genome (*10, 87)*. In-vitro studies of L1 ORF2 activity have had little success, however an excellent *in-vivo* transposition assay has been pioneered by John Moran. *In-silico* and *in-vivo* sequence analysis of L1 insertions reveals several interesting facts. L1 elements are typically flanked by short target site duplications, have an extended poly A tract at the 3' end, and target non-specific sites. These hallmarks indicate conservation in the DNA cleavage mechanism, reverse transcription of processed mRNAs, and a lack of target sequence specificity.

L1 elements are composed of two slightly overlapping ORFs. ORF I is composed of three conserved motifs, a coiled coil (C-C) domain, a central conserved (M) domain, and a carboxyl terminal domain (CTD). The C-C domain was recently shown to facilitate protein-protein interactions and deletion analysis of this domain abolished the proteins ability to form competent RNPs (*88)*. The CTD has been shown to bind RNA with high affinity and exhibit

nucleic acid chaperone activity associated with single stranded nucleic acids (88-92). The exact function of the M domain remains unclear. ORF1 protein, which is essential to, but insufficient for, retrotransposition is required for proper RNP formation (*88-92)*. The crystallization of ORF1p from Line1 reveals that the CTD acts as a trimer and confirms its suspected nucleic acid chaperone activity (*91, 93, 94)*.

ORF2 contains many regions of highly conserved amino acids. An APE domain occurs at the amino terminal end shortly followed by the reverse transcriptase. The APE domain of L1 has been shown to cleave its target site, priming it for TPRT (*96)*. Several structural studies have shed some light on the properties and activities of NLR APEs. The crystal structures of the APE endonuclease of TRAS and R1Bm confirm their phylogenetic identification and provide a detailed mechanism for how these endonucleases function to liberate their 3'-OH. Although R1 and Tras are site specific, they share a domain structure similar to L1. L1 usually prefers an A+T rich site, however swapping the APE of Line1 for that of another element changes this target site preference (*97)*. Cellular AP endonucleases are nickases and cannot perform double stranded cleavage.

Central to ORF2 is the reverse transcriptase domain whose activity was confirmed using an in-vivo transposition assay in multiple human cell lines (*94, 98-101)*. The properties of this reverse transcriptase have not been fully explored, but it is believed to be highly processive (*102)*. The carboxyl terminal end of the element encodes a cysteine histidine rich zinc finger like motif followed by a highly conserved RHN/D motif that is common to all NLRs. These highly conserved motifs are likely involved in RNA binding (*83, 103)*.

L1 integrates through a TPRT mechanism. An RNA transcript of the L1 element is translated and the newly translated protein shows cis preference for its own RNA (*104, 105)* (*106)*. At least one sub-unit of ORF2p and multiple sub-units of ORF1p bind the transcript (*93, 94)*. This integration competent RNP particle then seeks its site of integration. The APE endonuclease cleaves the bottom strand of the target DNA which primes TPRT. It is unclear if

the 5‘ end of the element is integrated through DNA repair mediated processes, recombination, or the L1 protein itself, and is a topic of controversy among LINE biologists. Whether second strand synthesis is performed from the second strand nick using the L1 reverse transcriptase or a cellular polymerase has yet to be answered (*11, 66, 107)*. During active transposition, the L1 element makes many more dsDNA breaks than successful transposition events, and new insertions have highly similar target site duplications (*108)*. This may be compelling evidence for a two sub-unit model of integration where one sub-unit performs bottom strand cleavage and TPRT, and the other sub-unit performs second strand cleavage and second strand synthesis. There are many unanswered questions in L1 biology, none more important than an in depth look at the complete integration mechanism.

Non-autonomous NLRs, often referred to as short interspersed nucleotide elements (SINEs), parasitize or hijack the L1 machinery to selfishly copy their own non-coding sequence. Some of these elements are quite successful. For example the Alu SINE from humans has reached more than one million copies or ~11% of the genome (*109, 110)*. SINEs are typically short ~200-500 bp, do not code for any functional proteins, and may have many subfamilies of varying age and divergence. Alu elements are SINEs derived from a 7SL RNA and can only be found in primates (*111-113)*. SINE integrations tend to be full length due to their short sequence and typically maintain a higher copy number than the autonomous form of the element.

1.7 RLE Bearing Non-LTR Retrotransposons

There are 7 major phylogenetic clades of RLE bearing NLRs. Many of these NLRs are site specific in nature and tend to target repetitive genome sequences as to maintain a higher copy number. By targeting repetitive sequences these site specific transposons can maintain themselves in a genome for a long period of evolutionary time without significantly harming the overall fitness of their host. All RLE bearing NLRs encode a highly similar central reverse transctiptase domain followed by a conserved cystidine histidine rich motif, a basic RHN patch,

and a restriction like endonuclease at their carboxyl terminus. Of These major clades are the ribosome sequence targeting elements such as R2, R4, and R8. R2 elements are described in greater detail below. R4 elements were first discovered in the parasitic nematode *Ascaris lumbricoides*, and may be found in other nematodes. R5 elements are more diverse in their set of hosts which include *Schmidtea mediterranea* and *Girardia tigrina*. R4 elements are translated as a single open reading frame while R5 elements are composed of two open reading frames. The Genie elements are among the most ancient NLRs and are found in the genomes of the genus *Girardia*. These elements contain two zinc finger motifs in their amino terminus and target a 771 nucleotide genomic repeat. CRE elements are very similar to Genie elements but are much longer. These transposons are found in nematodes and target the nematode specific spliced leader sequences. NeSL elements are unique in that they encode a ubiquitin-like protease domain in their amino terminus and target the spliced leader sequences of nematodes. NeSL elements have also been found in a highly divergent host, *Trichomonas vaginalis*. The final RLE bearing element is the previously described DualEN elements of *Chlamydomonas reinhardtii*.

R2 elements are the prototypical RLE bearing NLRs. They insert into the 28S rDNA locus of many animals such as cnidarians, horseshoe crab, fish, reptiles, and insects (*43, 114)*. The R2 element from the silk moth *Bombyx mori* (R2Bm) has served as the model system of NLR biochemical studies and its integration mechanism is likely to be recapitulated for all NLRs with some minor caveats (*64, 115)*. R2 integrates using a TPRT mechanism and shows a long history of vertical inheritance (*64, 116)*. Site specific NLRs target high copy repetitive loci in order to propagate themselves with a limited fitness cost to the host, allowing for these elements to be long term components of their host's genome. Very little biochemical studies have been performed on site specific NLRs with the exception of R2Bm. Therefore, the remainder of this review will focus on the evolution and biochemistry of R2 elements which are likely to be similar for all site specific NLRs.

With the exception of their amino terminal variability all R2 elements share the same domains. R2s encode a single ORF which may contain a variable number of zinc finger motifs followed by a myb domain in the N-terminus (*43)*. The reverse transcriptase domain is central to the ORF, and there are several conserved motifs in the carboxyl terminus including an R/HINALP motif, a cysteine histidine rich motif (CCHC), a charged patch RHN/D, and a site specific type II restriction like endonuclease domain (*67)*. Most full length R2 elements have a single ORF with a 5' and 3' untranslated region (UTR). One exception is a specific R2 element from the jewel wasp (R2Nvit-A) that has two open reading frames and is likely a chimeric R1/R2 element.

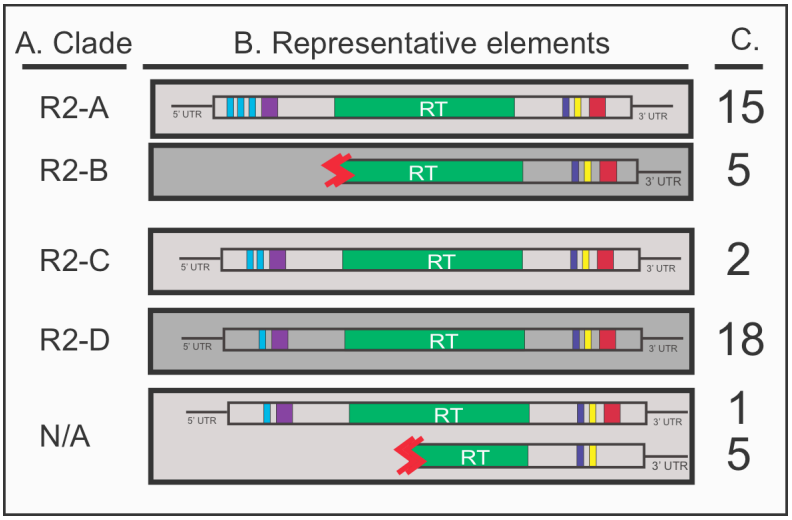


Figure 1.6 R2 Element Domain Structures. A. column A represents the R2 subclade each domain structure belongs to. B. representation of full length element of each subclade. C. column C represents the number of elements (full length and truncated) discovered from each subclade.

Phylogenetic analysis of the reverse transcriptase gene of R2 elements reveals that there are four clades (or sub-clades) of R2 (R2-A, R2-B, R2-C, and R2-D) (*43, 117, 118)*. While most hosts only carry a single type of R2, some genomes harbor elements from more than one of these clades (*43, 119)*. These clades are evolutionarily distinct and monophyletic with respect to their origins and all date back over 500 million years ago. The R2-A and R2-D sub-clades are

by far the most represented groups of R2 elements within the genomes of sequenced organisms (*43)*. Interestingly, the differing sub-clades all have amino terminal domain structures unique to their particular clade (see figure 1.6). Figure 1.6 illustrates the four R2 clades (left), their relative domain structures (center), and the number of elements discovered to date (right). The bottom panel includes element domain structures that have not yet been included in a phylogenetic analysis. R2-A clade members have three zinc finger motifs followed by a myb domain in their N-terminus. No R2-B clade member discovered has an intact N-terminus and R2-C clade members have two zinc finger motifs and a myb domain. R2-D clade members have one zinc finger and a myb domain in their N-terminus. So what drives all this variability in the amino terminal domain of R2? One likely hypothesis is that the common ancestor to modern day R2s may have had three zinc finger motifs and a myb domain (*43)*. As time passed elements which lost one or two of their zinc finger motifs by 5' truncation or mutation adapted to the loss and became competitive with the ancestral element. In some cases the ancestor was forced out of the genome, in other cases both elements were co-propagated within the same host (*43, 120)*. This situation is very likely and feasible, however there are other explanations to how the multiple clades of R2 elements arose. Unfortunately this concerted evolution occurred far back in evolutionary time and it is difficult to discern any clear relationships.

The R2 element's life cycle is more complex than that of the typical TE. Because these elements reside in a highly repetitive loci (the ribosomal repeat unit) they are subject to the same genomic recombination and deletion rates as that locus. The ribosomal locus is an extremely dynamic environment and genomic copy numbers can range dramatically between parents and their progeny. This can be advantageous or deleterious for R2. If unequal crossing over produces a chromosomal duplication then R2 may increase in copy number without retrotransposing. However, if unequal crossing over, or intrachromosomal recombination leads to deletion, then R2 copy number decreases (figure 7, panels A+B).

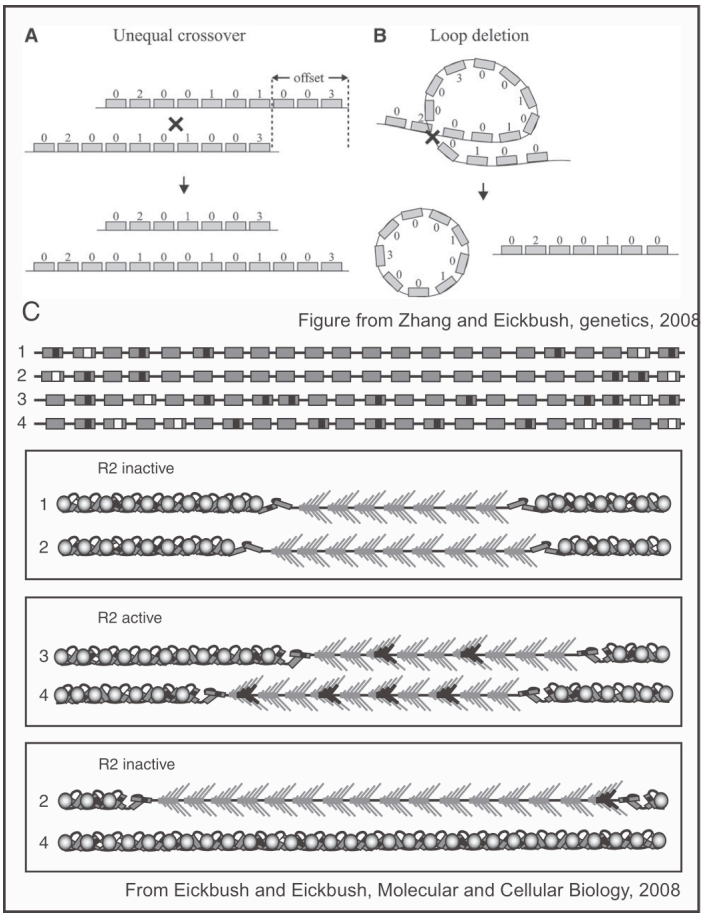


Figure 1.7: R2 Regulation and Ribosome Dynamics. A. Diagram showing ribosomal locus unequal crossing over intrachromosomally. 0= no insertions, 1=R1 insertion, 2=R2 insertion, 3=R1 and R2 insertions. B. Ribosomal locus interchromosomal recombination. C. Upper portion represents ribosomal repeats (gray rectangles) with R1 insertions (white) squares, or R2 insertions (Black squares). Lower panel illustrates R2 regulation by silencing of ribosome locus.

R2 persistence is described as a "running in place" model where the R2 element's transposition is cancelled out by ribosomal recombination, leaving overall genomic copy number approximately equal. Since R2 elements must maintain functionality many of the elements found in the genome are not highly divergent and still have intact ORFs. The ribosomal units are present in ~200 copies but only about 50 functional units are needed for sufficient host function. This allows the host to regulate the expression of ribosomal units with a high copy of TE insertions. As illustrated above in figure 7, as the host ribosomal units that are being transcribed

become inundated with insertions, the genome changes which units are actively being transcribed by repressing expression of TE-overwhelmed areas with chromatinization and begins transcribing different units.

Within the 5' UTR, the ORF and the 3' UTR of the R2Bm transcript are highly conserved RNA structures called protein binding motifs (PBM) (*121-124)* (Figure 1.8a). These motifs are putative binding sites for the R2Bm protein, and coordinate target site binding and integration reaction timing (*65, 121, 122, 125)*. It is believed that these RNA structures provide several other important regulatory roles for the R2Bm protein. First, the R2Bm protein does not seem to dimerize, meaning it shows no protein-protein interactions. Instead, R2Bm forms a psuedodimer ribonucleoprotein particle (RNP) by binding two sub-units of R2Bm protein to one RNA transcript, at the 5' PBM and at the 3' PBM. Second, R2 RNA serves as the timing mechanism during R2 integration, communicating the completion of TPRT and initiating second strand cleavage between the two sub-units of R2Bm protein. Lastly, it has been hypothesized for R2Bm and other NLRs that a form of non-traditional translation may be coordinated by RNA structures (*104, 105)*. The 5'PBM may also serve as an internal ribosome initiation site, which would account for the numerous NLR elements that do not begin translation from the canonical start methionine (*121)*.

R2 elements are thought to carry an internal promoter to initiate transcription like other NLRs, however it is currently unclear whether R2 transcripts are produced by RNA polymerase II or III. Since R2 elements are imbedded within coding region of the 28S rDNA gene they are often times co-transcribed with ribosomal genes presumably by RNA polymerase III (*126)*. It was recently shown that the 5' end of the R2 element RNA includes a self-cleaving ribozyme (*127)*. This ribozyme has sequence similarity to the *hepatitis delta virus* RNA cleaving ribozyme. It is now believed that R2 elements are co-transcribed with rRNA genes and cleave their transcripts free via self-cleavage as illustrated in figure 1.8.b (*127)*.

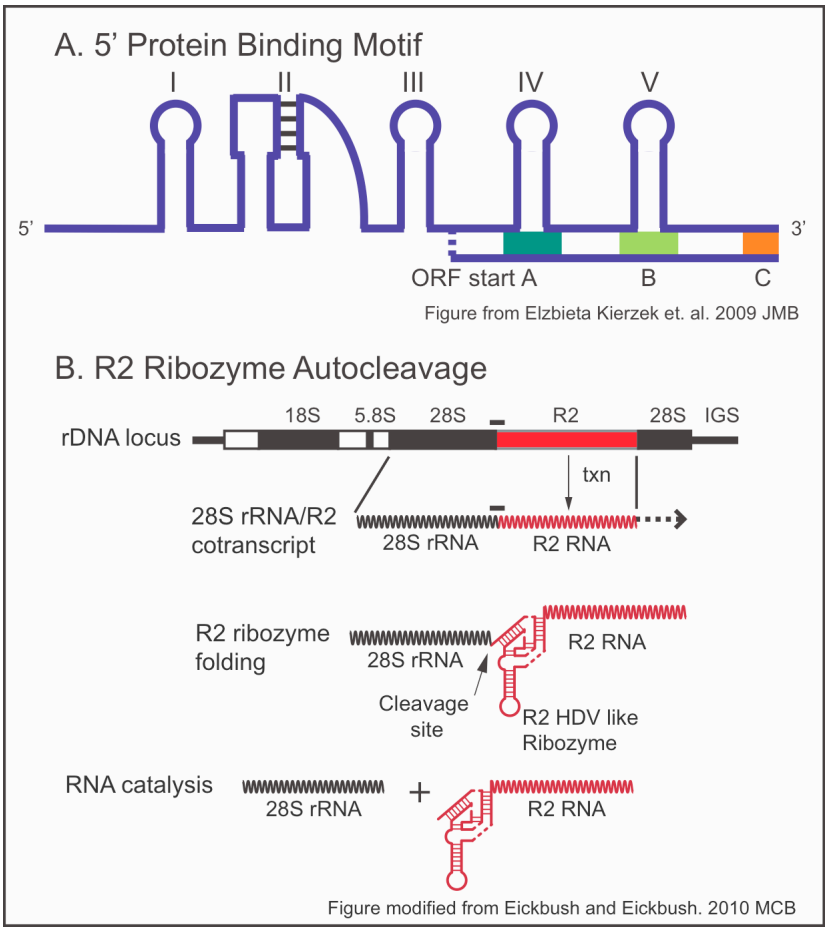


Figure 1.8: R2 RNA Structures. A. Panel A shows the five (i-v) RNA structures of the 5' PBM. Depicted below iv and v is the ORF start site. B. Panel B illustrates the steps of R2 ribozyme folding and autocleavage from the rRNA/R2 cotranscript. Black zig zag lines represent rRNA and red zig zag lines represent R2 RNA. The R2 ribozyme is not predicted to be within the 5' PBM.

Past and current experiments show that the complex structures of the R2 RNA allow for specific binding of the R2 protein to its own RNA during translation. Footprint analysis of the R2Bm protein shows protein binding of the 28S rDNA gene spans ~60bp (Figure 1.9 Dimer) (*128)*. Binding of the 5' PBM by R2Bm protein coordinates the RNP complex to bind target DNA downstream of the insertion site (*125)*. Footprints of the zinc finger motif and myb domain show that these motifs are responsible for this downstream target binding (Figure 1.9 panel N-

terminal peptide) (*65)*. The protein sub-unit bound to the 3‘ PBM binds the target DNA up to the integration site spanning a region of ~40bp (Figure 1.9 Monomer). In the absence of the 5’ PBM the R2 protein preferentially binds upstream of the target insertion site. Binding of R2Bm to the 3‘ PBM has no effect on upstream target site preference. It is clear from these experiments that there are two DNA binding domains in the R2Bm protein, one domain is composed of the amino terminal zinc finger and Myb motifs and the other is currently unidentified.

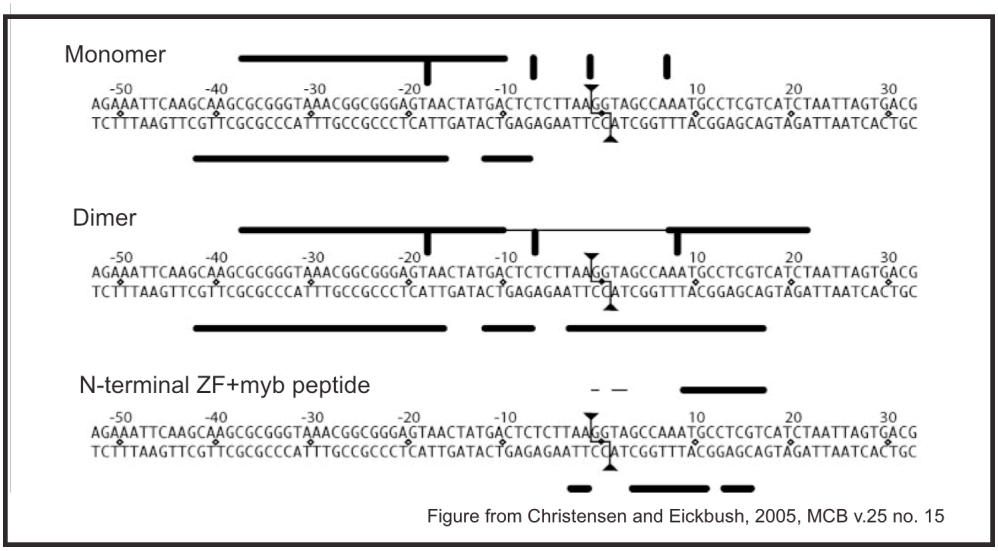


Figure 1.9: R2Bm Target Site Footprint. Summary of the R2Bm protein footprints. The 2-bp staggered cleavage by R2 is indicated by the lines ending in triangles. Nucleotide postitions are numbered relative to the dyad cleavage site, with negatice numbers correspondind to upstream sequences. Thick horizontl lines represent areas of greatest protection by the R2 protein, while thinner horizontal lines indicate weaker protection. Thick vertical lines are DNase I-hypersensitive sites.

Once translated, two sub-units of R2 protein bind to R2 RNA, one at the 5’ PBM and one at the 3’ PBM. This RNP particle will bind to ~60bp of the 28S rDNA gene (*128)*. The protein subunit bound to the 5‘ PBM preferentially binds to the downstream side of the target (*125)*. The protein sub-unit bound to the 3‘ PBM binds to the target DNA upstream of the integration site (figure 1.10.A). Once bound to the DNA, the upstream sub-unit cleaves the bottom strand of DNA exposing a free 3’-OH to prime cDNA synthesis (*64, 129, 130)*. The upstream sub-unit performs TPRT and as the reaction nears completion the shortening RNA

template is released from the sub-unit of protein which is still bound downstream. This release of the 5' PBM activates second strand cleavage. The final step in the reaction is second strand synthesis which is yet to be observed with confidence in-vitro (Figure 1.10, panel B) (*65, 116, 121, 122, 125, 128, 131, 132)*. The R2Bm RT has all the biochemical requirements to perform second strand synthesis, including performing DNA templated DNA synthesis. In fact, R2Bm can polymerize this reaction more efficiently than when using an RNA template *In-vitro*. The R2Bm RT can displace RNA from an RNA:DNA hybrid quite readily, and it is highly processive (*133)*.

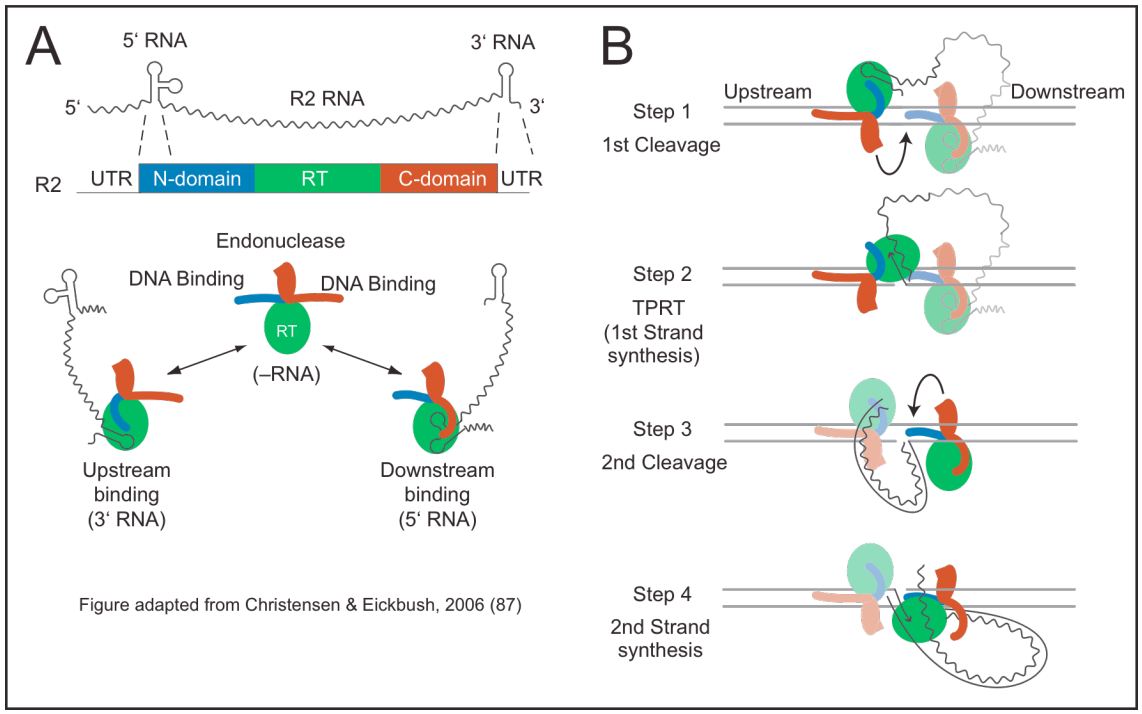


Figure 1.10: R2Bm Integration Model. (A) The R2Bm protein contains an N-terminal DNA-binding domain (blue arm) consisting of zinc finger and myb motifs (not shown), a central reverse transcriptase (RT) domain (green), and a C-terminal domain consisting of an endonuclease domain (red oval) and a proposed DNA binding domain (red arm). R2 protein bound to the 3' UTR RNA exposes the upstream DNA binding domain. Protein bound to the 5' RNA exposes the N-terminal DNA-binding domain to bind the downstream site. (B) R2 integration is proposed to be catalyzed by two subunits in four steps. Step 1: The endonuclease (red oval) from the upstream subunit is responsible for first strand cleavage. Step 2: The RT (green oval) of the upstream subunit catalyzes first strand TPRT. Step 3: The downstream subunit cleaves the second DNA strand. Second strand cleavage does not occur until reverse transcription strips away the 5'□□ RNA region bound to this subunit. Step 4: The downstream subunit provides the polymerase to perform second strand DNA synthesis. Step 4 has not yet been shown to occur in vitro.

CHAPTER 2

DNA BINDING PROPERTIES OF THE R2LP AMINO TERMINUS

2.1 Background

Significant discoveries in the field of genomics have been made over the past half century. One of these major discoveries was that of transposable elements (TEs) and their impact on genome structure and evolution. TEs or jumping genes cause mutations and genetic diseases by copying their own sequences within a host genome without regard for host gene function. Most eukaryotic genomes studied to date contain transposable elements.

The R2 elements are a family of non-Long Terminal Repeat retrotransposons which specifically target the 28S rDNA genes. These retrotransposons encode a variable number of zinc finger motifs, a DNA binding myb domain, a central reverse transcriptase domain, a carboxyl terminal cystidine histidine rich motif, and a restriction like endonuclease. R2 elements are primarily found within the genomes of insects and are estimated to date back almost 900 million years. By targeting the highly repetitive ribosomal DNA (rDNA) locus R2 elements can disrupt gene function without causing damage to the overall fitness of the host due to the high copy nature of the ribosomal repeat region (*115)*. This clever target site preference enables R2 elements to be maintained in host genomes by vertical inheritance for hundreds of millions of years (*43)*. R2 elements have been shown to integrate via the target primed reverse transcription mechanism (TPRT) where the endonuclease domain creates a free 3' OH that is used to prime cDNA synthesis by the RT domain using element RNA as a template.

R2 is an excellent model to study the TRPT mechanism for two primary reasons. First, R2 protein may be expressed in bacteria and purified for *in-vitro* studies unlike the LINE-1 ORF2 protein. Second, because the R2 elements are site specific, their integration mechanism

may be studied *In-vitro* using a small specific target site. A comprehensive understanding of the R2 element integration mechanism is likely to be recapitulated throughout all TPRT elements.

There are four phylogenetically distinct clades of R2 elements that share fundamental hallmarks including integration into the 28S rDNA gene, lack of introns, a short poly A tail, and very slight sequence divergence within a single genome. These hallmarks suggest a reverse transcription integration mechanism.

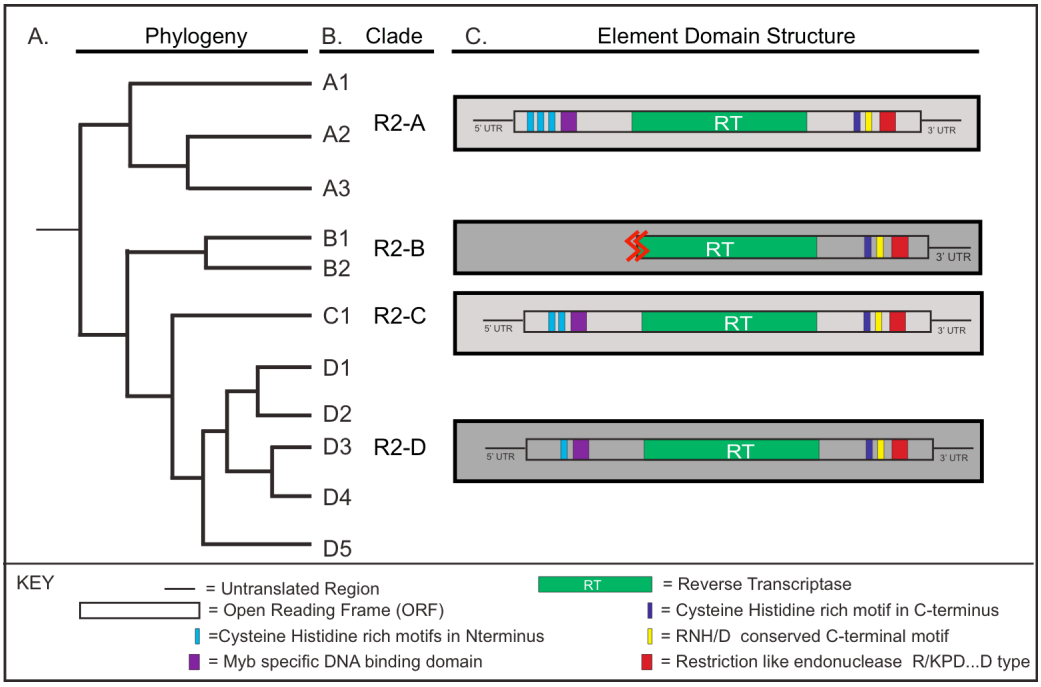


Figure 2.1. R2 Element Domain Structures. A. column A represents the R2 subclade phylogeny based on bayesian analysis. each domain structure belongs to. B. representation of full length element of each subclade. C. column C represents the number of elements (full length and truncated) discovered from each subclade.

An intriguing observation emerged from the phylogenetic studies of the reverse transcriptase domain. All R2 A clade elements contain three amino terminal zinc finger motifs, while R2 B elements have no full length representatives, R2 C elements contain two zinc fingers, and R2 D elements encode only a single zinc finger motif. It is possible that the ancestral R2 element contained three zinc finger motifs and the subsequent divergence led to

loss of these zinc fingers in clades other than R2 A. It is unclear what competitive advantage variation within the zinc finger motifs would convey to R2 elements targeting the same DNA site.

Zinc finger motifs are short cystidine histidine rich motifs that commonly bind DNA, or occasionally RNA and protein (*135, 136)*(*137)*. Zinc fingers typically bind DNA weakly across approximately 3-5 nucleotides (*138)*. Because of the weak DNA binding affinity of zinc fingers they usually operate in coordination with two or more motifs and/or other DNA binding motifs such as a myb or a helix turn helix domain (*139, 140)*. All R2 elements contain an amino terminal myb domain located just downstream of the zinc finger motifs. Myb domains are a family of strong DNA binding proteins. These domains are composed of three alpha helices with conserved large hydrophobic amino acids. Helices one and two coordinate helix three into the major groove of the DNA(*141)*. The third helix is responsible for nucleotide recognition and may bind from 5-10 nucleotides(*142)*.

Data on the R2Bm element, an R2 D clade member with one zinc finger motif, suggests that the protein functions as a dimer with two subunits binding to the target site DNA(*116)*. One subunit binds upstream of the target insertion site and recognizes ~ 40 nucleotides. This DNA binding activity has not yet been assigned to a region of the R2Bm protein. The second subunit binds across ~15nt of the target DNA downstream of the target insertion site. This binding activity has been assigned to the myb and zinc finger motif of the R2Bm protein(*116)*. It has been hypothesized that all R2 element myb domains bind to the target DNA downstream of the target insertion site, however no DNA targeting experiments have been performed on an R2 A, R2 B, or R2 C element.

We are interested in studying the amino terminal variation and its potential impact on R2 element targeting. We specifically want to characterize the role of the variable number of zinc finger motifs in the R2 amino terminus from an R2 A clade member using EMSA and DNA footprinting analysis. By fully understanding the R2 element targeting mechanism we hope to gain insight to the R2 and TPRT integration mechanism. We also would like to explore the

possibility of engineering R2 into a site specific targeting vector. We plan to map the nucleotides recognized by each motif of the R2 A clade element R2Lp, from *Limulus polyphemus,* zinc finger and myb motifs using EMSAs and DNA footprinting. This data will be compared to the data previously produced for the R2Bm element (an R2 D clade member) in hopes of discerning any differences in R2 element target site recognition properties. We hope to assign a larger portion of the bound target DNA to motifs found within the amino terminus. We hypothesize that the R2Lp myb will bind to downstream target site DNA as the R2Bm element. Our second hypothesis is that the zinc finger motifs will extend the footprint across the target site in the 5' direction.

2.2 Experimental methods

2.2.1 *In-Vitro* Protein Expression Vector Design

In order to express, purify, and assay our constructs we had to create a gateway compatible protein expression vector. pDESTTAP (abbreviated pDTAP) is a ligation independent destination vector with an amino terminal TAP tag (6X His and malE). pDESTTAP was generated by cloning the malE gene and a gateway destination cassette into the pET45b plasmid (Novagen 71327-3). pMALc4x (New England Biolabs E8000S) provided the template for the malE gene and pDEST17 (Invitrogen 11803012) served as template for the destination cassette. Primers designed to the malE coding sequence of pMAL*c4x*, Forward - ACTGCATACGGTACCATGAAAATCGAAGAAGGTAAACTGGTAATCTGG and Reverse - ACTGCATACAAGCTTCAATCCTTCCCTCGATCCCGAGG contained *KpnI* and *HindIII* restriction site respectively. Primers designed to the destination cassette region of pDEST17, Forward #1 - ACTGCATACAAGCTTTCGAATCAACAAGTTTGTACAAAAAAGCTGAACG and Reverse - ACTGCATACCTCGAGATCAACCACTTTGTACAAGAAAGCTGAACGAG, contain *HindIII*, and *XhoI* restriction sites respectively. DNA was produced by pcr followed by purification and restriction digestion. pET45b plasmid was restricted with *KpnI* and *XhoI*

restriction enzymes. Restricted insert DNAs were ligated into restricted pET45b simultaneously using T4 DNA ligase (Promega M1801) and a 3:3:1 ratio of *malE*:destination cassette:restricted plasmid at a final concentration of ~70fmol total DNA. Ligated plasmids were transformed into Oneshot *ccdB* survival chemically competent *E. coli* (Invitrogen A10460) and incubated overnight at 37C under selection of 100ug/mL Ampicillin. Transformants were screened with pcr and sequence verified. To create frameshift versions of pDESTTAP we used a second forward primer, Forward #2 - ACTGCATACAAGCTTCAATCAACAAGTTTGTACAAAAAAGCTGAACG, for the destination cassette pcr (designated pDESTTAPb or pDTAPb, 1 nucleotide frameshift). To produce a 2nt frame shift we restricted pDESTTAP with *Hindiii*, created blunt ends with DNA polymerase, and religated plasmids (designated pDESTTAPc). See appendix section A.1 for pDESTTAP vector map and A.2 for pDESTTAP sequence.

2.2.2 *In-Vitro* Protein Expression Constructs

In order to assay the nucleic acid binding properties of R2Lp we generated several *in-vitro* protein expression constructs. These constructs contain a varying number of the conserved zinc finger motifs as well as isolated domains. Naming of clones is as follows; R2Lp ZF1-myb codons 67-280, R2Lp ZF2-myb codons 116-280, R2Lp ZF3-myb codons 154-280, R2Lp ZF1-3 codons 67-183, and R2Lp myb codons 199-280. R2Lp ZF1-myb gene was cloned into pET28a using pcr and restriction digestion, forward - CGAGATCCGCATATGAGAAAGGTGGCATGTGACTTGTGTTCTAAAG with *NdeI* site and reverse CGAGATCGCGGATCCTTAGTCCAACTCCTCGTACTCCTCG with *BamHI* site. Ligation reactions were transformed into XL-1 Blue competent *E. coli* (agilent 200259). All other clones were generated using the gateway system from Invitrogen. R2Lp ZF2-myb forward - CACCGAAACTCAGGCATGCTGCACATATTGC, R2Lp ZF3-myb forward - CACCAGCAATTTCTTGTGTGATCTTTGCAATGATAG, and R2Lp myb forward - CACCCGCCCTCGCCAGGTAGTG. R2Lp ZF1-myb reverse primer was used for all genes listed above. R2Lp ZF1-3 forward - CACCCCGTGCGTAACTGAGGGTAGGTTTG and reverse -

GCAGATCCGGGATCCTCACCTTGAACAAGGATGCTTATGACGCTTATG. All genes except R2Lp ZF1-myb were cloned into the donor vector pENTR/TEV/D-TOPO (Invitrogen K2535-20) and transformed into Oneshot Top10 chemically competent *E. coli* (Invitrogen C4040-10). Clones were pcr screened and sequence verified then recombined into the destination vectors pDEST17 (Invitrogen 11803-102) or pDESTTAP (see 2.2.1) for expression. R2Lp ZF2-myb was expressed in pDEST17 and pDESTTAP. R2Lp ZF3-myb was expressed in pDEST17. R2Lp Myb and R2Lp ZF1-3 clones were expressed in pDESTTAPc and pDESTTAP respectively. All clones listed above were expressed using Arctic express RIL DE3 *E. coli* (Stratagene 230193). The R2Bm myb clone was created from a synthetic codon optimized gene as template and cloned codons 138-228 using forward - CACCGAAACCAATACCGACGCCGCTCC and reverse -CTACGGCTCTGCGCTGCAACC. Pcr products were cloned into pENTR/TEV/D-TOPO then pDESTTAP and expressed in Arctic Express DE3 *E. coli* (Stratagene 230192). All destination clones were screened with pcr and sequence verified. To serve as a control we generated a pDESTTAP version of the LR clonase positive control reaction GUS gene.

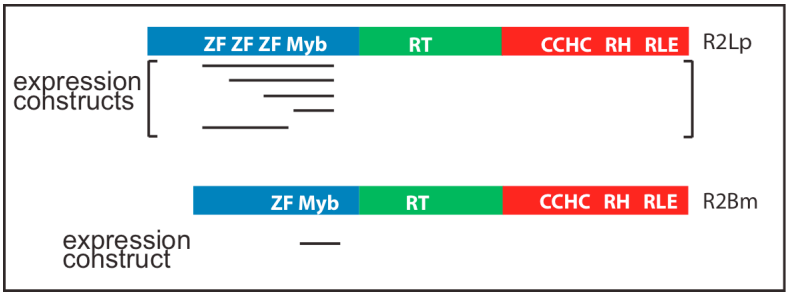
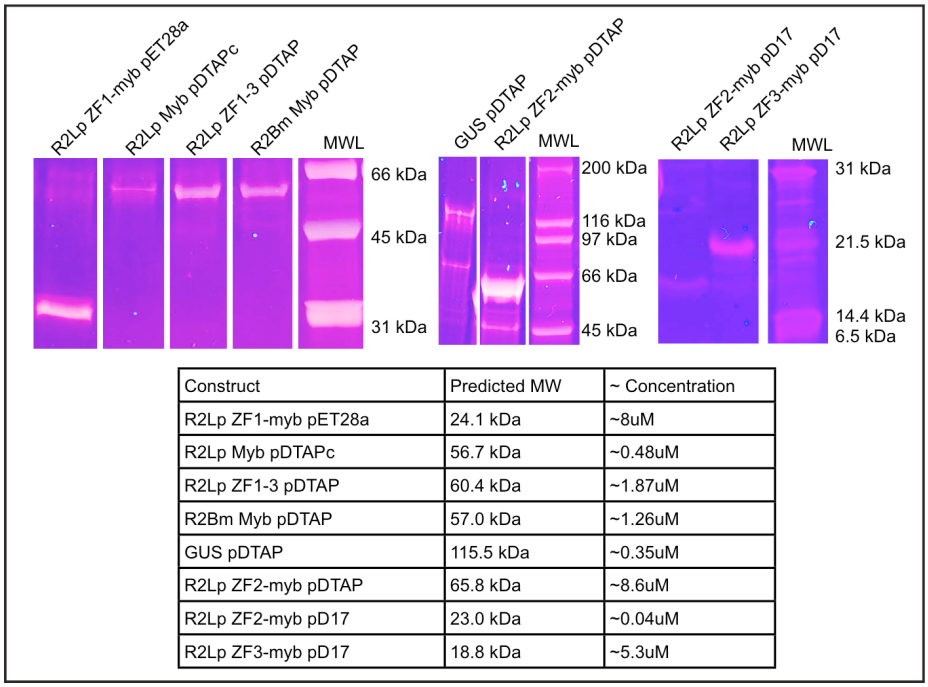


Figure 2.2: R2 Element Expression Domains. R2Lp (top) and R2Bm (bottom) element domain structures are illustrated with blue amino termini, green reverse transcriptase domains, and red carboxyl termini. Black horizontal bars represent protein expression constructs of both elements.

2.2.3 Protein Expression, Purification, and Quantification

200mL Luria Broth protein expression cultures were inoculated from 240uL of saturated subculture. Expression cultures were grown to A_{600}O.D. of 0.6 at 37C with 240rpm agitation. Cultures were then cooled to 12C on ice water and induced to express protein with 0.1mM

IPTG. Cultures were incubated at 12C 240rpm for 24Hrs and harvested with centrifugation at 4000XG for 20min at 4C. Cell pellets were resuspended in 50mL 10mM Tris HCl pH 7.5 and centrifuged again. Rinsed pellets were used immediately or stored at -80C.



Construct	Predicted MW	~ Concentration
R2Lp ZF1-myb pET28a	24.1 kDa	~8uM
R2Lp Myb pDTAPc	56.7 kDa	~0.48uM
R2Lp ZF1-3 pDTAP	60.4 kDa	~1.87uM
R2Bm Myb pDTAP	57.0 kDa	~1.26uM
GUS pDTAP	115.5 kDa	~0.35uM
R2Lp ZF2-myb pDTAP	65.8 kDa	~8.6uM
R2Lp ZF2-myb pD17	23.0 kDa	~0.04uM
R2Lp ZF3-myb pD17	18.8 kDa	~5.3uM

Figure 2.3: Purified Proteins. Shown above are purified proteins from in-vitro expression constructs. All gels are 15% PAC (19:1), standard laemmli SDS PAGE system. Above the gel lanes are the construct name. Ladder lanes are labelled MWL and sizes are marked to the right of each gel.

Cell pellets were resuspended in 2.5mL of a 50% glycerol, 100mM Hepes pH7.5, 5mM *beta*-mercapto ethanol solution containing 2mg/mL of hen egg white lysozyme (Amresco 0663) and incubated on ice for 30min with gentle hand warming every 10 minutes. 13.2mL of lysis solution (100mM Hepes pH 7.5, 1M NaCl, 5mM *beta*-mercapto ethanol, and 0.2% triton X-100) was gently added to the resuspension and inverted several times to mix. Cell lysate was then centrifuged at 33,000 rpm and 2C under vacuum for 20 hrs. Supernatant was decanted and allowed to gravity flow through a prewashed (3mL of 50mM Hepes pH 7.5, 500mM NaCl, 0.02% triton X-100, 5mM Imidazole pH 7.5, and 2mM *beta*-mercapto ethanol) talon metal affinity resin (Clontech 635501) column. Resin bound protein was washed with increasingly stringent

solutions of column buffer, wash 1 (1.2mL of prewash with 10mM imidazole), wash 2 (1.2mL of prewash with 20mM Imidazole), wash 3 (1mL of prewash with 300mM NaCL and 30mM Imidazole), wash 4 (600uL of prewash with 300mM NaCl and 40mM Imidazole) and eluted (300uL of prewash with 300mM NaCl and 60mM imidazole). Elutant was diluted to 0.5X with 100% glycerol and stored at -20C(*116)*.

Protein concentration was determined using laemelli PAGE and a BSA stardard (Biorad 500-0202) of know concentrations. Strip densitometry measurements determined band intensity of both Sypro orange (Biorad 170-3120) and Comassie blue R-250 (Amresco 0472-10G) stained gels using ImageJ(*143)*. A linear regression analysis was used to approximate protein concentration.

2.2.4 DNA Target Preparation

To assay DNA binding activity we produced P-32 end labeled substrates *in-vitro.* For preliminary binding experiments and footprint analysis we will use a pcr generated 150 mer (80/70, forward- GTGATTTCTGCCCAGTGCTCTGAATGTC reverse-GATAGGGACAGTGGGAATCTCGTTAATCCATTC), for competitive target site binding we used a set of 49mer half site target oligos (49/0 ATTCAAGCAAGCGCGGGTAAACGGCGGGAGTAACTATGACTCTCTTAAG) and (0/49 GTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATTAA). As a non specific 49mer we used oligos containing pBluescriptII polylinker (CTAGTGGATCCCCCGGGCTGCAGGAATTCGATATCAAGCTTATCGATAC). Target DNA was radiolabeled using gamma-ATP and T4 polynucleotide kinase (Promega M4101) (*116, 128)*. Complementary unlabeled target oligos will be incubated with labeled oligo and incubated at 70C for 10 minutes followed by slow cooling to room temperature in 1X STE (100mM NaCl, 10mM Tris HCl pH 7.5, 1mM EDTA pH 8.0). Annealed oligos are mixed with glycerol to 13% final and band purified by native 1X TBE pH8.3 PAGE (7% 19:1 polyacrylamide, 89mM Tris base, 89mM Boric acid, 3mM EDTA, pH8.3). Oligos were extracted from gel slices with 400uL

of crush and soak buffer (0.3M Sodium Acetate pH5.5, 0.5% SDS, 0.5mM EDTA pH 8.0) and incubated at room temperature for 4-8 hours with gentle agitation. Extracted target sites were chloroform (Amresco X205-450mL) extracted then precipitated. Samples were resuspended in 80uL of 1X TE pH 8.0 (10mM Tris, 1mM EDTA). DNA concentrations were approximated from previous work (*65, 128)*. Primers were labeled as listed above then used to generate labeled target by pcr and purified as listed above.

Missing nucleoside footprint target DNA was created by cleaving 70uL (~9.75pmol) target DNA mixed in 140uL 10mM Tris HCl pH7.5 with 35uL 1mM Fe($EDTA_2$), 35uL 10mM Ascorbate, and 35uL of 0.02% H_2O_2 for 2 minutes at room temperature(*144, 145)*. Cleavage reaction was quenched by addition of 120uL of 100% glycerol and incubation on ice water. Reaction was precipitated with glycogen and resuspended in 70uL 1X TE pH8.0. Appendix A.3 contains more information about the missing nucleoside target DNA preparation.

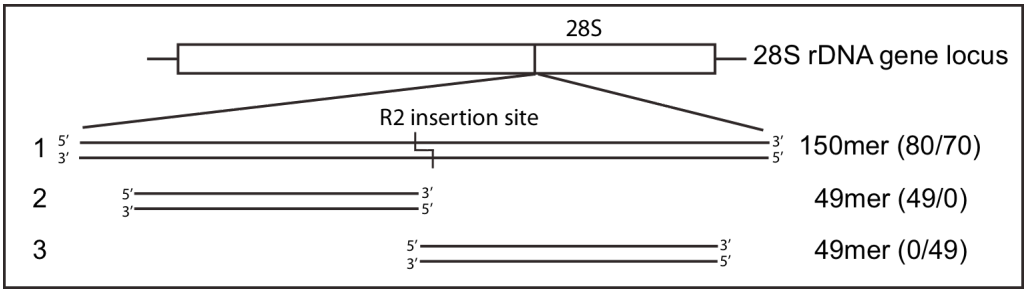


Figure 2.4: DNA Target Sites from the 28S rDNA Gene. Above is a representation of the 28S section of the ribosomal gene locus. The vertical line inside the 28S gene box is the R2 insertion site. Below an expanded view of the R2 target site. The zig-zag line represents the insertion site. To the right is a description of the target length and (#bp before/#bp after) insertion site.

2.2.5 Electro-Mobility Shift Assays

EMSAs were performed in 13 uL reactions with final conditions of 10mM Tris pH 7.5, 75mM NaCl, 3mM MgCl2, 0.5mM CaCl2, 0.01% Triton 100X, and 1mM DTT. All full target substrate EMSA reactions contained 1uL of target site prep (~5nM) and 3uL of protein. EMSAs were assayed in 0.5X TBE pH8.3 PAGE gels (44.5mM Tris base, 44.5mM Boric Acid, and 1.5mM EDTA) and run ~5cm at 300V. Gels were dried and visualized on a phophurimaging

screen for 1Hr. Half site EMSAs contained ~5nM labeled target substrate. Competitive half site EMSAs contained ~10ng (24nM) labeled target DNA and ~100ng sheared DNA.

2.2.6 DNase I and Missing Nucleoside Footprints

DNase footprint binding reactions were performed in 30uL volumes (10mM Tris HCl ph 7.5, 75mM NaCl, 3mM $MgCl_2$, 0.5mM $CaCl_2$, 0.01% triton X-100, 1mM DTT, and 1.65% glycerol final) containing 3uL of protein (18% glycerol, 25mM Hepes, 150mM NaCl, 0.01% triton, and 2mM DTT). Reactions were incubated at room temperature for 25 minutes and then subjected to DNA cleavage by DNase I enzyme (1U/uL, Promega M610A). 3uL of a 1:250 diluted DNase I in 1X DNase reaction buffer was incubated with binding reactions for 2 minutes at room temperature. Cleavage reactions were quenched in ice water slurry and mixed with 8uL 50% glycerol solution at 1X binding reaction conditions. Samples were allowed to separate by native EMSA PAGE (7% PAC 19:1 and 0.5X TBE pH8.3 for R2Lp ZF1-myb and reference, and 5% PAC 29:1and 0.5X TBE pH8.3 for R2Lp myb and R2Bm myb). Gels were not dried down, bound and free fractions of samples were extracted from EMSA gel as explained above(*128, 146)*. Samples were precipitated with 10ug sheared denatured DNA and resuspended in a 0.5X TE buffer. Samples were quantified by scintillation counting and lyophilized for 45 minutes at 45C. Samples were resuspended to a final activity of 4,000 counts/min/uL. An Adenine + Guanosine DNA molecular weight ladder was created by using the Maxam and Gilbert method. Fractions were allowed to separate over a 1X TBE pH 8.8 large denaturing polyacrylamide gel (6% 19:1 polyacrylamide, 8M Urea, 133mM Tris base, 89mM Boric acid, 3mM EDTA) at constant temperature of 55C. Footprint gels were dried and exposed to phosphurimager plates and autoradiography film.

Missing nucleoside footprint EMSAs differ from DNase I footprint EMSA reactions only by the addition of 13% glycerol in the binding reaction and the use of hydroxyl radical treated DNA. All extraction, quantification, and visualization techniques are identical to DNase I footprints, however missing nucleoside samples were resuspended at 10,000 counts/min/uL

after lyophilization. A detailed version of the DNase I and missing nucleoside footprint reaction optimizations may be viewed in appendix A.2 and A.3.

2.3 Results

2.3.1 R2Lp ZF1-Myb Clone DNA Binding Properties

R2Lp ZF1-myb DNA binding properties were first assayed using EMSAs with full target site DNA. Reactions containing only protein storage buffer (protein -) and gels run with the GUS control gene (GUS pDTAP) served as assay controls. Figure 2.5 panel A shows R2Lp ZF1-myb protein titrated from 3.5nM to 0.13nM to ascertain optimum DNA binding capacity. GUS protein was titrated from 81nM to 0.33nM in order to give highest possible contaminating shift data in panel B. Radio labeled target DNA was held at a constant concentration of ~5nM for all reactions. R2Lp ZF1-myb protein shifts nearly all target DNA and forms low concentration dimer complexes at 3.5nM, however only monomer complexes are seen at concentrations of 1.16nM and lower. Percentage of shifted target DNA seems to correlate strongly with protein concentration. GUS protein shifts nearly all target DNA to the gel well and within a smear band at 81nM, but DNA shifts seem unresponsive to protein concentrations at 27nM and lower suggesting a low level contaminant shift of target DNA. Shifted DNA height of GUS gene runs to a different height than R2Lp ZF1-myb protein shifts suggesting that the R2Lp protein does not contain the same contaminant.

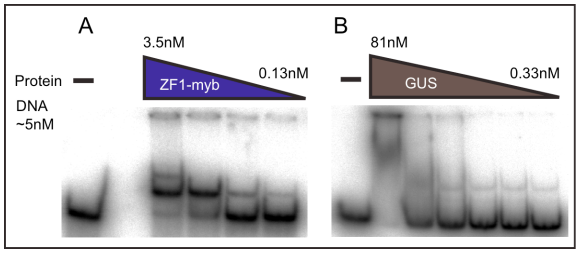


Figure 2.5: R2Lp ZF1-Myb Full Target EMSA. Native 6% PAC (19:1), 1X TBE pH8.3. Panel A shows R2Lp ZF1-myb protein EMSA. Titrtation of protein is depicted as a decreasing triangle above gel image with high and low concentrations indicated. Protein negative lanes are below the horizontal black bar. Panel B shows the control GUS gene in similar fashion. All lanes contain ~ 5nM DNA.

In order to test which region of the target site DNA the R2Lp ZF1-myb clone was binding we performed a set of half target site DNA EMSAs. Half site DNA contained either 49 bp of target DNA upstream or downstream of the R2 element insertion site. EMSAs were performed with ~5nM labeled target DNA containing a protein titration of R2Lp ZF1-myb protein. Protein storage buffer and a non-specific DNA served as reaction controls.

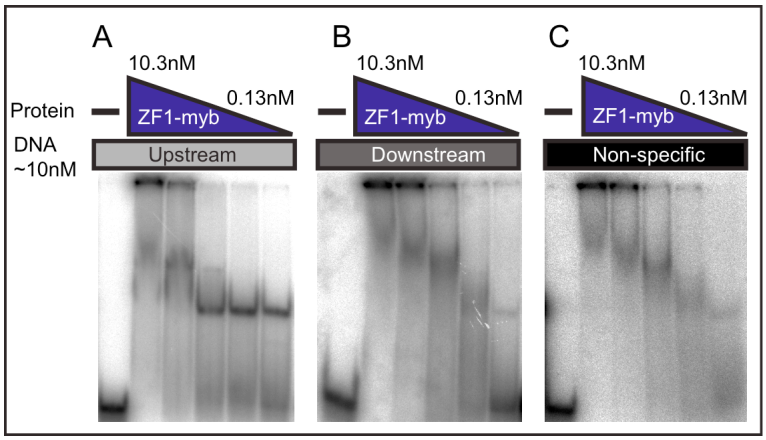


Figure 2.6: R2Lp ZF1-MybTarget Half Site EMSA. Native 7% PAC (19:1), 1X TBE pH8.3 . Panel A shows R2Lp ZF1-myb protein EMSA with upstream target DNA. Panel B shows downstream target DNA and panel C shows a non specific target DNA. Titrtation of protein is depicted as a decreasing triangle above gel image with high and low concentrations listed. Protein negative lanes are below the horizontal black bar. All reactions contain ~10nM DNA.

In panel A we test the ability of the R2Lp ZF1-myb protein to bind upstream target site DNA. Protein shifted target DNA to the well and formed strong dimer complexes at concentrations of 10.3 and 3.5nM. Protein at concentrations of 1.16nM and lower give a strong monomer complex that responds well to protein concentration. Panel B shows downstream target DNA binding of R2Lp ZF1-myb, however no clear monomer bands were seen at any protein concentration. A strong well complex and a dimer sized smear were observed at the higher protein concentrations in the downstream target DNA binding reactions. Panel C contains the same protein titrations as panels A and B but using a labeled non specific 49bp target DNA. Results of panel C are very similar to panel B. These results suggest that the R2Lp ZF1-myb protein bind preferentially to the upstream target site in a base pair specific manner, but binds the other two targets in non specific manner at concentrations above a 2:1 target DNA

to protein ratio. To summarize the R2Lp ZF1-myb protein appears to bind the R2 element target site upstream of the insertion site.

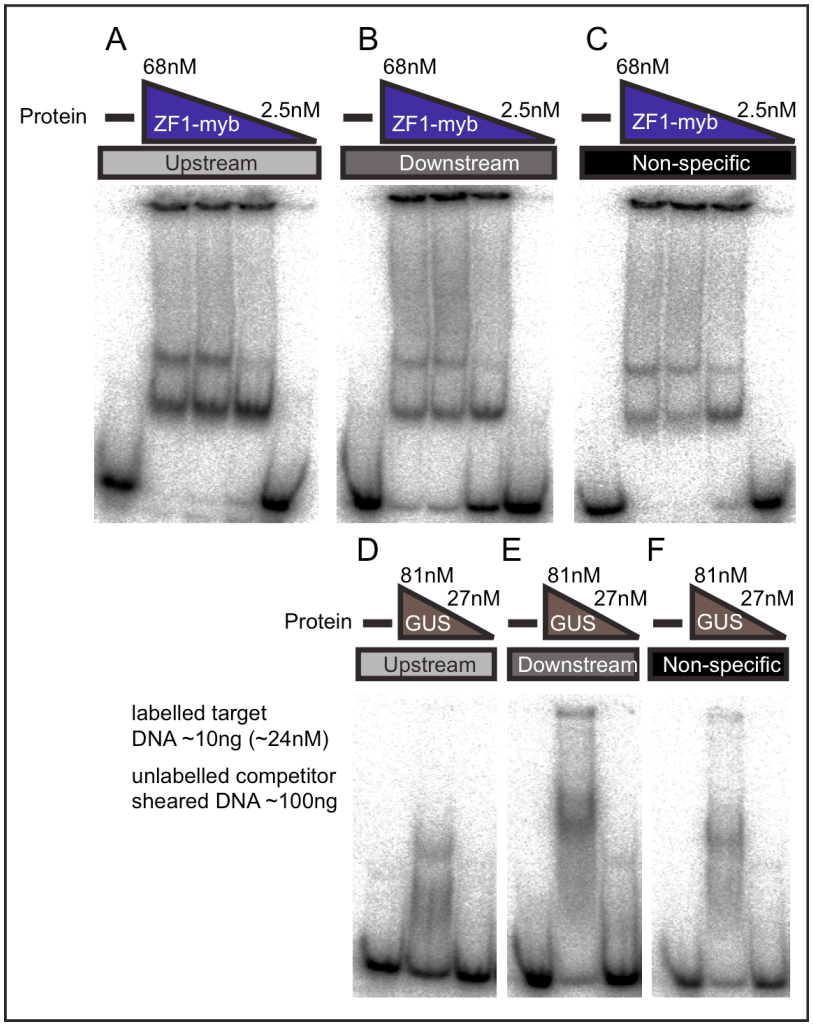


Figure 2.7: R2Lp ZF1-Myb Target Half Site Competition EMSA. Native 5% PAC (29:1), 0.5X TBE pH8.3. Panel A shows R2Lp ZF1-myb protein EMSA with upstream target DNA, panel B shows downstream target DNA, and panel C shows a non specific target DNA. Titrtation of protein is depicted as a decreasing triangle above gel image with high and low concentrations listed. Protein negative lanes are below the horizontal black bar. All reactions contain ~10ng labelled DNA and 100ng unlabelled competitor DNA. Panels D, E, and F are similar to A, B, and C except GUS protein was added.

In order to confirm that the R2Lp ZF1-myb protein binds the upstream target DNA in a base pair specific manner we performed a competitive half site EMSA using a labeled specific

target site and sheared denatured salmon sperm DNA as an unlabeled non specific DNA competitor in 10 fold excess. We titrated R2Lp ZF1-myb protein using protein storage buffer as a loading control. The GUS gene served as a vector control, and we used a labeled non-specific DNA as a double negative control.

Panel A shows the R2Lp ZF1-myb protein binding to upstream target and Panel D shows the GUS control shift. All shift reactions in panel A show strong monomer, dimer, and well complexes except at 2.5nM (~10 fold excess labeled target DNA). Shift complex response to protein concentration suggests that the binding reaction is strong and base pair specific. Panel B shows R2Lp ZF1-myb protein binding to downstream target site DNA and panel E shows the GUS control reaction. Panel B shift reactions appear very similar to panel A with slightly less of the total DNA shifted. These results suggest that the R2Lp ZF1-myb protein may be binding downstream target site DNA tightly. Panel C shows the R2Lp ZF1-myb protein EMSAs using a labeled non-specific DNA substrate and panel F shows the GUS control reaction. This panel also appears similar to panels A and B. Panel C target shifts illustrate that the R2Lp ZF1-myb protein is binding the non-specific DNA target tightly. These results suggest that the R2Lp protein may be binding to all target DNA substrates with a high affinity, however the target sites do not share any regions of high sequence homology. It appears that the R2Lp ZF1-myb clone is binding target DNA in a non-sequence specific manner but tightly in this experiment.

2.3.2 R2Lp ZF1-myb Footprint Analysis

Based on early experimental results we decided that EMSA reactions cannot on their own explain our hypotheses. As an alternative method we chose to perform footprinting analysis to obtain higher resolution observations of our proteins activity.

We began footprinting analysis with a low resolution DNase I footprint. Full target site DNA was incubated with R2Lp ZF1-myb protein and then subjected to digestion with DNase I enzyme. Purified fractions of bound (monomer) and free (unbound) DNA complexes were

denatured and separated over a large denaturing polyacrylamide gel. Target site DNA cleaved by formic acid (Maxam and Gilbert method) at adenine and guanidine residues serves as a molecular weight ladder. Target DNA subjected to digestion in the absence of protein serves as a digestion control. Panel A shows R2Lp ZF1-myb protein bound and free fractions on top strand labeled DNA with both reference DNA and DNA ladder. The protein bound specifically to nt -41 to -33, nt -20 to -13, and nt -10 to -5. These regions of bound DNA are represented as vertical red bars in the diagram. Hypersensitive sites were created at -21, -12, -4, +4, and +5 and are represented as short horizontal blue bars in the diagram. Panel B shows DNase I footprint of R2Lp ZF1-myb protein on bottom strand labeled DNA. Protein footprinted a region of ~35 nt on the bottom strand including regions from nt -41 to -32, nt -24 to -17, and -14 to -7. Hypersensitive sites were recorded at nt -26,-25, -16, -6, -5, and -4. Panel C summarizes the data mapped over the target site DNA sequence (the A+G ladder is mapped one nucleotide short of the DNA sequence). Due to the method of footprint analysis, all nucleotides that show a footprint are covered by protein therefore prohibiting the DNase I enzyme from creating cut target. However, this does not mean that every nucleotide that has been footprinted is bound specifically by the protein. These low resolution results suggest that the protein binds to the upstream region of the target DNA from nt -42 to -5.

To obtain a higher resolution footprint we employed the missing nucleoside technique, in which the fenton reaction creates and recycles hydroxyl radicals to excise nucleotides and fragments the backbone of the DNA target. Panel A shows R2Lp ZF1-myb protein missing nucleoside footprint on top strand labeled DNA. In missing nucleoside footprints the protein will show a zone of clearing in the footprint region of the bound fraction and overrepresentation within the free fraction of purified DNA. Bound lanes show a clear footprint from nt -41 to -34 and from nt -21 to -19 with a hypersensitive site at -18. These regions are accentuated by darkening of the free bands relative to the reference DNA bands. Panel B shows the R2Lp ZF1-myb protein footprinted on the bottom strand of target DNA. Protein bound lanes show two

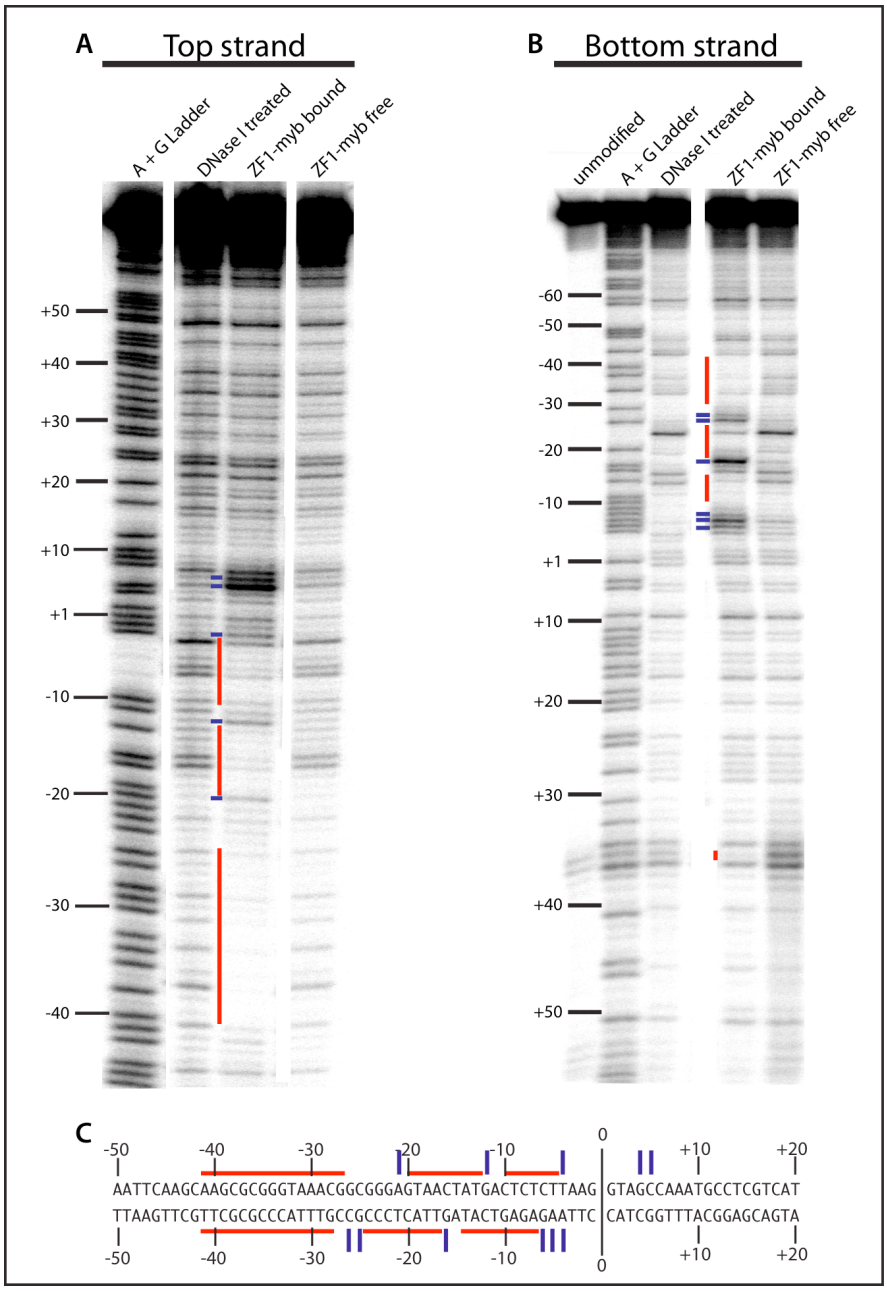


Figure 2.8: R2Lp ZF1-Myb DNase I Footprint. Shown above is a 1X TBE pH8.8 denaturing polyacrylamide gel, 6% PAC (19:1), DNaseI footprint of the R2Lp ZF1-myb protein. Panels A and B show top and bottom strand labelled DNA respectively. A) lane1 is Adenosine + Guanosine target DNA ladder, lane 2 Is DNase I treated reference DNA, lanes 3 and 4 show bound and free fraction DNA respectively. Areas of footprinting are shown as red vertical bars to the left of bound DNA lanes. Hypersensive sites are shown as short horizontal bars to the left of bound DNA lanes. Panel B is similar to panel A but with bottom strand labelled DNA and lane 1 is unmodified DNA. Panel C shows target site DNA sequence with numbering.

regions of footprinting from nt -38 to -31 and nt -22 to -18. Panel C illustrates the footprinted region over and under the target site sequence. Results from this experiment show a strong nucleotide binding of base pairs -40 to -31 and -22 to -18 of the target site DNA.

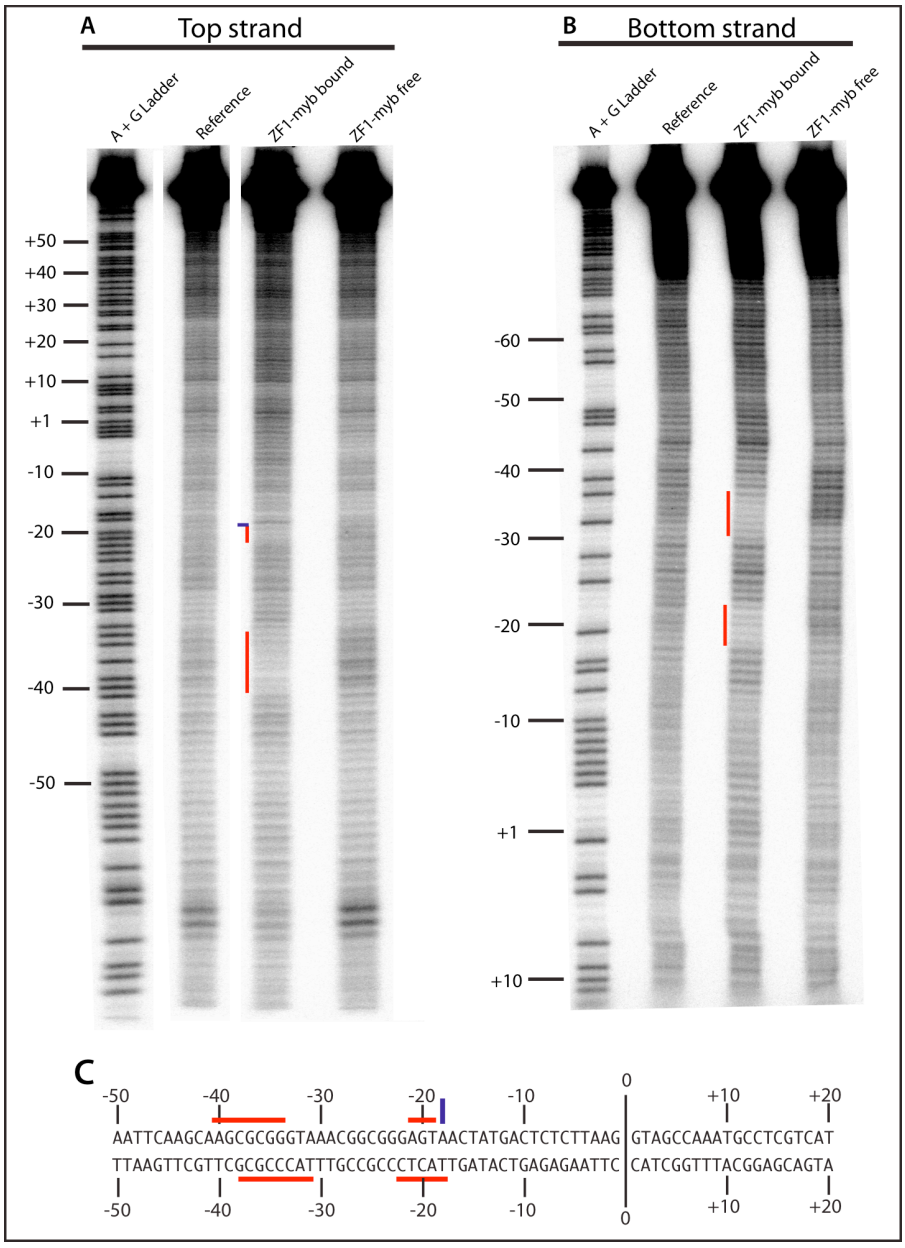


Figure 2.9: R2Lp ZF1-Myb Missing Nucleoside Footprint. Figure 2.9 is similar to figure 2.8 with the exception of hydroxyl radical treated DNA for missing nucleoside footprints.

2.3.3 R2Lp and R2Bm Myb Domain DNA Binding Activity

Data from the aforementioned experiments shows a clear picture of the R2Lp DNA binding activity and maps it to a region of ~35 nucleotides upstream of the target insertion site. Although this data is significant it does not show which DNA binding motif is responsible for binding specific nucleotides of the mapped target region. For this reason we have decided to perform the analysis again with a second set of clones specific to the R2Lp myb domain, the R2Bm myb domain, and the R2Lp ZF1-3 motifs. We will proceed with the data from the R2Lp myb and R2Bm myb domains first.

To first test the ability of our sub-clones to bind the target site DNA we performed EMSAs using full target site substrate. Proteins were titrated over an 81-fold change in concentration ranging from 2.5-fold excess protein to 33-fold excess DNA target. Protein storage buffer serves as the unbound DNA control and GUS control gene shifts can be viewed above in figure 2.5. Panel A shows the R2Lp myb domain protein titrated from 12nM to 0.15nM concentrations. Protein negative lane DNA runs at an aberrant molecular weight due to gel smiling. The R2Lp myb domain forms strong monomer complexes that shift DNA in a protein concentration dependent manner. Panel B shows the R2Bm myb domain titration. R2Bm myb was titrated at a slightly higher concentration due to weaker DNA binding capacity and ranges from 32nM to 0.4nM or from 6-fold excess protein to 12-fold excess DNA target. R2Bm myb protein forms a slight well complex at the highest concentrations but forms monomers of differing sizes as protein is titrated. This phenomenon is believed to be due to dilution of imidazole from protein purification. These results indicate that the R2Bm myb domain is binding to target DNA and forms monomer complexes well. This experiment shows that both the R2Lp and R2Bm myb domains bind target site DNA, however the binding of the R2Lp myb to target DNA appears to be stronger.

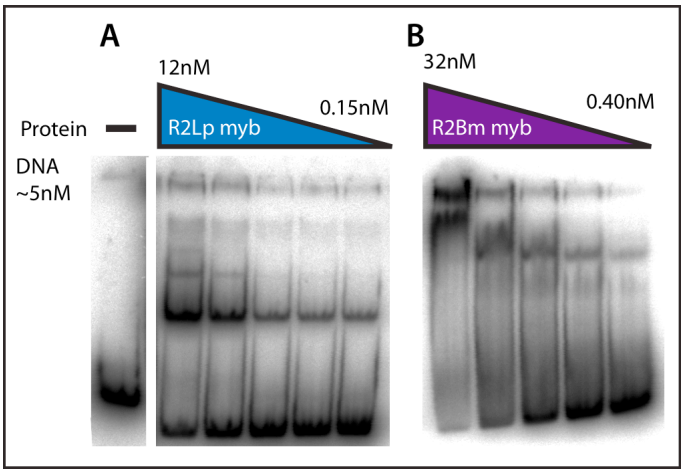


Figure 2.10: R2Lp Myb and R2Bm Myb Full Target EMSA. Native 7% PAC (19:1), 0.5X TBE pH8.3. Panel A shows R2Lp myb protein and panel B shows R2Bm Myb protein EMSAs. Titrtation of protein is depicted as a decreasing triangle above gel image with high and low concentrations indicated. Protein negative lanes are below the horizontal black bar. All lanes contain ~ 5nM DNA.

To test the specificity of the R2Lp and R2Bm myb domains we performed competitive DNA binding EMSAs using labeled specific target DNA and unlabeled competitor DNA. Protein negative lanes and a labeled non-specific DNA substrate serve as reaction controls and the GUS control gene reactions may be viewed above in Figure 2.7. Figure 2.11 panels A, B, and C show R2Lp myb protein bound to upstream, downstream, and non specific target DNAs respectively. In panel A we observed the formation of strong dimer and monomer complexes at high concentrations of protein leading to just monomer complexes at lower protein concentrations. Panel B shows a weaker binding capacity than panel A, suggesting a higher degree of specificity to bind upstream DNA. Panel C appears very similar to panel A and suggests the strong DNA binding affinity of the R2Lp protein to the non-specific DNA substrate. Panels D, E, and F show R2Bm myb protein bound to upstream, downstream, and non-specific DNAs respectively. In panels D, E, and F we observe the formation of monomer complexes to approximately the same concentrations independent of target DNA. This suggests that the R2Bm myb DNA binding capacity is relatively non-specific in nature. Panel E shows a slightly

higher affinity for downstream target DNA in terms of total DNA target shifted compared with panels D and F. In summary these experiments show that the R2Lp myb has a higher affinity for binding upstream target DNA than the other two targets, however the R2Bm myb domain seems to have nearly identical affinities for all three target substrates.

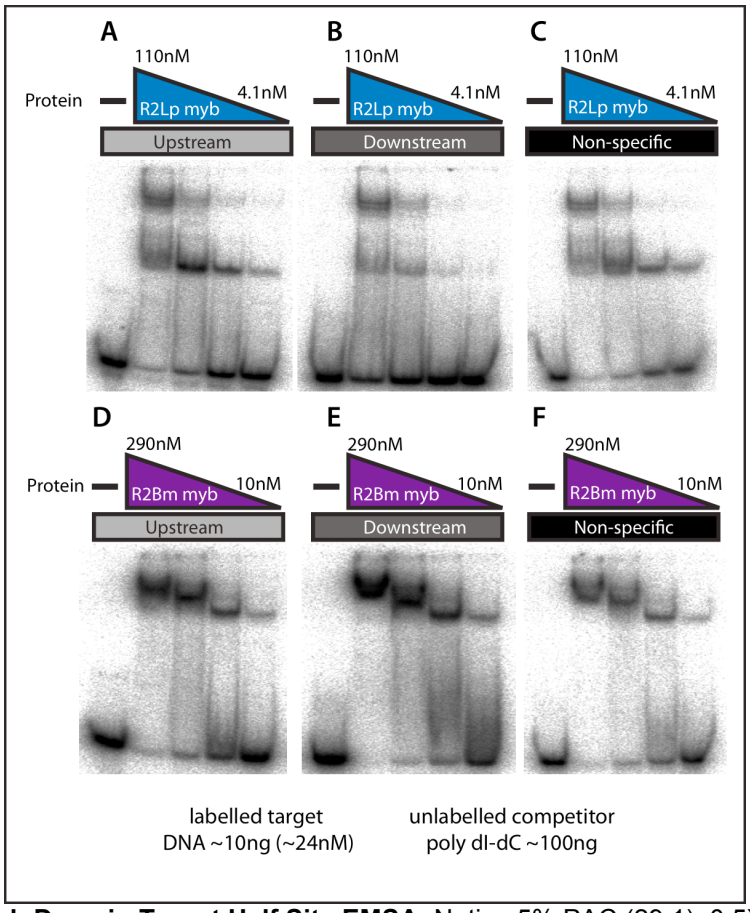


Figure 2.11: Myb Domain Target Half Site EMSA. Native 5% PAC (29:1), 0.5X TBE pH8.3. Panel A-C show R2Lp Myb protein EMSA with upstream, downstream, and non specific target DNA respectively. Titrtation of protein is depicted as a decreasing triangle above gel image with high and low concentrations listed. Protein negative lanes are below the horizontal black bar. All reactions contain ~10ng labelled DNA and 100ng unlabelled competitor DNA. Panels D, E, and F are similar to A, B, and C except R2Bm Myb protein was added.

2.3.4 R2Lp Myb and R2Bm Myb Footprint Analysis

In order to obtain a higher resolution of the DNA binding specificity of both the R2Lp and R2Bm myb domains, were performed both DNase I and missing nucleoside DNA footprints as explained above for R2Lp ZF1-myb.

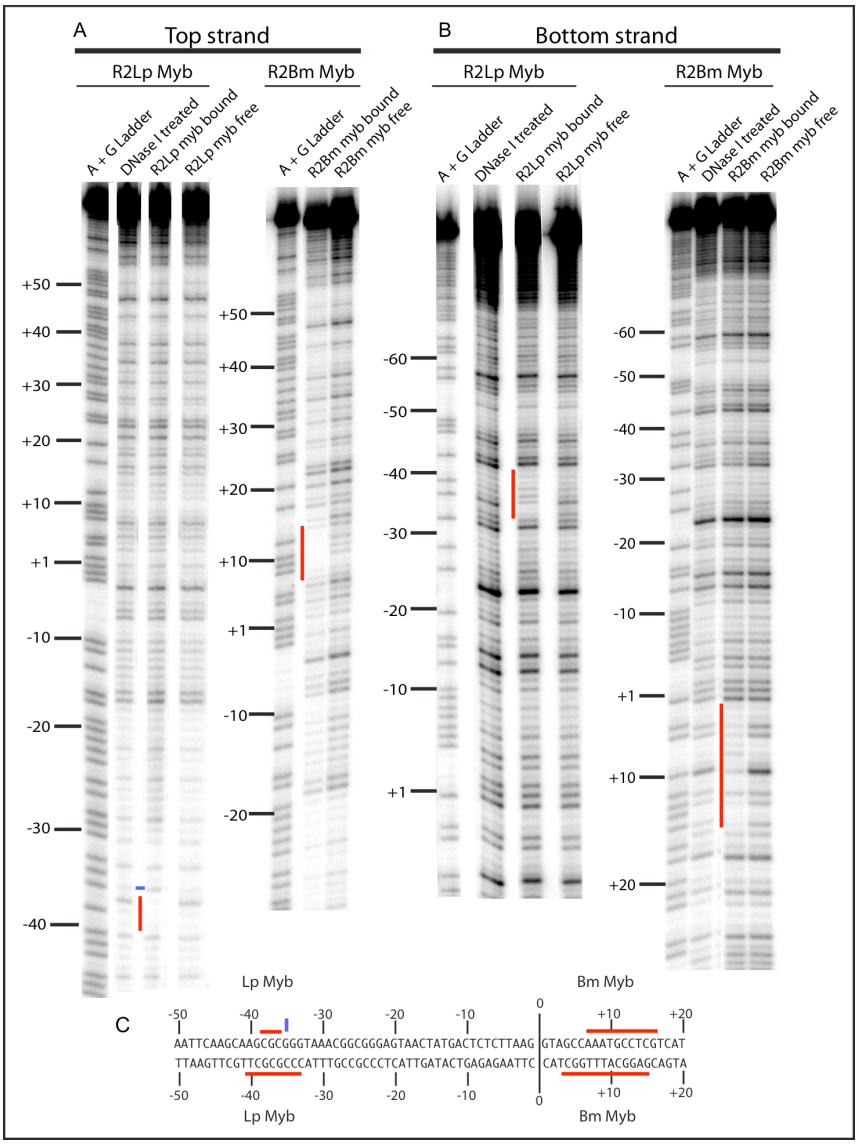


Figure 2.12: R2Lp Myb and R2Bm Myb DNase I Footprint. Shown above is a denaturing large polyacrylamide gel, 6% PAC (19:1) and 1X TBE pH8.8, DNaseI footprint. Panels A and B show top and bottom strand labelled DNA respectively. A) Left side is R2Lp Myb, right side is R2Bm Myb. lane1 is Adenosine + Guanosine target DNA ladder, lane 2 Is DNase I treated reference DNA, lanes 3 and 4 show bound and free fraction DNA respectively. Areas of footprinting are shown as red vertical bars to the left of bou nd DNA lanes. Hypersensive sittes are shown as short horizontal bars to the left of bound DNA lanes. Panel B lane order is the same as panel A. Panel C shows target site DNA sequence with numbering. Regions of footprinting and sensitivity are mapped above and below target sequence with red bars and blue bars as in panels A and B.

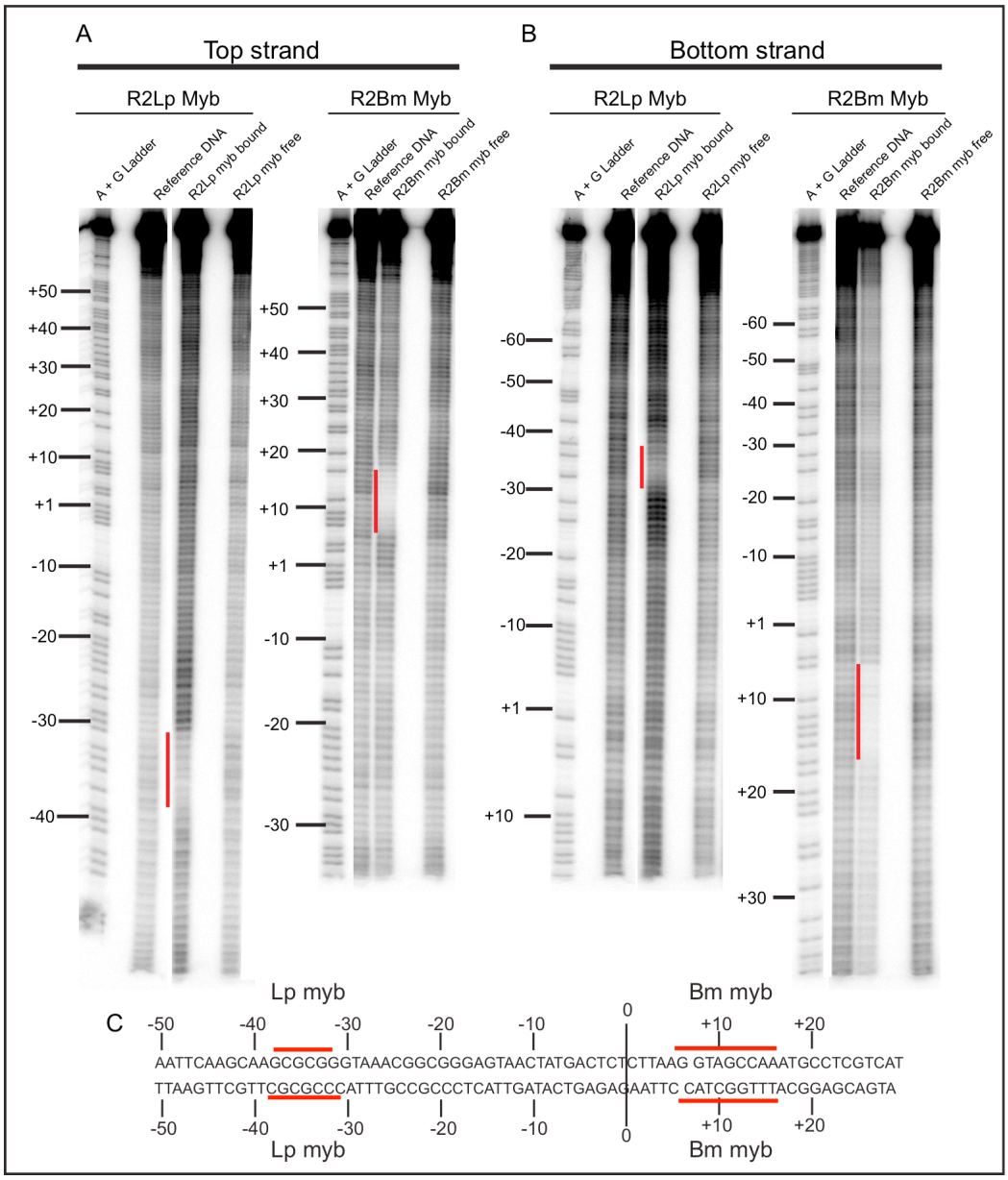


Figure 2.13: R2Lp Myb and R2Bm Myb DNase I Footprint. Shown above is a denaturing large polyacrylamide gel, 6% PAC (19:1) and 1X TBE pH8.8, missing nucleoside footprint. Figure layout is identical to figure 2.12.

DNase I footprints were performed and annotated as for R2Lp ZF1-myb protein. Figure 2.12 panel A shows the R2Lp myb domain (left) and shows R2Bm myb domain (right) footprinted on the top strand of the target site DNA. R2Lp myb shows a small footprint of 3 nt

from -38 to -36 with a hypersensitive site at -35. In contrast the R2Bm myb domain shows a longer footprint from nt +7 to +16 with no hypersensitive sites. Panel B shows R2Lp myb (left) and R2Bm (right) footprinted along the bottom strand of the target site DNA. We observed ~8 nt of binding from nt -41 to -34 with no hypersensitive sites for the R2Lp myb but observed footprinting from nt +3 to +15 for the R2Bm myb. When footprinting is mapped along the target site sequence we can clearly visualize the binding to both top and bottom strand DNA for each clone. As DNA contains approximately 10 nt per helical turn and myb domains typically bind via an alpha helix within the major groove, it was expected to observe a single footprinted region of ~10 nt or shorter. These results clearly show that the R2Bm and R2Lp myb domains bind to different regions of the target site, +3 to +16 and -41 to -34 respectively, which is predicted to be relevant to the insertion mechanism.

Missing nucleoside footprints of R2Lp and R2Bm myb domains were used to provide a high-resolution view of bound nucleotides. In Figure 2.13, panel A shows R2Lp myb (left) and R2Bm myb (right) on top strand labeled DNA. The R2Lp bound lane was slightly overloaded and as such bands appear darker throughout the gel lane. It can be seen that nucleotides -37 to -32 show significant binding. The R2Bm myb domain binding mapped to nt +6 to +16. Both regions of bound DNA are accompanied by overrepresentation of DNA fragments in the free fraction. Panel B shows R2Lp myb (left) and R2Bm myb (right) footprints on bottom strand labeled DNA. The R2Lp myb bound lane was slightly overloaded compared to free lane and appears slightly darker throughout the gel. R2Lp myb shows binding from nucleotides -38 to -31 with overrepresentation of same nucleotides in the free fraction. R2Bm myb bound lane was slightly under loaded compared to reference and free DNA, giving bands a lighter appearance. Binding was mapped from nt +6 to +16. Missing nucleoside footprint data is mapped over and under target site sequence below the gel image. These data are consistent with DNase I footprint data and with previous findings. It appears that the R2Lp myb domain is responsible for the 5’ most target site footprint of the R2Lp ZF1-myb protein. It is possible that the increased

size of the R2Lp ZF1-myb footprint is due to the zinc finger motifs. This experiment clearly shows that the R2Lp myb binds to the target site DNA upstream of the insertion site while the R2Bm myb binds to target site DNA downstream of the insertion site.

2.3.5 Zinc Finger Contribution To Target Site Recognition

Since we tested all three zinc finger motifs with the myb domain in the R2Lp ZF1-myb clone and we tested the myb domain by itself with the R2Lp myb clone, we hypothesized that the increased footprint size in the larger clone was due to the zinc finger motifs. These motifs are expected to bind nt -33 to -5 of the target DNA.

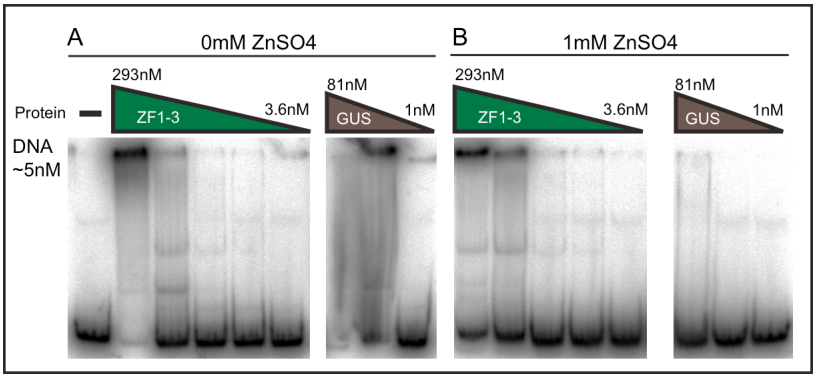


Figure 2.14: R2Lp ZF1-3 Full Target Site EMSAs. Show above is a full target site NATIVE EMSA (5% PAC 29:1, 0.5X TBE pH 8.3) comparing R2Lp ZF1-3 protein (left) to GUS control gene (right) at 0mM zinc (panel A) and 1mM zinc (panel B). Protein titrations and concentrations are indicated above lanes. All reactions contain ~5nM DNA.

To test our hypothesis we created a sub-clone to the R2Lp ZF1-3 motifs in the pDTAP expression vector. R2Lp ZF1-3 was tested for DNA binding activity using EMSAs with full target DNA. Figure 2.14 panel A shows R2Lp ZF1-3 clone (left) and GUS control gene (right) in the absence of zinc ion in binding buffer. Panel B is similar to panel A with the exception that binding reactions contained 1mM zinc ion. In the absence of zinc, we see a well complex form at high protein concentrations and several light bands of various size at lower protein concentrations. Although we see some bands formed, there does not appear to be strong DNA binding ability when compared to the GUS control gel shift. In panel B reactions performed in the presence of 1mM zinc ion do not appear to produce significant band shifts when compared

to the GUS control gene. Although reactions performed in the presence of zinc do produce shifted bands, the percentage of shift is low considering the overabundance of protein and will not shift enough total DNA to perform footprints. We performed EMSA reactions testing several parameters to optimize binding and shifting. In all we tested 5 different gel buffer systems, Tris-glycine, 1X Tris borate EDTA, 0.5X Tris borate EDTA, 1X Tris borate, and 0.5X Tris acetate EDTA. We also tested differing levels of NaCl (35-125mM) and Zn ion levels (0-50mM) in the binding reactions. In all we never observed significant DNA binding activity. There are several explanations for weak DNA binding activity. First, we could have produced a non-functional clone. We performed EMSA reactions on 4 clones that contained 2 different start and stop sites. These proteins were tested once under common EMSA conditions with no significant results. It is also possible that the zinc finger motifs do not bind DNA tight enough to produce a concise band. On several occasions we observed a reasonable amount of DNA shifted and smeared throughout the gel lane. In summary we were unable to produce a quality band shift with the R2Lp ZF1-3 clones.

To test whether the zinc fingers contribute to target site binding, we decided on a second approach. We produced zinc finger deletion mutant expression constructs named R2Lp ZF2-myb and R2Lp ZF3-myb. These constructs were assayed for DNA binding ability using full target site EMSAs. Figure 2.15 panel A shows R2Lp ZF2-myb and panel B shows R2Lp ZF3-myb. As controls we used protein minus lanes with storage buffer substituted. Panel A shows R2Lp ZF2-myb protein forming strong well complexes at 25nM and 8.3nM concentrations. Clear monomer bands become apparent when DNA is in excess. In panel B the R2Lp ZF3-myb protein shifts similar to panel A with protein forming a well complex when in excess and forming monomer bands at lower protein concentrations. Unlike R2Lp ZF1-3 protein, these zinc finger deletion mutants shift DNA with a high enough affinity for footprinting analysis.

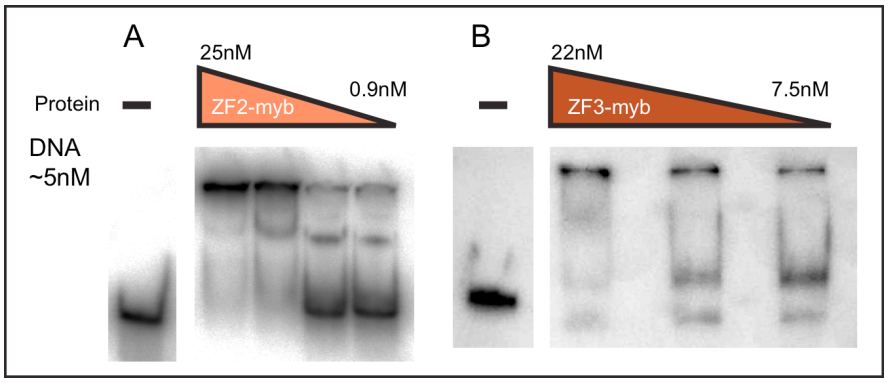


Figure 2.15: R2Lp ZF2-Myb and R2Lp ZF3-Myb EMSAs. Show above is a full target site NATIVE EMSA. Panel A contains R2Lp ZF2-myb (pDTAP version) and was run in a 6% PAC 19:1, 1X TBE pH 8.3 gel. All lanes contain ~5nM DNA. Panel B contains R2Lp ZF3-Myb pD17 protein and was performed and run at DNase I footprint conditions. Protein concentrations and titrations are shown above gel lanes.

We decided to perform DNase I footprints of the zinc finger deletion mutants in order to map the zinc finger contribution to target site recognition. Figure 2.16 panel A shows R2Lp ZF2-myb (left) and R2Lp ZF3-myb (right) proteins footprinted on top strand target DNA. R2Lp ZF2-myb protein shows strong DNA binding from nt -38 to -30 and nt -19 to -16 with two hypersensitive sites at nt -12 and -11. R2Lp ZF3-myb protein footprinted approximately the same region with binding from nt -38 to -27 and from nt -19 to -16 with no hypersensitive sites. Panel B shows R2Lp ZF2-myb (left) and R2Lp ZF3-myb (right) footprinted on bottom strand target DNA. R2Lp ZF2-myb protein bound nt -41 to -27 and nt -23 to -15 with no hypersensitive sites. R2Lp ZF3-myb protein footprinted nt -41 to -27 with two hypersensitive sites at nt -25 and -24. Given the complete data set it appears that the R2Lp ZF 2 motif is responsible for binding the target site at nt -19 to -16 and the R2Lp ZF3 motif is responsible for binding nt -30 to -27 with the rest of the footprint region accounted for by the myb domain. It is not entirely clear whether the R2Lp ZF1 motif contributes to DNA binding, as it did not produce a footprint in a complementary region in the missing nucleoside experiments. To summarize, it appears that the R2Lp zinc finger motifs contribute to target site recognition in a 3' to 5' orientation with

respect to the protein motifs, and that all the motifs strongly bind to nucleotides upstream of the target insertion site.

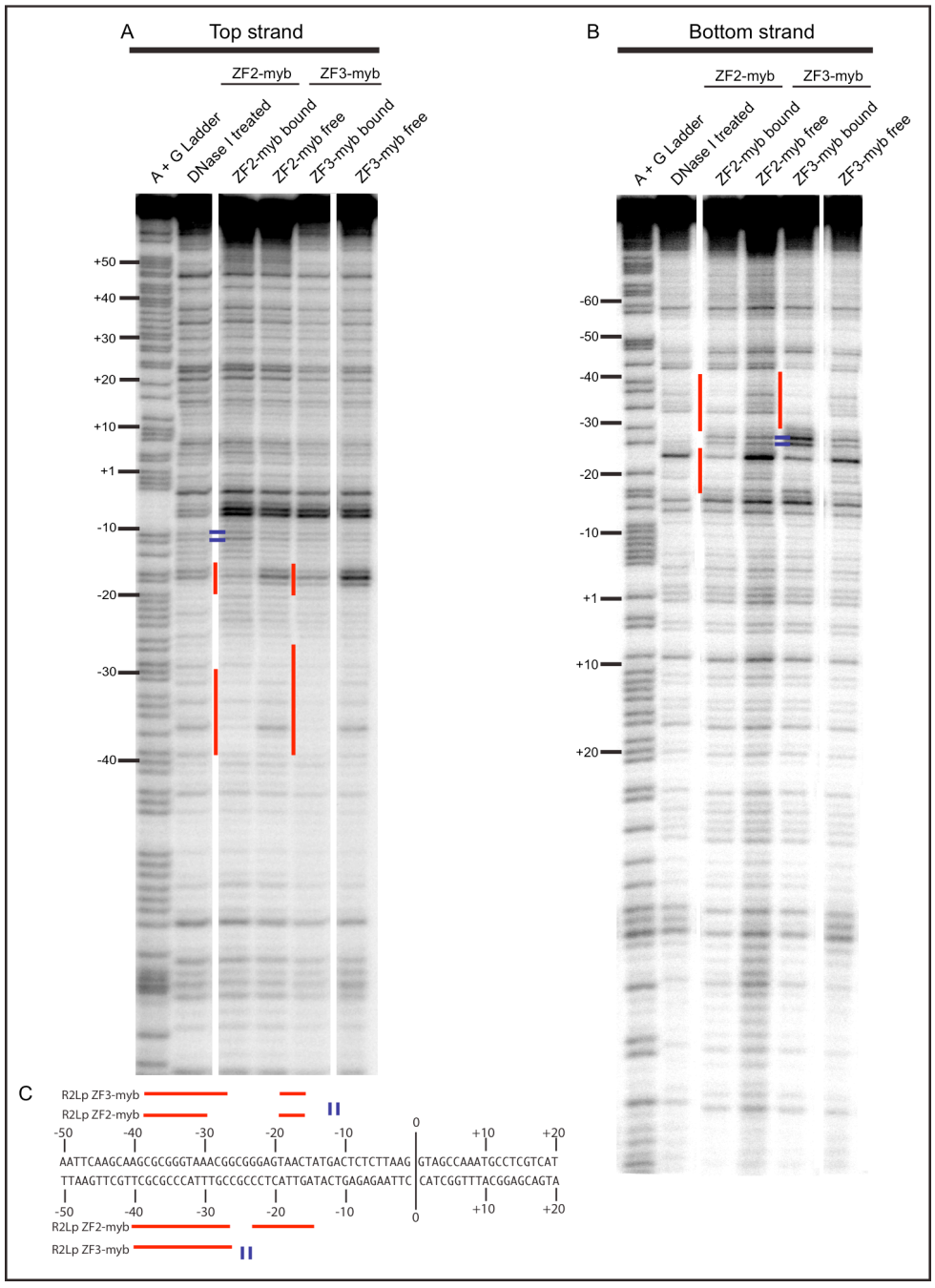


Figure 2.16: R2Lp ZF2-Myb and R2Lp ZF3-Myb DNase I Footprint. Large denaturing 6% PAC (19:1), 1X TBE pH 8.8 gel. Panel A Top strand labelled DNA land 1 A+ G ladder, lane 2 DNase I treated DNA, lane 3 and 4 ZF2-Myb pD17 bound and free fractions respectively, and lanes 5 and 6 ZF2-Myb pD17 bound and free frations respectively. Panel B is identical to panel A with bottom strand labelled DNA. Panel C summarizes footprint data.

2.4 Discussion

Results from our EMSA analysis revealed that the R2Lp full amino terminus has a high affinity for its target site. Although the R2Lp full amino terminus contains three zinc finger motifs and a myb domain, we were unable to confirm that the R2Lp zinc finger motifs have a strong affinity for target site DNA on their own. The myb domain of both R2Lp and R2Bm show strong DNA binding capacity. Inability of the zinc finger motifs of R2Lp to bind target DNA could be explained by alternative function or poor construct design. It is hypothesized that R2 elements have at least two RNA binding domains and it may be possible that the extra zinc finger motifs participate in this role (*65, 125)*. It is also possible that the zinc finger motifs recognize a protein such as the nucleosome on which the target DNA is coiled. Even though the zinc finger motifs did not show strong DNA binding ability in the EMSA reactions, they do show binding in our footprint analysis.

The R2Lp ZF1-myb's low resolution data shows three regions of binding, from nt -40 to -24, nt -20 to -15, and nt -12 to -5 with hypersensitive sites at -21, -14, -4, and +3 to +5. High-resolution data confirms region one from nt -38 to -31 and region two from nt -20 to -17. Disagreements between low-resolution and high-resolution data sets could be explained if one or more of the zinc finger motifs do not bind target DNA. Perhaps the zinc finger motifs, due to their charged nature and close proximity with target DNA, interfere with DNase I cleavage during low resolution footprinting, but are not strong enough to produce a footprint with high resolution techniques.

In contrast to our original hypothesis, the R2Lp myb domain binds to a different region of the28S rDNA gene than the R2Bm myb domain (*116)*. Both low and high-resolution footprint data show that the R2Lp myb binds from 39 to 32 nucleotides upstream (Figure 2.18.D) of the R2 element insertion site, while the R2Bm myb binds from 6 to 16 nucleotides downstream (Figure 2.18.E) of the insertion site. We did not expect to find such a drastic difference in the myb domain function between these two elements.

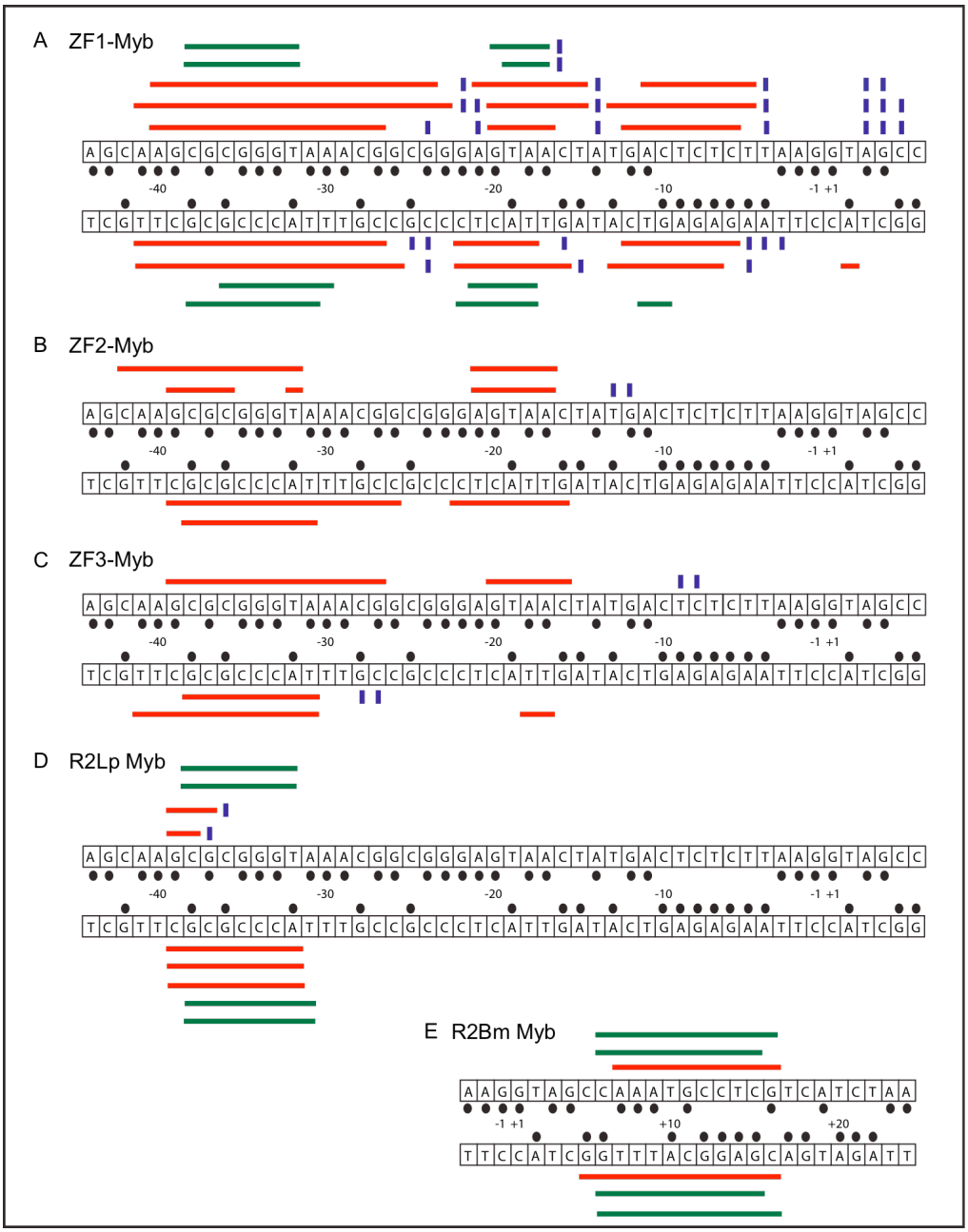


Figure 2.17: Footprint Data Summary. Shown above is an illustration of all DNA footprint data. Target site sequence is depicted as for the region of footprinting. Between top and bottom strand se3quence, black ovals represent A+G ladder bands and numbers represent nucleotides up or down stream from R2 element insertions site. Red horizontal bars represent DNase I footprint, Blue vertical bars represent hypersensitive sites, and green horizontal bars represent missing nucleoside footprint data. Multiple sets of bars represent replicates of experiments. Panel A contains data from R2Lp ZF1-Myb protein, panel B from R2Lp ZF2-Myb, Panel C from R2Lp ZF3-Myb, panel D from R2Lp Myb, and panel E R2Bm Myb.

The R2Lp ZF deletion constructs, R2Lp ZF2-myb and R2Lp ZF3-myb, show nearly identical footprinting patterns. Both constructs show binding in region one from nt -39 to -32 and in region two from nt -20 to -17. Some of the experimental replicates show larger regions of binding. Given that these constructs both contain the myb domain and zinc finger motif 3 and bind region one and region two of the full amino terminal footprint we can draw two conclusions. First, zinc finger 1 motif is likely responsible for the full amino terminus footprint region 3. Second, the zinc finger 2 motif does not appear to be involved in DNA binding. An interesting observation of the R2 A clade zinc finger motifs is that zinc fingers 1 and 3 are both CCHH type while zinc finger number 2 is a CCHC type (*67)*. Although no RNA binding experiments were performed using these constructs, the carboxyl terminal CCHC motif is proposed to bind RNA.

We are unable to confidently assign a function to zinc finger 2 with these data. We have confirmed that the R2Lp myb domain binds from nt -38 to -31, the R2Lp zinc finger number 3 binds nt -21 to -17, and the R2Lp zinc finger motif 1 may bind nt -12 to -5. In contrast the R2Bm myb domain binds the target DNA from nt +6 to +16.

From studies of the R2 target locus and evolution of the R2 elements, it is unclear why there would be two different targeting mechanisms for recognizing the same insertion site (*41)* (*43)*. Furthermore, the advantages of targeting the same insertion site in different fashions seem puzzling. It is clear that the R2 A and R2 D clade elements diverged approximately 900 million years ago and have distinct reverse transcriptase phylogenies. It is also clear that the R2 A elements encode three zinc finger motifs while R2 D elements encode a single zinc finger motif. If we choose the most parsimonious explanation, that the ancestral state of R2 elements was the R2 A clade, then the zinc finger motifs were lost in the R2 C and R2 D clades. During this divergence, the myb domain must have gained its specificity for the new site downstream of the insertion. Several hosts currently harbor R2 elements from both the R2 A and R2 D clade (*67)*. Perhaps competition between these elements into the same target site prompted a competitive evolutionary trend. If divergence of R2 elements was limited to DNA targeting and not RNA

recognition, it would be possible for one element's RNA (i.e. an R2 D member) to parasitize the ancestral element's (i.e. an R2 A member) protein during TPRT. This would allow the R2 D sequence to diverge with less constraint on function. Although this hypothesis is feasible, it is unlikely that the R2 A element would retrotranspose R2 D RNA often enough for genome maintenance, given the R2 element cis binding preference (*125)*. R2 elements are known to be regulated by epigenetic modifications in drosophila flies (*120)*. It has been show that the ribosome locus is not transcribed in its entirety, rather only a small subset of rDNA repeats are transcribed and the others are left in a highly condensed state. It is possible that R2 A and R2 D elements escape competitive transposition by virtue of host regulation. This hypothesis would explain how two R2 elements could have evolved without direct competition. There are still several unanswered questions in R2 element biology including the exact full mechanism of targeting, especially the role of zinc finger motif 2 from R2 A members.

CHAPTER 3

FUTURE EXPERIMENTS AND CONCLUSIONS

3.1 DNA Targeting

A complete targeting model for R2 elements is an essential remaining question to R2 element biology. We still have an incomplete understanding of the upstream DNA binding domain of R2Bm. We must also confirm the R2Lp target site footprint in the context of full length protein, similar to the experiments performed on R2Bm. We wish to understand whether R2 elements have one or two DNA binding domains, that is, whether R2 always functions as a dimer. We also wish to understand whether the R2 DNA binding domain is modular and can be manipulated to target a new site.

It has always been our goal to engineer R2 elements into a potential gene-targeting vector. Engineered R2 elements could be used to produce specific gene knockouts or for gene therapeutics. R2 elements recognize their own transcripts through a specific 5' and 3' RNA motif. Manipulation of the encoded ORF is not predicted to affect RNA binding by the protein. RNA could be engineered to contain both the 5' and 3' RNA motifs as well as the desired ORF. This would allow us to mobilize a foreign transcript to a new site without having to insert the R2 protein coding sequence. If the R2 element DNA binding domain is modular, then we would hope to engineer a new series of DNA binding motifs that will target a new site of our choosing. This could mean that engineering of different zinc finger and myb motifs into the amino terminus of R2 elements could result in a novel target site for the element.

At this time these experiments and hopes for R2 are much more advanced than the state of knowledge for these elements. There are some major hurdles to jump on the road to using R2 elements as a gene-targeting vector. Although there are many major steps to overcome, the benefits of such a tool could be pivotal.

3.2 R2 as a Competitive TE Integration Model

As transposable element research has become more advanced some interesting questions have emerged. For one, how do elements evolve during active competition for similar target sites, and what factors determine the winner. There are not many established systems that would allow for such research, however R2 elements could be a great model to use. There are several examples of hosts with multiple lineages of R2 elements (*43)*. Since R2 elements are presumed to be functionally active when found in a genome, it is conceivable that these elements are currently competing for an identical target site. It could be interesting to view the dynamics of a competing set of TEs and their impact on the host genome.

Studies of this type are useful for several reasons, first, they may shed light on whether it is a successful TE which drives another less successful TE to extinction. Second, this system could allow us to visualize the effectiveness of host defenses in combating TE insertions. Third, competitive insertion experiments may give us some insight as to how many transposable elements may be transpositionaly active before host fitness becomes too compromised to remain viable. Finally, this system may allow us to gain new hypotheses about how the R2 system evolved and why these elements have been so successful throughout time.

3.3 Concluding Remarks

In conclusion, the R2 elements represent a unique site-specific set of non-LTR retrotransposons that have maintained active transposition for close to one billion years. These elements are highly evolved and their integration mechanism is believed to be conserved across may RT containing systems. A greater understanding of R2 element integration will likely impact several fields of molecular biology including NLRs, Retroviruses, and the DGRs.

APPENDIX A

MOLECULAR TOOLS AND REACTION DIAGNOSTICS

A.1 pDESTTAP Vector

A.1.1 pDESTTAP Vector Map

The pDESTTAP vector is a low copy protein expression Gateway destination vector that encodes a tandem affinity 6X His tag/Maltose binding protein (MBP) tags. The vector backbone is pET45b(+) (Novagen). The MBP tag was amplified from the pMALc4X vector (New England Biolabs) and inserted into the pET45b vector between the KpnI and HindIII restriction sites. The sequence amplified included the FactorXa protease cleavage site. The destination cassette was amplified from the pDEST17 vector (Invitrogen). The sequence was amplified after the pDEST17 6X His Tag and continued to the 3' end of the attR2 site. This cassette was inserted in to the pET45b vector between the HindIII and XhoI restriction sites. The Destination cassette is in frame with the pET45b carboxyl terminal S-tag. This vector has Ampicillin resistance and is ~8100 nt in length. Vector map shows an expanded multiple cloning site that highlights significant regions to protein expression. Sequence landmarks are list to the middle left of the figure. Underneath the sequence landmarks are the nucleotides for sequencing. Beneath the vector map is a nucleotide and 3 letter amino acid representation of the tags and sequence fusions. pDESTTAPc was constructed by restricting the *HindIII* site and creating blunt ends with *Taq* polymerase followed by ligation. Subsequently pDESTTAPc is 8101nt in length.

A.1.2 pDESTTAP Vector Sequence

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTT
ACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTC
CCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTT
TAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGG
TTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCAC
GTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATT
CTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAA
CAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCTGGCGGCACGATGGC
ATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCA
ATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACC
TATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAA
CTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCAC
GCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAA
GTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGT
AAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTG
TCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTA
CATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAG
AAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACT
GTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAG
AATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGC
CACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTC
AAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCT
TCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCG
CAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATCA
TGATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGA
AAAATAAACAAATAGGTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTC
AGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCT
GCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTAC
CAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCT
AGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCT
CTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTG

CACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCT
ATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCA
GGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTAT
AGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGG
GGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGC
TGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTAC
CGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAG
TGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTA
TTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGC
CAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAA
CACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCT
GTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCG
AGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGT
TCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGC
GGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATT
TCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTAC
TGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGAT
GCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGT
AGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCA
GGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTT
GTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGT
GATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGG
AGCACGATCATGCTAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGC
TCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCG
TTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATC
GGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCA
CCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGC
AAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGC
GGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATGTCCGCACCAA
CGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCA
ACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGG
ACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATA
TTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAG
CGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTC
ATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGA
ACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGA
TCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGC
CGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAA
TCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATC
AGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCC
GCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACC
ACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACT
GGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAA
AGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGC
ATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGG
TGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCA
CGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGT
CGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGT
CCGGCGTAGACCGCTGCTGCGAAATTTGAACGCCAGCACATGGACTCGAGGATCGAGATC
GATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATT
CCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGCACATCACCACCA
CCATCACGTGGGTACCATGAAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGAT

AAAGGCTATAACGGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAG
TCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCG
ATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCT
GTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTACCTGGGA
TGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTG
ATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTG
GATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAACCGTAC
TTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAG
TACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTG
GTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTG
CCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCG
ACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAA
ACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGC
AAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGAC
AAACCGCTGGGTGCCGTAGCGCTGAAGTCTTACGAGGAAGAGTTGGTGAAAGATCCGCG
GATTGCCGCCACTATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGAT
GTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGAC
TGTCGATGAAGCCCTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAACAACAATAAC
AATAACAACAACCTCGGGATCGAGGGAAGGATTGAAGCTTTCGAATCAACAAGTTTGTACA
AAAAAGCTGAACGAGAAACGTAAAATGATATAAATATCAATATATTAAATTAGATTTTGCATA
AAAAACAGACTACATAATACTGTAAAACACAACATATCCAGTCACTATGGCGGCCGCATTA
GGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGGATTTTGAGTTAGG
ATCCGTCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATAC
CACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTC
AATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAA
AATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCC
GGAATTCCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGT
TACACCGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACG
ATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGC
CTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTT
TCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATG
GGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCAT
GCCGTCTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATG
AGTGGCAGGGCGGGGCGTAAAGATCTGGATCCGGCTTACTAAAAGCCAGATAACAGTATG
CGTATTTGCGCGCTGATTTTTGCGGTATAAGAATATATACTGATATGTATACCCGAAGTATG
TCAAAAAGAGGTGTGCTATGAAGCAGCGTATTACAGTGACAGTTGACAGCGACAGCTATCA
GTTGCTCAAGGCATATATGATGTCAATATCTCCGGTCTGGTAAGCACAACCATGCAGAATG
AAGCCCGTCGTCTGCGTGCCGAACGCTGGAAAGCGGAAAATCAGGAAGGGATGGCTGAG
GTCGCCCGGTTTATTGAAATGAACGGCTCTTTTGCTGACGAGAACAGGGACTGGTGAAAT
GCAGTTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGATGTACAG
AGTGATATTATTGACACGCCCGGGCGACGGATGGTGATCCCCCTGGCCAGTGCACGTCTG
CTGTCAGATAAAGTCTCCCGTGAACTTTACCCGGTGGTGCATATCGGGGATGAAAGCTGG
CGCATGATGACCACCGATATGGCCAGTGTGCCGGTCTCCGTTATCGGGGAAGAAGTGGCT
GATCTCAGCCACCGCGAAAATGACATCAAAAACGCCATTAACCTGATGTTCTGGGGAATAT
AAATGTCAGGCTCCCTTATACACAGCCAGTCTGCAGGTCGACCATAGTGACTGGATATGTT
GTGTTTTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATTTAATATATTGATATTTAT
ATCATTTTACGTTTCTCGTTCAGCTTTCTTGTACAAAGTGGTTGATCTCGAGTCTGGTAAAG
AAACCGCTGCTGCGAAATTTGAACGCCAGCACATGGACTCGTCTACTAGCGCAGCTTAATT
AACCTAGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG
GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

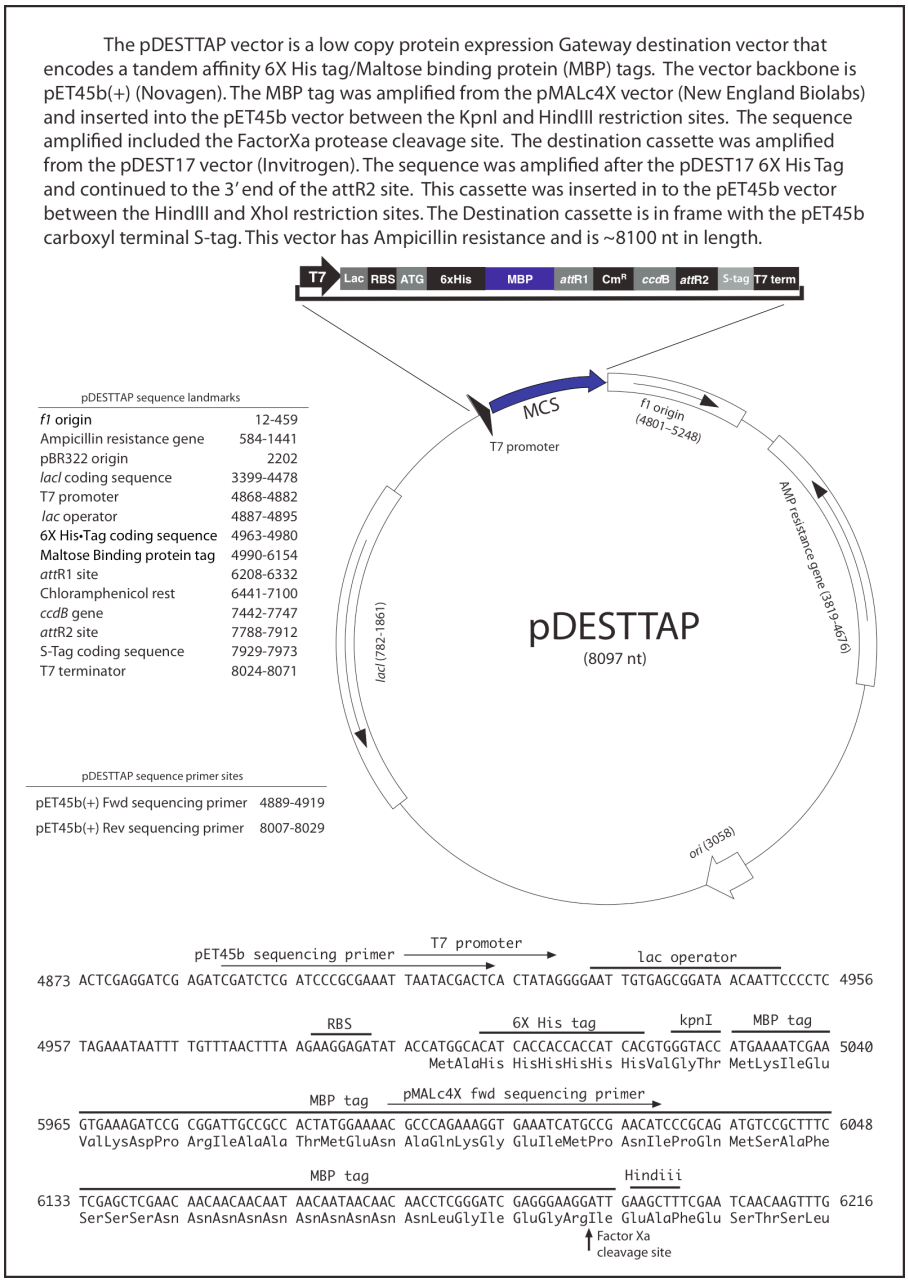


Figure A.1: pDESTTAP Bacterial Protein Expression Vector Map. Shown above is the pDESTTAP protein expression vector map. Center left, pDESTTAP sequence landmarks, includes a list of the nucleotides for important vector sequence regions. Below are the nucleotides of sequencing primers. center right, pDESTTAP vector map, included are the origin of replication, antibiotic resistance genes, and multiple cloning site. Above the vector map is an expanded view of the multiple cloning site. Bottom center, an illustration of pDESTTAP vector nucleotide sequence and amino acid translation highlighting tags and protease cleavage sites.

A.2 Footprint Diagnostics

In order to adapt established protocols to our specific needs we performed a diagnostic DNase I DNA cleavage assay and a hydroxyl radical DNA cleavage assay. All buffer conditions were as listed in the experimental methods section (2.2.6). DNase I cleavage reactions were performed at full scale with 1uL target DNA per reaction. Hydroxyl radical diagnostics were performed on 5uL of DNA for easy scale up.

A.2.1 DNAse I Cleavage Diagnostic

Figure shows a DNAse I diagnostic cleavage reaction. Each lane was treated with 3uL of DNase I(1U/uL) diluted to the value listed at the top of the lane. Gel shown is a 6% PAC (19:1) TBE pH 8.8 large denaturing gel. It was decided that 3uL of a 250 fold dilution of DNase I (lane 6, 243) was optimum to cleave a 150 nt target probe at 2% glycerol in 2 minutes at 25C.

A.2.2 Hydroxyl Radical/Missing Nucleoside Cleavage Diagnostic

Figure shows a 6% PAC (19:1) TBE pH 8.8 large denaturing gel of hydroxyl radical treated target DNA probe. All buffer concentrations were held constant at conditions listed in 2.2.4. Hydrogen peroxide was titrated from a concentration of 0.633% (lane 1) to 0.0084% (Lane 16) in a 3/4 dilution series (75uL of hydrogen peroxide solution into 25uL of water). It was decided that 0.02% (lane 13) hydrogen peroxide was optimum for cleaving a 150nt probe in 2 minutes at 25C.

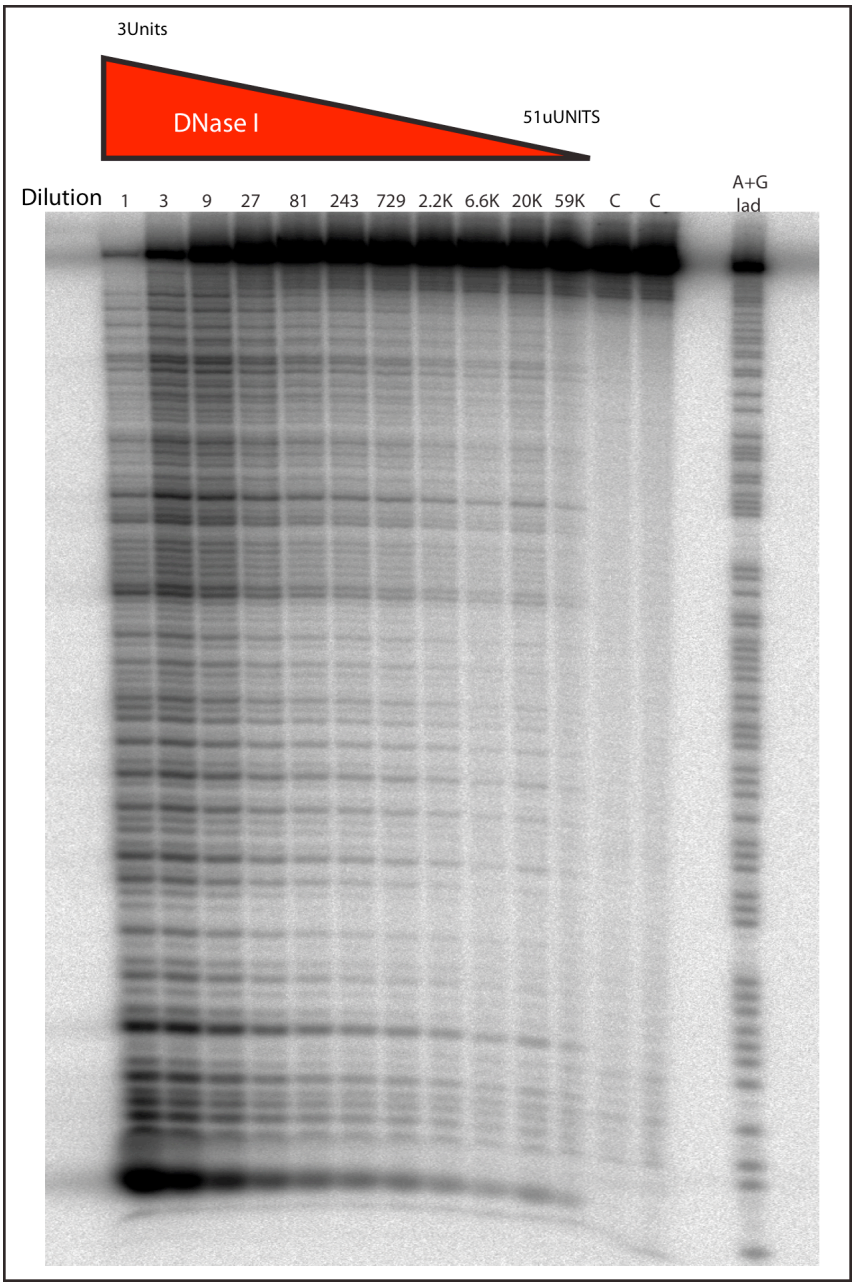


Figure A.2: DNase I Cleavage Diagnostic. Show above is a DNA probe cleavage diagnostic using a titration of DNase I enzyme. enzyme was diluted to values listed above lanes from ~3Units to 51uUnits. Lanes C contain untreated DNA. A+G lad is an Andenosine + Guanosine Molecular weight ladder.

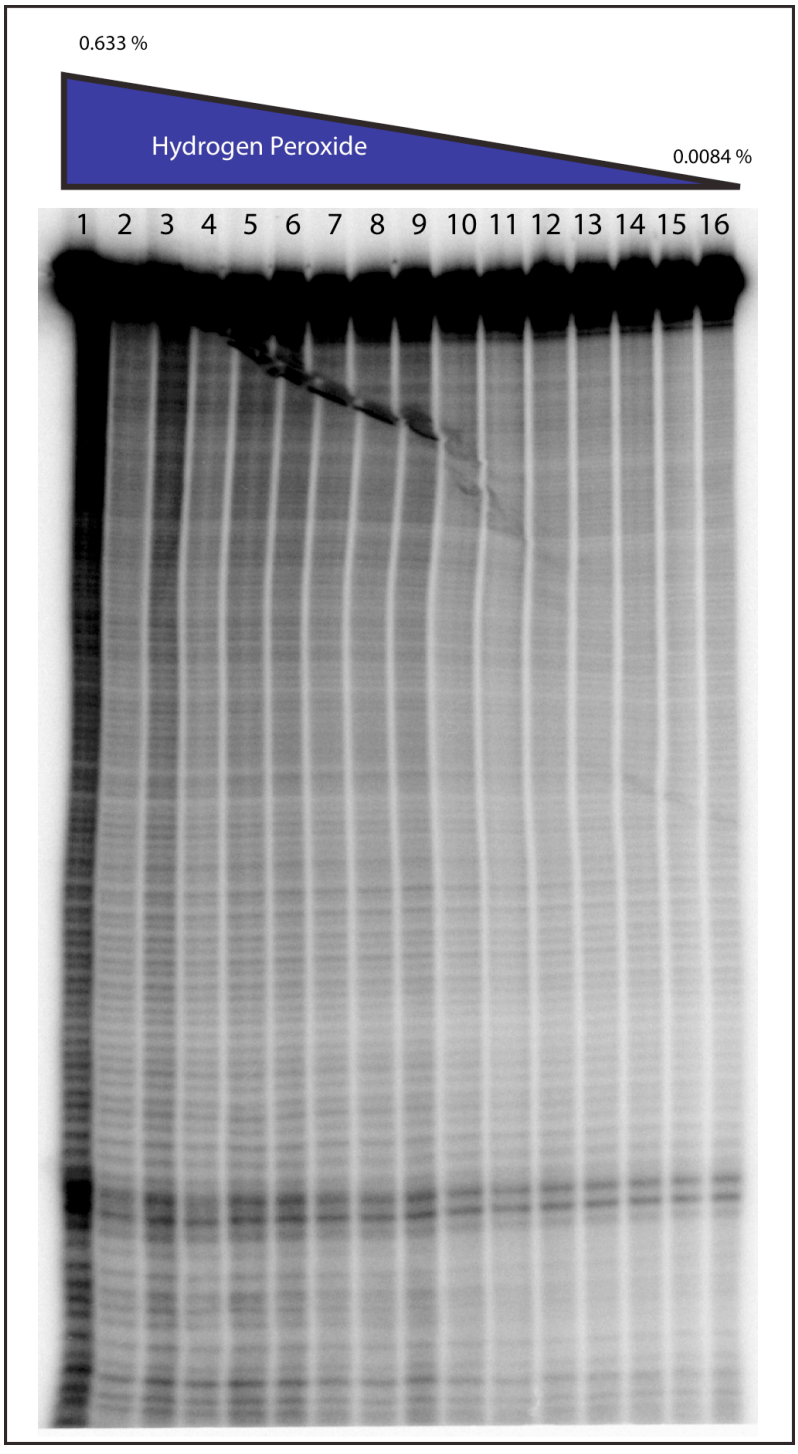


Figure A.3: Missing nucleoside Cleavage Diagnostic. Shown above is a DNA probe cleavage diagnostic reaction containing a titration of hydrogen peroxide. lanes 1-16 contain a 3/4 dilution series of hydrogen peroxide beginning at 0.633% and ending at 0.0084%.

APPENDIX B

ZINC FINGER DIAGNOSTICS AND FOOTPRINT REPLICATES

B.1 R2Lp ZF1-3 EMSAs

As mentioned in section 2.3.6 we were unable to perform a successful EMSA reaction with the R2Lp ZF1-3 clone regardless of reaction and gel buffer system used. First we tried to perform 7% PAC (19:1) EMSAs in quadruplicate using four different buffer systems (A=1X Tris/Glycine, B=1X Tris/Borate, C= 1X Tris/Borate/EDTA, and D=0.5X Tris/Borate/EDTA) while titrating differing concentrations of zinc (figure B.1). Zinc concentrations and protein dilutions are listed above gel lanes. DNA concentration was held constant at ~5nM. ZF1-3 protein dilution as follows; 1 ~293nM, 3~97.6nM, 9 ~32.5nM, 27 ~10.9nM, 81 ~3.6nM, and 243 ~1.2nM. Panel B shows a slight monomer band at 1/27 and 1/81 diluted protein in 0mM zinc. Panel C shows a slight monomer and dimer band that run at different heights depending of zinc concentration. Panel D shows similar results to panel C.

Second we attempted to assay the effect of polyacrylamide matrix on protein DNA complex migration by using 5% PAC (29:1) EMSAs (figure B.2). Gel buffers were A= 0.5X Tris/Borate/EDTA and B=1X Tris/Borate.Results were similar to figure B.1.D and B.1.B respectively.

Third we attempted to test the effect of 1X Tris/Glycine (A) and 0.5X Tris/Acetate/EDTA (B) at conditions similar to the previous experiment (figure B.3). In these reactions we included the GUS control gene for comparison, diluted concentrations are 1 ~81nM, 9 ~9nM, and 81 ~ 1nM. Left side of gels contain no zinc and right side of gels contain 1mM zinc. Left side of gels contain no zinc while the right side of the gel contained ~1mM zinc. Gus control genes show similar DNA shifting at all conditions compared to R2Lp ZF1-3 protein. No significant binding was observed for R2Lp ZF1-3 protein.

Finally we attempted to test the affect of different NaCl concentrations in the presence (A) 0r absence (B) of zinc on protein binding (figure B.4). We tested 35mM (left side) 75mM (center) and 125mM (right side) NaCl concentrations. Protein dilutions and presence of zinc are indicated above gel. DNA binding ability appears similar to figure B.3 and was concluded as lacking significance. We decided not to pursue any more trouble shooting with this construct.

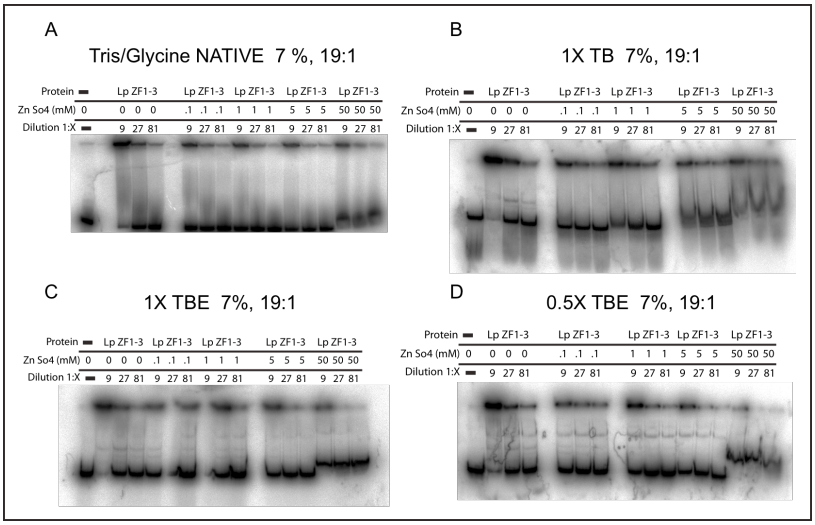


Figure B.1: R2Lp ZF1-3 Clone Diagnostic EMSAs #1. Shown above is the first set of ZF1-3 diagnostic EMSAs. Panels A-D contain the same lane contents, ~5nM DNA probe, 0-50mM zinc, and titrated ZF1-3 protein using four different gel buffer systems in 7% PAC (19:1). Concentration of zinc and protein dilution added are indicated above gel lanes. A is a 1X Tris/Glycine buffered gel, B is a 1X Tris/Borate buffered gel, C is a 1X Tris/Borate/EDTA buffered gel, and D is a 0.5X Tris/Borate/EDTA buffered gel.

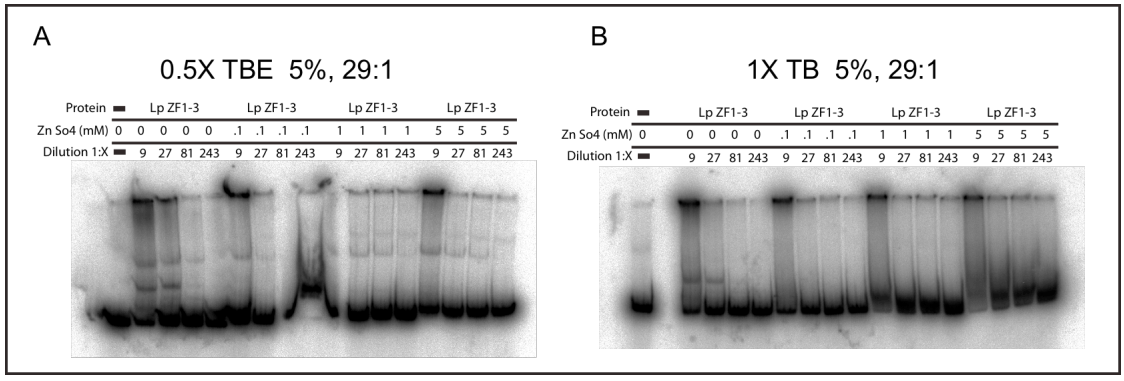


Figure B.2: R2Lp ZF1-3 Clone Diagnostic EMSAs #2. Shown above is the second set of ZF1-3 diagnostic EMSAs. Panels A and B contain the same lane contents, ~5nM DNA probe, 0-5mM zinc, and titrated ZF1-3 protein using two different gel buffer systems in 5% PAC (29:1). Concentration of zinc and protein dilution added are indicated above gel lanes. A is a 0.5X Tris/Borate/EDTA buffered gel and B is a 1X Tris/Borate buffered gel.

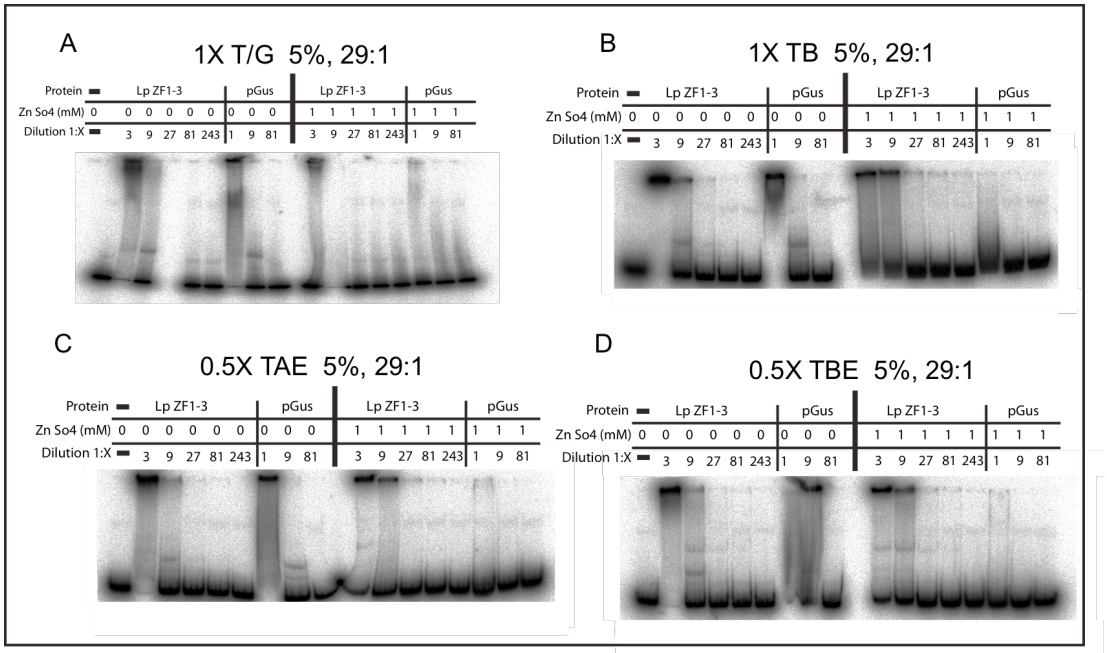


Figure B.3: R2Lp ZF1-3 Clone Diagnostic EMSAs #3. Shown above is the third set of ZF1-3 diagnostic EMSAs. Panels A and B contain the same lane contents, ~5nM DNA probe, 0 or 1mM zinc, and titrated ZF1-3 or GUS control gene protein using four different gel buffer systems in 5% PAC (29:1). Concentration of zinc and protein dilution added are indicated above gel lanes. A is a 1X Tris/Glycine buffered gel, B is a 1X Tris/Borate buffered gel, C is a 0.5X Tris/Acetate/EDTA buffered gel, and D is a 0.5 X Tris/Borate/EDTA buffered gel.

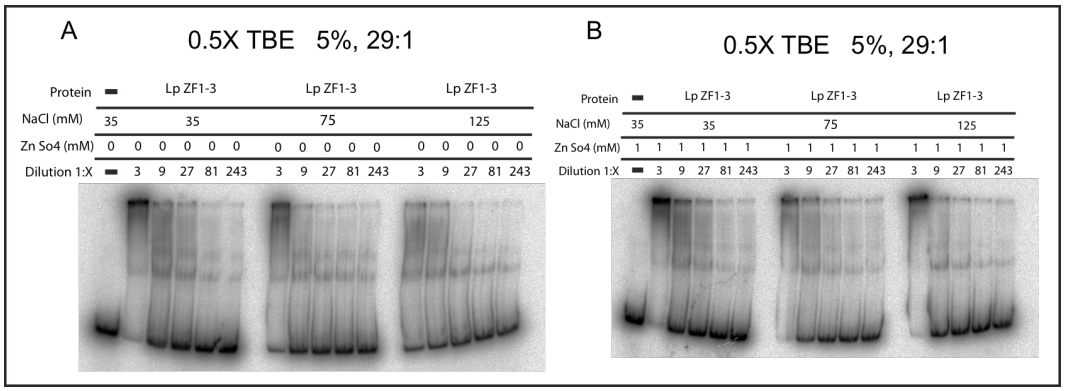


Figure B.4: R2Lp ZF1-3 Clone Diagnostic EMSAs #4. Shown above is the fourth and last set of ZF1-3 diagnostic EMSAs. Panels A and B contain the same DNA concentration (~5nM) and titrated ZF1-3 protein in a 0.5 Tris/Borate/EDTA 5% PAC (29:1) gel. Panel A contains 0mM zinc and a NaCl at three differnt concentrations, 35mM (left), 75mM (center), and 125mM (right) while panel B contains 1mM zinc and the same NaCl concentrations. Concentration of zinc, NaCl, and protein dilution added are indicated above gel lanes.

B.2 Footprint Replicates

Within the results section of chapter 2 (2.3) Footprint gels were cropped and organized for clarity. Included below are the annotated forms of unmodified footprint gel images. All gels are 6% PAC (19:1) 1X TBE pH 8.8 large denaturing gels. From top to bottom of all gel figures is the DNA stand, lane contents, gel image, and footprint data summary. A comprehensive and summarized view of all footprint data may be viewed in the discussion section of chapter 2 (section 2.4, figure 2.18).

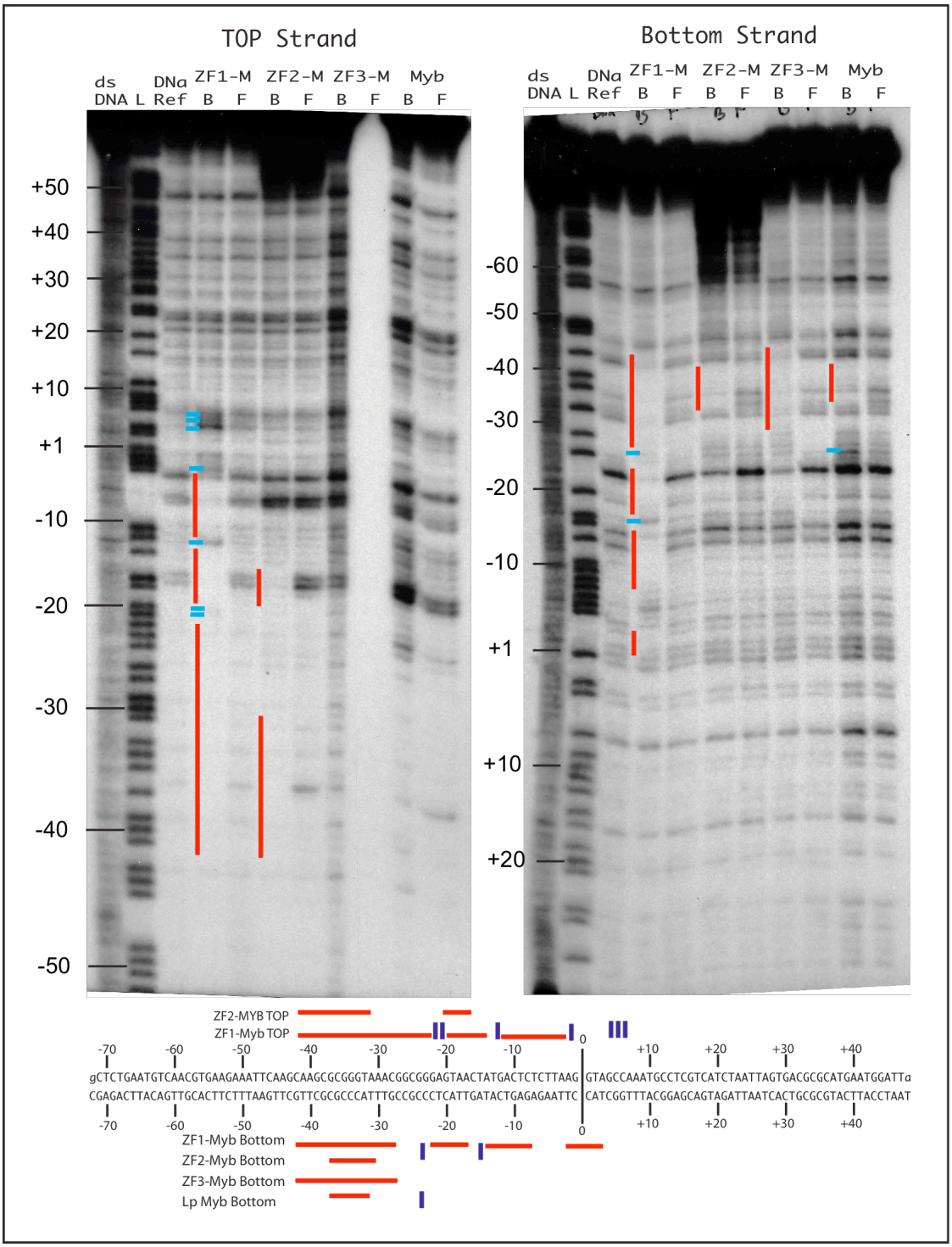


Figure B.5: DNase Footprint #1. Shown above is the first DNase I footprint. Top strand DNA left and bottom strand DNA right. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction. Protein constructs used are marked above lane. Footprinting is summarized at the bottom.

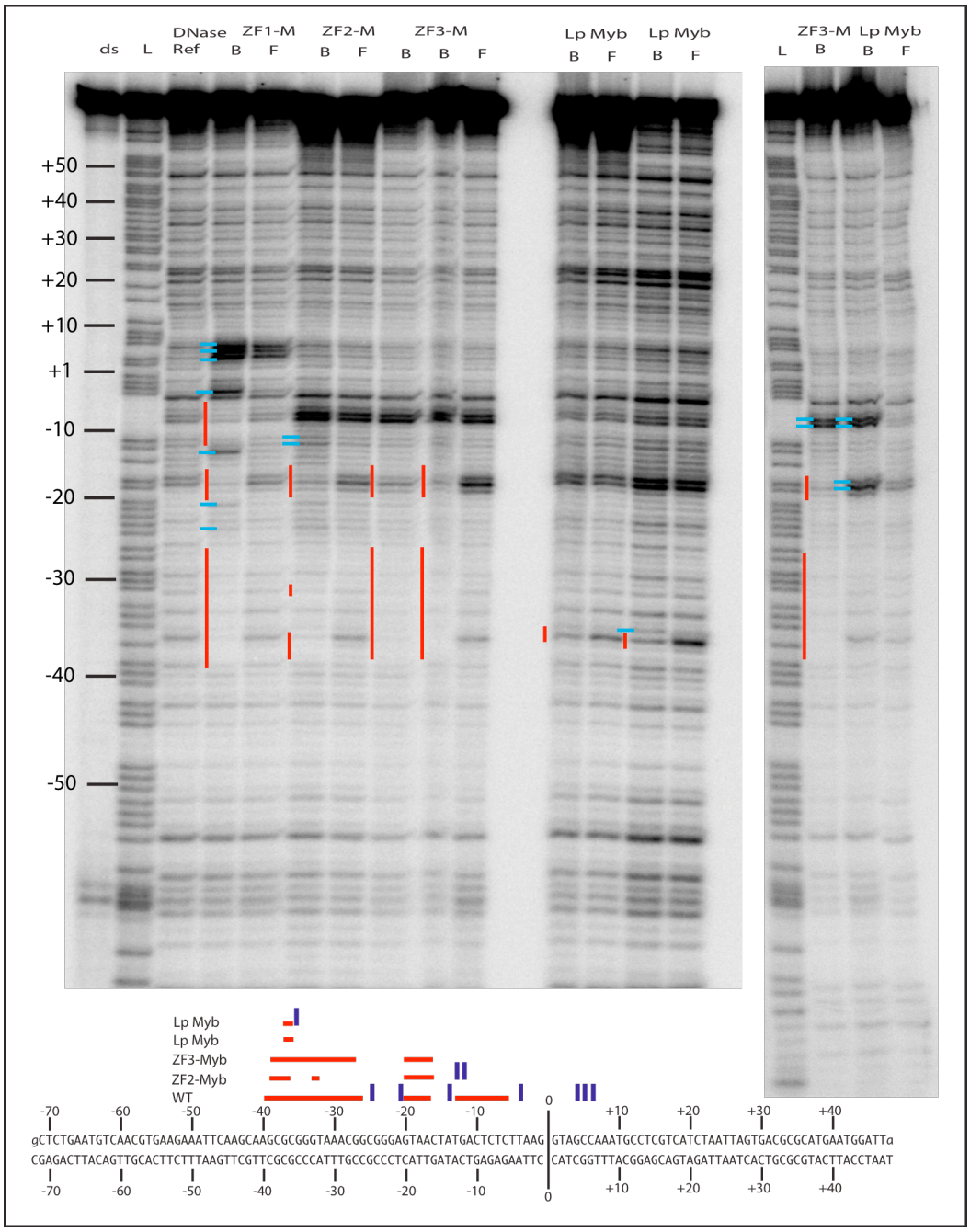


Figure B.6: DNase Footprint #2 Top Strand. Shown above is the top strand of the second DNase I footprint. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction. Protein constructs used are marked above lane. Footprinting is summarized at bottom of figure.

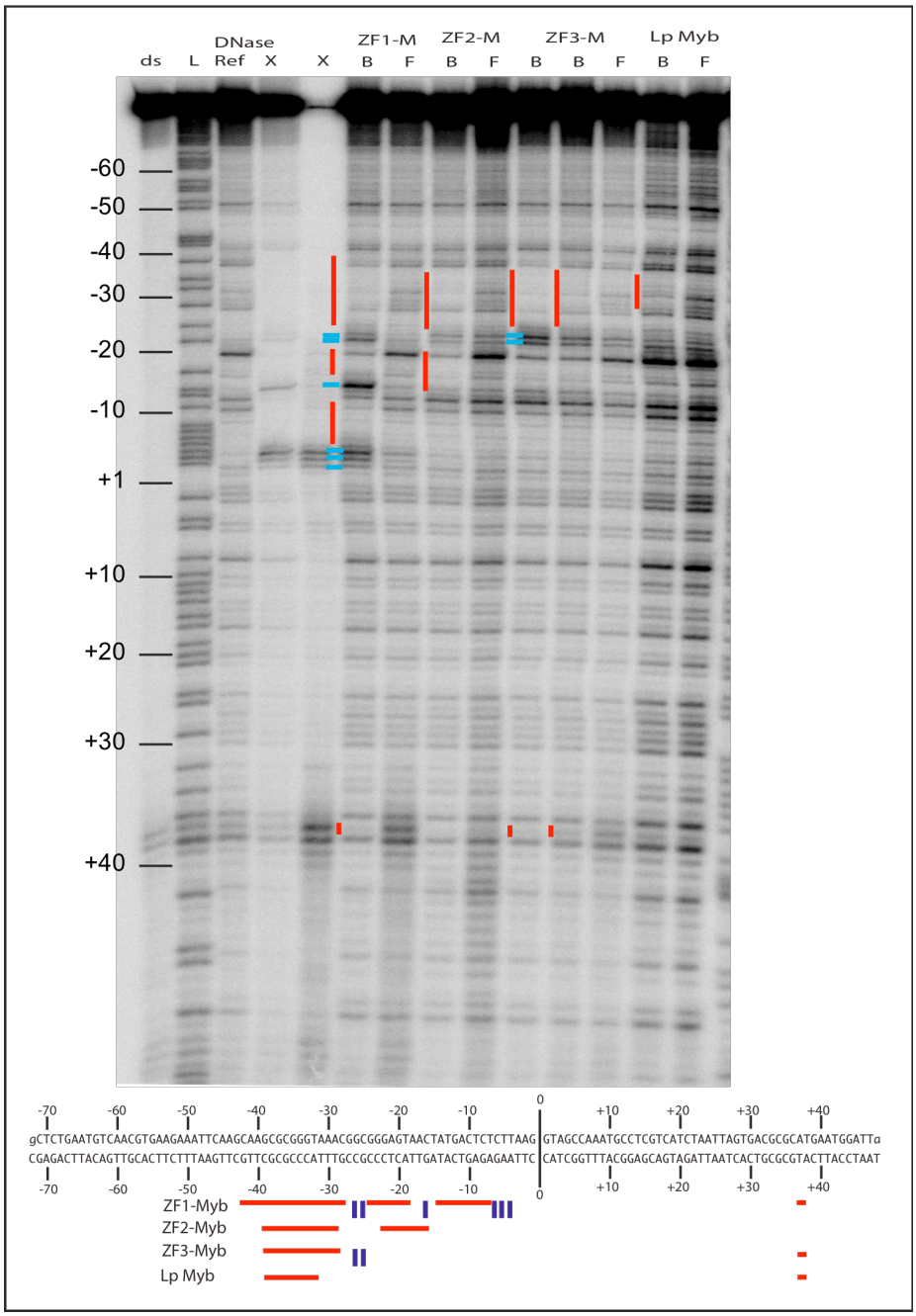


Figure B.7: DNase Footprint #2 Bottom Strand. Shown above is the bottom strand of the second DNase I footprint. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction, X = unloaded or misloaded lane. Protein constructs used are marked above lane. Footprinting is summarized at bottom of figure.

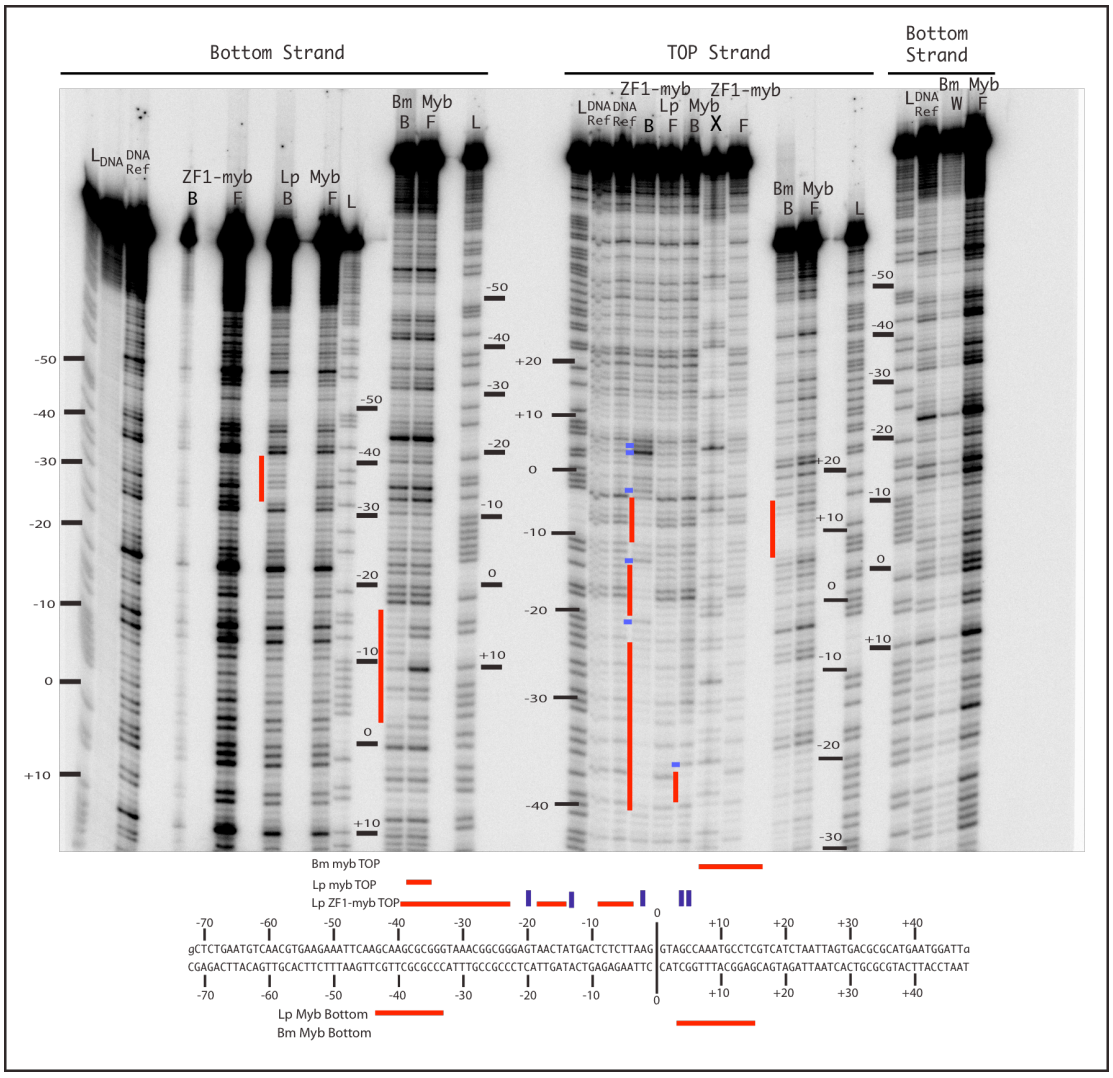


Figure B.8: DNase Footprint #3. Shown above is the third DNase I footprint. Top strand DNA middle right and bottom strand DNA left and far right. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction, X = unloaded or misloaded lane. Protein constructs used are marked above lane. Footprinting is summarized at the bottom. Top strand section lanes as follows, 1 L, 2+3 DNase reference, 4 ZF1-M B, 5 Lp Myb F, 6 Lp Myb B, and 7 ZF1-Myb B.

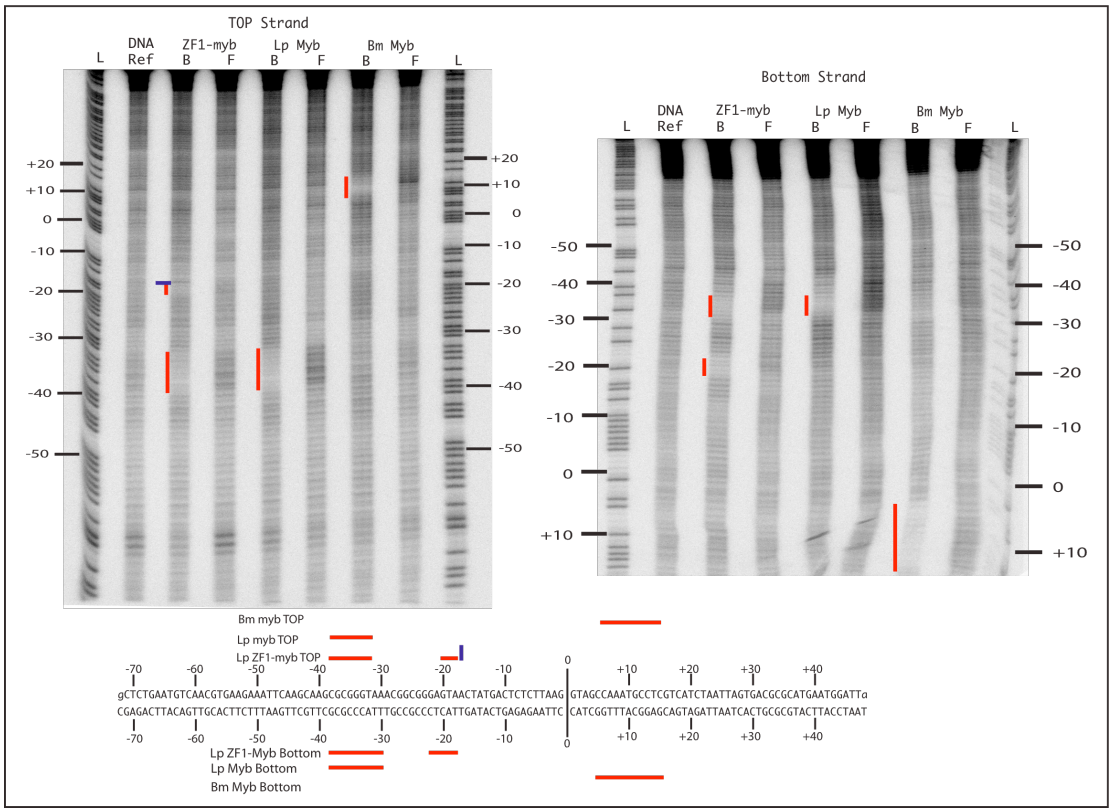


Figure B.9: Missing Nucleoside Footprint #1. Shown above is the first missing nucleoside footprint. Top strand DNA left and bottom strand DNA right. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction. Protein constructs used are marked above lane. Footprinting is summarized at the bottom.

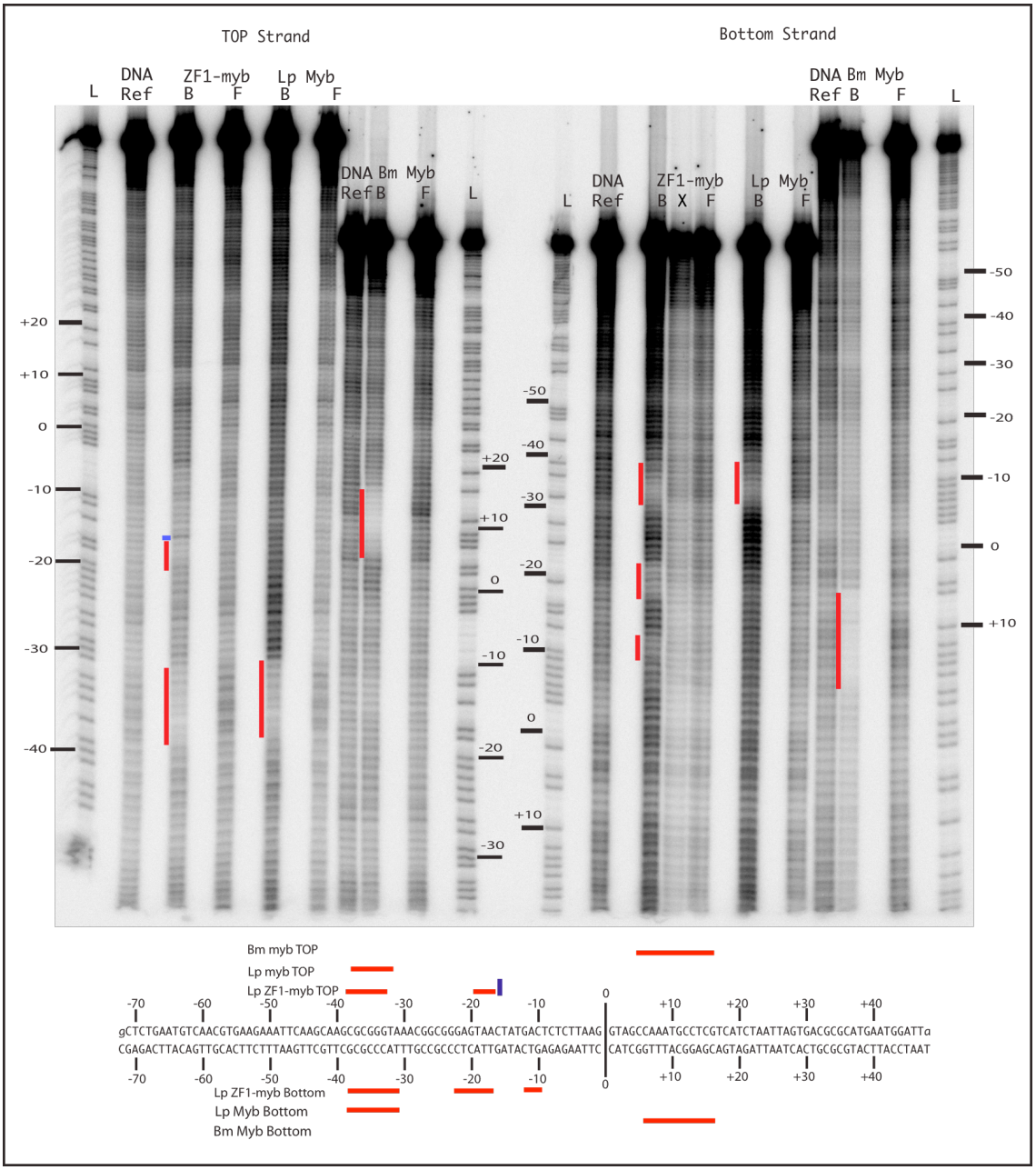


Figure B.10: Missing Nucleoside Footprint #2. Shown above is the second missing nucleoside footprint. Top strand DNA left and bottom strand DNA right. A+G ladder nucleotides are marked according to the target sequence diagram (bottom). L = A+G ladder, B = bound fraction, F = free fraction, X = unloaded or misloaded. Protein constructs used are marked above lane. Footprinting is summarized at the bottom.

REFERENCES

1. E. V. Koonin, T. G. Senkevich, V. V. Dolja, *Biol Direct* **1**, 29 (2006).
2. E. V. Koonin, *Curr Biol* **15**, R167 (2005).
3. I. H. G. S. Consortium, *Nature* **409**, 860 (2001).
4. I. H. G. S. Consortium, *Nature* **431**, 931 (2004).
5. E. Birney *et al.*, *Nature* **447**, 799 (2007).
6. H. H. J. Kazazian, J. V. Moran, *Nat Genet* **19**, 19 (1998).
7. A. Zuccolo *et al.*, *BMC Evol Biol* **7**, 152 (2007).
8. M. G. Kidwell, D. R. Lisch, in *Transposable Elements as Sources of Genomic Variation*, Craig, NL, Craigie, R, Gellert, M, A. M. Lambowitz, Eds. (ASM Press, Washington, DC, 2002), pp. 59-92.
9. M. G. Kidwell, *Genetica* **115**, 49 (2002).
10. S. M. Lutz, B. J. Vincent, H. H. J. Kazazian, M. A. Batzer, J. V. Moran, *Am J Hum Genet* **73**, 1431 (2003).
11. P. L. Deininger, M. A. Batzer, *Genome Res* **12**, 1455 (2002).
12. I. Arkhipova, M. Meselson, *Proc Natl Acad Sci U S A* **97**, 14473 (2000).
13. I. R. Arkhipova, H. G. Morrison, *Proc Natl Acad Sci U S A* **98**, 14497 (2001).
14. A. K. Khatua, H. E. Taylor, J. E. Hildreth, W. Popik, *Virology* **400**, 68 (2010).
15. A. E. Hulme, H. P. Bogerd, B. R. Cullen, J. V. Moran, *Gene* **390**, 199 (2007).
16. Y. L. Chiu, W. C. Greene, *Philos Trans R Soc Lond B Biol Sci* (2008).
17. J. Roman-Gomez *et al.*, *Oncogene* **24**, 7213 (2005).
18. V. P. Belancio, D. J. Hedges, P. Deininger, *Nucleic Acids Res* **34**, 1512 (2006).
19. D. J. Obbard, K. H. Gordon, A. H. Buck, F. M. Jiggins, *Philos Trans R Soc Lond B Biol Sci* **364**, 99 (2009).
20. N. Yang, H. H. J. Kazazian, *Nat Struct Mol Biol* **13**, 763 (2006).
21. M. Ghildiyal *et al.*, *Science* **320**, 1077 (2008).
22. A. Girard, G. J. Hannon, *Trends Cell Biol* (2008).
23. J. N. Volff, *Bioessays* **28**, 913 (2006).
24. D. J. Obbard, D. J. Finnegan, *Curr Biol* **18**, R561 (2008).
25. H. P. Cam, K. Noma, H. Ebina, H. L. Levin, S. I. Grewal, *Nature* **451**, 431 (2008).
26. E. A. Gladyshev, I. R. Arkhipova, *Proc Natl Acad Sci U S A* **104**, 9352 (2007).
27. H. Takahashi, S. Okazaki, H. Fujiwara, *Nucleic Acids Res* **25**, 1578 (1997).
28. T. Anzai, H. Takahashi, H. Fujiwara, *Mol Cell Biol* **21**, 100 (2001).
29. H. Fujiwara, M. Osanai, T. Matsumoto, K. K. Kojima, *Chromosome Res* **13**, 455 (2005).
30. Y. Kubo, S. Okazaki, T. Anzai, H. Fujiwara, *Mol Biol Evol* **18**, 848 (2001).
31. R. Cordaux, S. Udit, M. A. Batzer, C. Feschotte, *Proc Natl Acad Sci U S A* **103**, 8101 (2006).
32. C. Casola, A. M. Lawing, E. Betran, C. Feschotte, *Mol Biol Evol* (2007).
33. C. Feschotte, *Nat Rev Genet* **9**, 397 (2008).
34. V. V. Kapitonov, J. Jurka, *PLoS Biol* **3**, e181 (2005).
35. C. P. Lu, J. E. Posey, D. B. Roth, *Nucleic Acids Res* (2008).
36. C. Feschotte, *Proc Natl Acad Sci U S A* **103**, 14981 (2006).
37. T. Vandendriessche, Z. Ivics, Z. Izsvak, M. K. Chuah, *Blood* (2009).
38. Z. Ivics, Z. Izsvak, *Curr Gene Ther* **6**, 593 (2006).
39. W. C. r. Lathe, T. H. Eickbush, *Mol Biol Evol* **14**, 1232 (1997).

40. J. K. n. Pace, C. Gilbert, M. S. Clark, C. Feschotte, *Proc Natl Acad Sci U S A* **105**, 17023 (2008).
41. H. S. Malik, W. D. Burke, T. H. Eickbush, *Mol Biol Evol* **16**, 793 (1999).
42. C. Feschotte, E. J. Pritham, *Annu Rev Genet* **41**, 331 (2007).
43. K. K. Kojima, H. Fujiwara, *Mol Biol Evol* **22**, 2157 (2005).
44. X. Zhang, M. T. Eickbush, T. H. Eickbush, *Genetics* (2008).
45. T. H. Eickbush, H. S. Malik, in *Origins and Evolution of Retrotransposons*, Craig, NL, Craigie, R, Gellert, M, A. M. Lambowitz, Eds. (ASM Press, Washington, DC, 2002), pp. 1111-1146.
46. C. Bartolome, X. Bello, X. Maside, *Genome Biol* **10**, R22 (2009).
47. C. Gilbert, J. K. Pace, C. Feschotte, *Commun Integr Biol* **2**, 117 (2009).
48. M. J. Curcio, K. M. Derbyshire, *Nat Rev Mol Cell Biol* **4**, 865 (2003).
49. S. Doulatov *et al.*, *Nature* **431**, 476 (2004).
50. D. M. Simon, S. Zimmerly, *Nucleic Acids Res* **36**, 7219 (2008).
51. M. J. Curcio, M. Belfort, *Proc Natl Acad Sci U S A* **104**, 9107 (2007).
52. A. M. Lambowitz, S. Zimmerly, *Annu Rev Genet* **38**, 1 (2004).
53. T. M. Bryan, T. R. Cech, *Curr Opin Cell Biol* **11**, 318 (1999).
54. T. R. Cech, T. M. Nakamura, J. Lingner, *Biochemistry (Mosc)* **62**, 1202 (1997).
55. J. V. Moran *et al.*, *Mol Cell Biol* **15**, 2828 (1995).
56. S. Zimmerly, J. V. Moran, P. S. Perlman, A. M. Lambowitz, *J Mol Biol* **289**, 473 (1999).
57. B. Medhekar, J. F. Miller, *Curr Opin Microbiol* **10**, 388 (2007).
58. T. H. Eickbush, V. K. Jamburuthugoda, *Virus Res* (2008).
59. I. R. Arkhipova, K. I. Pyatkov, M. Meselson, M. B. Evgen'ev, *Nat Genet* **33**, 123 (2003).
60. G. T. Lyozin *et al.*, *J Mol Evol* **52**, 445 (2001).
61. K. I. Pyatkov, I. R. Arkhipova, N. V. Malkova, D. J. Finnegan, M. B. Evgen'ev, *Proc Natl Acad Sci U S A* **101**, 14719 (2004).
62. Y. E. Xiong, T. H. Eickbush, *Cell* **55**, 235 (1988).
63. J. Yang, H. S. Malik, T. H. Eickbush, *Proc Natl Acad Sci U S A* **96**, 7847 (1999).
64. D. D. Luan, M. H. Korman, J. L. Jakubczak, T. H. Eickbush, *Cell* **72**, 595 (1993).
65. S. M. Christensen, T. H. Eickbush, *Mol Cell Biol* **25**, 6617 (2005).
66. J. V. Moran, *Genetica* **107**, 39 (1999).
67. W. D. Burke, H. S. Malik, J. P. Jones, T. H. Eickbush, *Mol Biol Evol* **16**, 502 (1999).
68. E. Berezikov, A. Bucheton, I. Busseau, *Genome Biol* **1**, RESEARCH0012 (2000).
69. J. K. Biedler, Z. Tu, *BMC Evol Biol* **7**, 112 (2007).
70. W. D. Burke, H. S. Malik, S. M. Rich, T. H. Eickbush, *Mol Biol Evol* **19**, 619 (2002).
71. W. D. Burke, F. Muller, T. H. Eickbush, *Nucleic Acids Res* **23**, 4628 (1995).
72. W. D. Burke, D. Singh, T. H. Eickbush, *Mol Biol Evol* **20**, 1260 (2003).
73. A. Gabriel *et al.*, *Mol Cell Biol* **10**, 615 (1990).
74. K. K. Kojima, H. Fujiwara, *Genome Res* **15**, 1106 (2005).
75. K. K. Kojima, T. Matsumoto, H. Fujiwara, *Mol Cell Biol* **25**, 7675 (2005).
76. M. Komatsu, K. Shimamoto, J. Kyozuka, *Plant Cell* **15**, 1934 (2003).
77. H. S. Malik, T. H. Eickbush, *Mol Biol Evol* **15**, 1123 (1998).
78. H. S. Malik, T. H. Eickbush, *Genetics* **154**, 193 (2000).
79. T. Matsumoto, M. Hamada, M. Osanai, H. Fujiwara, *Mol Cell Biol* **26**, 5168 (2006).
80. O. Novikova *et al.*, *BMC Evol Biol* **7**, 93 (2007).
81. J. Permanyer, R. Gonzalez-Duarte, R. Albalat, *Genome Biol* **4**, R73 (2003).
82. I. Schon, I. R. Arkhipova, *Gene* **371**, 296 (2006).
83. B. K. Thompson, S. M. Christensen, (2010).
84. V. Zupunski, F. Gubensek, D. Kordis, *Mol Biol Evol* **18**, 1849 (2001).
85. K. Usdin, P. Chevret, F. M. Catzeflis, R. Verona, A. V. Furano, *Mol Biol Evol* **12**, 73 (1995).

86. S. L. Gasior *et al.*, *Gene* **390**, 190 (2007).
87. M. C. Seleme *et al.*, *Proc Natl Acad Sci U S A* **103**, 6611 (2006).
88. S. L. Martin, J. Li, J. A. Weisz, *J Mol Biol* **304**, 11 (2000).
89. K. Januszyk *et al.*, *J Biol Chem* **282**, 24893 (2007).
90. V. O. Kolosha, S. L. Martin, *J Biol Chem* **278**, 8112 (2003).
91. S. L. Martin, *J Biomed Biotechnol* **2006**, 45621 (2006).
92. S. L. Martin, F. D. Bushman, *Mol Cell Biol* **21**, 467 (2001).
93. D. A. Kulpa, J. V. Moran, *Hum Mol Genet* **14**, 3237 (2005).
94. D. A. Kulpa, J. V. Moran, *Nat Struct Mol Biol* **13**, 655 (2006).
95. S. L. Martin, D. Branciforte, D. Keller, D. L. Bain, *Proc Natl Acad Sci U S A* **100**, 13815 (2003).
96. Q. Feng, J. V. Moran, H. H. J. Kazazian, J. D. Boeke, *Cell* **87**, 905 (1996).
97. H. Takahashi, H. Fujiwara, *EMBO J* **21**, 408 (2002).
98. B. Brouha *et al.*, *Proc Natl Acad Sci U S A* **100**, 5280 (2003).
99. N. G. Coufal *et al.*, *Nature* (2009).
100. B. A. Dombroski *et al.*, *Mol Cell Biol* **14**, 4485 (1994).
101. D. M. Sassaman *et al.*, *Nat Genet* **16**, 37 (1997).
102. G. J. Cost, Q. Feng, A. Jacquier, J. D. Boeke, *EMBO J* **21**, 5899 (2002).
103. S. R. Heras, M. C. Lopez, J. L. Garcia-Perez, S. L. Martin, M. C. Thomas, *Mol Cell Biol* **25**, 9209 (2005).
104. R. S. Alisch, J. L. Garcia-Perez, A. R. Muotri, F. H. Gage, J. V. Moran, *Genes Dev* **20**, 210 (2006).
105. P. W. Li, J. Li, S. L. Timmerman, L. A. Krushel, S. L. Martin, *Nucleic Acids Res* **34**, 853 (2006).
106. W. Wei *et al.*, *Mol Cell Biol* **21**, 1429 (2001).
107. J. N. Athanikar, T. A. Morrish, J. V. Moran, *Nat Genet* **32**, 562 (2002).
108. S. L. Gasior, T. P. Wakeman, B. Xu, P. L. Deininger, *J Mol Biol* **357**, 1383 (2006).
109. M. A. Batzer, P. L. Deininger, *Nat Rev Genet* **3**, 370 (2002).
110. K. Han *et al.*, *Genome Res* **15**, 655 (2005).
111. E. Ullu, C. Tschudi, *Nature* **312**, 171 (1984).
112. M. S. Comeaux, A. M. Roy-Engel, D. J. Hedges, P. L. Deininger, *Genome Res* **19**, 545 (2009).
113. P. L. Deininger, J. V. Moran, M. A. Batzer, H. H. J. Kazazian, *Curr Opin Genet Dev* **13**, 651 (2003).
114. T. H. Eickbush, in *R2 and Related Site-Specific Non-Long Terminal Repeat Retrotransposons*, Craig, NL, Craigie, R, Gellert, M, A. M. Lambowitz, Eds. (ASM Press, Washington, DC, 2002),
115. T. H. Eickbush, B. Robins, *EMBO J* **4**, 2281 (1985).
116. S. M. Christensen, A. Bibillo, T. H. Eickbush, *Nucleic Acids Res* **33**, 6461 (2005).
117. D. G. Eickbush, W. C. r. Lathe, M. P. Francino, T. H. Eickbush, *Genetics* **139**, 685 (1995).
118. D. G. Eickbush, T. H. Eickbush, *Genetics* **139**, 671 (1995).
119. K. K. Kojima, K. Kuma, H. Toh, H. Fujiwara, *Mol Biol Evol* **23**, 1984 (2006).
120. D. E. Stage, T. H. Eickbush, *Insect Mol Biol* **19 Suppl 1**, 37 (2010).
121. E. Kierzek *et al.*, *J Mol Biol* **390**, 428 (2009).
122. A. M. Ruschak *et al.*, *RNA* **10**, 978 (2004).
123. T. Anzai, M. Osanai, M. Hamada, H. Fujiwara, *Nucleic Acids Res* **33**, 1993 (2005).
124. M. Osanai, H. Takahashi, K. K. Kojima, M. Hamada, H. Fujiwara, *Mol Cell Biol* **24**, 7902 (2004).
125. S. M. Christensen, J. Ye, T. H. Eickbush, *Proc Natl Acad Sci U S A* **103**, 17602 (2006).
126. D. G. Eickbush, T. H. Eickbush, *Mol Cell Biol* **23**, 3825 (2003).
127. D. G. Eickbush, T. H. Eickbush, *Mol Cell Biol* (2010).

128. S. Christensen, T. H. Eickbush, *J Mol Biol* **336**, 1035 (2004).
129. D. D. Luan, T. H. Eickbush, *Mol Cell Biol* **16**, 4726 (1996).
130. D. D. Luan, T. H. Eickbush, *Mol Cell Biol* **15**, 3882 (1995).
131. E. Kierzek *et al.*, *Nucleic Acids Res* **36**, 1770 (2008).
132. D. H. Mathews, A. R. Banerjee, D. D. Luan, T. H. Eickbush, D. H. Turner, *RNA* **3**, 1 (1997).
133. A. Kurzynska-Kokorniak, V. K. Jamburuthugoda, A. Bibillo, T. H. Eickbush, *J Mol Biol* **374**, 322 (2007).
134. J. A. George, W. D. Burke, T. H. Eickbush, *Genetics* **142**, 853 (1996).
135. R. S. Brown, *Curr Opin Struct Biol* **15**, 94 (2005).
136. A. V. Giesecke, R. Fang, J. K. Joung, *Mol Syst Biol* **2**, 2006.2011 (2006).
137. S. S. Krishna, I. Majumdar, N. V. Grishin, *Nucleic Acids Res* **31**, 532 (2003).
138. S. A. Wolfe, L. Nekludova, C. O. Pabo, *Annu Rev Biophys Biomol Struct* **29**, 183 (2000).
139. M. E. Churchill, T. D. Tullius, A. Klug, *Proc Natl Acad Sci U S A* **87**, 5528 (1990).
140. J. M. Matthews, M. Sunde, *IUBMB Life* **54**, 351 (2002).
141. K. Ogata *et al.*, *Proc Natl Acad Sci U S A* **89**, 6428 (1992).
142. K. Ogata *et al.*, *Nucleic Acids Symp Ser* 201 (1993).
143. M. D. Abràmoff, Magalhães, P.J. and Ram, S.J, *Biophotonics international* **11**, 36 (2004).
144. T. D. Tullius, B. A. Dombroski, *Proc Natl Acad Sci U S A* **83**, 5469 (1986).
145. T. D. Tullius, B. A. Dombroski, M. E. Churchill, L. Kam, *Methods Enzymol* **155**, 537 (1987).
146. B. Brenowitz, D. F. Senear, R. E. Kingston, in *DNase I Footprint Analysis of Protein-DNA Binding*, F. M. Ausubel, Ed. (John Wiley and Sons, Inc, Hoboken, NJ, 2003), pp. 12.4-.5.
147. A. Gabriel *et al.*, *Mol Cell Biol* **10**, 615 (1990).
148. W. D. Burke, F. Muller, T. H. Eickbush, *Nucleic Acids Res* **23**, 4628 (1995).
149. H. Takahashi, S. Okazaki, H. Fujiwara, *Nucleic Acids Res* **25**, 1578 (1997).
150. H. S. Malik, T. H. Eickbush, *Mol Biol Evol* **15**, 1123 (1998).
151. W. D. Burke, H. S. Malik, J. P. Jones, T. H. Eickbush, *Mol Biol Evol* **16**, 502 (1999).
152. J. Yang, H. S. Malik, T. H. Eickbush, *Proc Natl Acad Sci U S A* **96**, 7847 (1999).
153. H. S. Malik, W. D. Burke, T. H. Eickbush, *Mol Biol Evol* **16**, 793 (1999).
154. E. Berezikov, A. Bucheton, I. Busseau, *Genome Biol* **1**, RESEARCH0012 (2000).
155. H. S. Malik, T. H. Eickbush, *Genetics* **154**, 193 (2000).
156. Y. Kubo, S. Okazaki, T. Anzai, H. Fujiwara, *Mol Biol Evol* **18**, 848 (2001).
157. V. Zupunski, F. Gubensek, D. Kordis, *Mol Biol Evol* **18**, 1849 (2001).
158. T. H. Eickbush, H. S. Malik, in *Origins and Evolution of Retrotransposons*, Craig, NL, Craigie, R, Gellert, M, A. M. Lambowitz, Eds. (ASM Press, Washington, DC, 2002), pp. 1111-1146.
159. W. D. Burke, H. S. Malik, S. M. Rich, T. H. Eickbush, *Mol Biol Evol* **19**, 619 (2002).
160. J. Permanyer, R. Gonzalez-Duarte, R. Albalat, *Genome Biol* **4**, R73 (2003).
161. M. Komatsu, K. Shimamoto, J. Kyozuka, *Plant Cell* **15**, 1934 (2003).
162. W. D. Burke, D. Singh, T. H. Eickbush, *Mol Biol Evol* **20**, 1260 (2003).
163. K. K. Kojima, H. Fujiwara, *Genome Res* **15**, 1106 (2005).
164. K. K. Kojima, T. Matsumoto, H. Fujiwara, *Mol Cell Biol* **25**, 7675 (2005).
165. I. Schon, I. R. Arkhipova, *Gene* **371**, 296 (2006).
166. T. Matsumoto, M. Hamada, M. Osanai, H. Fujiwara, *Mol Cell Biol* **26**, 5168 (2006).
167. O. Novikova *et al.*, *BMC Evol Biol* **7**, 93 (2007).
168. J. K. Biedler, Z. Tu, *BMC Evol Biol* **7**, 112 (2007).
169. T. H. Eickbush, V. K. Jamburuthugoda, *Virus Res* (2008)

BIOGRAPHICAL INFORMATION

Blaine Thompson was born and raised in Amarillo Texas where he graduated from Amarillo High School in 2002. He attended Amarillo Community College for 2 years before he transferred to the University of Texas at Arlington to complete his Bachelors of Science degree in Microbiology and minor in Chemistry. He continued his graduate studies at the University of Texas Arlington where he acquired a Masters of Science in Biology degree. Blaine plans to gain research experience as a research assistant in academic laboratories before advancing to a career in Industry.