# SINE MOBILIZATION ANALYSIS TO IDENTIFY STRUCTURAL CHARACTERISTIC THAT ASSISTS WITH CYTOPLASMIC MOBILIZATION TO THE RIBOSOME

Jordyn M. Hutchison

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

December 2020

Committee:

Julia Halo, Advisor

Raymond Larsen

Vipaporn Phuntumart

## ABSTRACT

Short Interspersed Nuclear Elements (SINEs) are non-autonomous mobile elements present in eukaryotic genomes. SINEs are mobilized in the form of a transcribed RNA intermediate in a 'copy and paste' process referred to as retrotransposition, by making use of the enzymes encoded by an autonomous element pair, or a Long Interspersed Nuclear Element (LINE). In turn, the active LINE is able to mobilize its own mRNA (termed retrotransposition *in cis*) or a SINE RNA (*in trans*). SINE insertions have been linked to not just genetic diversity but also naturally and artificially selected phenotypes in humans and non-human mammals. Within these host genomes, SINEs differ in their activities and evolutionary histories. Unlike a LINE, which contains an RNA II polymerase promoter, a SINE is known for its internal polymerase III promoter which has played a crucial role in its structure and gives it an advantage to utilizing the mobile proteins from LINE by attaching onto the ribosome via conserved structure involving signal recognition particles (SRP). The active and most predominant SINE in human genomes, *Alu*, is derived from 7SL RNA; the active SINE in canine genomes, SINE_Cf is derived from a $tRNA^{Lys}$, and active SINEs in murine genomes, B1 and B2, are derived from either 7SL and tRNA, respectively, in an evolutionary independent manner. Members of the currently active *Alu*/L1 pair, of the specific *Alu*Y (*i.e.*, for 'young') and L1Hs (*i.e.*, for human specific) have been very well studied and their mobilization detailed in an elegant model. *Alu* RNA adopts the conserved structure and cellular binding partners as it's ancestral 7SL RNA molecule, SRP9/14, which direct the *Alu* intermediate to the ribosome. In the currently understood *Alu*/L1Hs model, while docked on the ribosome, the *Alu* RNA intermediate 'piggybacks' the translated L1Hs ORF2p at the ribosome, which functions to transport the intermediate to the nucleus and reverse transcribe it into a new locus as double-stranded DNA. *Alu* retrotransposition has be shown in a

cellular *ex vivo* assay using the SINE element marked with a reporter gene, *neo*$^{R}$, in which mobilization is assessed by the relative presence of *neo*$^{R}$ cellular foci present following L1Hs-driven retrotransposition *in trans*. We hypothesize by exploring relative activities of non-*Alu* SINE elements, we will be able to identify structural characteristics and potential sites of host interaction involved in mobilization and ribosome association. Using the described *ex vivo* assay, we have cloned tRNA SINE elements and preliminarily assessed their mobilization. This work should shed light on the evolution and patterns of ribosomal access of non-7SL SINEs in mammalian genomes.

# ACKNOWLEDGMENTS

I would first like to acknowledge my advisor, Dr. Julia Halo, who introduced me to the wonderful world of SINEs and has shown me what happens when you chase your passions. I am so grateful for her patience and support while completing this project. My committee members, Dr. Ray Larsen and Dr. Vipa Phuntumart have been incredible guides through my time at Bowling Green State University by being attentive and always having the door open for me to stop by to discuss my research and anything else I got carried away with.

In the Halo lab, I was so unbelievably blessed to have two amazing women to work with me: Abigail Jarosz and Madison Baltzly. Abbey was always there for any questions I had and she undoubtedly made my two years a thousand times better. I have also been so lucky to have Maddie working on this project with me and I am so glad she's going to be continuing this project during her time as a Master's student at BGSU. These women are some of the most intelligent, loving, and wonderful people I have ever known. I am so fortunate to not only call them my lab-mates but also my great friends.

Finally, thank you to my incredible parents and siblings, who are the most supportive and loving people. Mom, thank you for talking to me every day, giving a comforting shoulder, and giving me a good laugh whenever I needed it. Dad, thank you for spending many nights talking about this grey world with me and showing me unconditional fatherly love. Matt and Riley, I am so thankful you two are my brothers and I will always have your back. I hate to be cheesy but, I have met my life-long friends during my 6 years at BGSU and without them I can honestly say I would not be where I am today. Chris, I am so glad I met you that day in chemistry lab and I cannot wait to see what a wonderful life of love and happiness we have ahead of us. Rebel, thank you for rescuing me. To summarize my time at BGSU: The Office- Season 9, Episode 23, 45:07.

TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION

## Transposons

Transposable elements, or transposons, comprise of nearly half the human genome. Though previously often referred to as "junk DNA", many of these elements are now known to have a great role in genetic diversity, physiology, and evolution (Kurth, 2006, Hancks, 2012, Lisch and Burns, 2018, Platt et al., 2018, Hancks and Kazazian, 2012). Transposons can be further categorized into one of two classes (I and II), defined as being able to mobilize to a new genomic location via, respectively, a DNA or RNA intermediate (Figure 1). Class II transposons, or DNA transposons (Figure 1A), comprise about 3% of the human genome. These transposons utilize a "cut and paste" mechanism to excise DNA and move it to another location in the genome, achieved through an encoded transposase enzyme within the element (Kramerov and Vassetzky, 2011, Platt et al., 2018). These elements are found in both eukaryotes and prokaryotes however, in mammals there are currently no known active DNA transposons except for one species of bats, *Myotis lucifugus* (Campos-Sanchez et al., 2014, Mitra et al., 2013).

Class I transposons (Figure 1B) are mobilized via an RNA intermediate, and therefore are referred to as 'retrotransposons'. The mobilization of the RNA intermediate occurs through a "copy and paste" mechanism, during which the source element remains in place and new insertions are generated at a different chromosomal location. The mechanism of this mode of replication is referred to as reverse transcription and produces new insertions that are flanked by short sequence repeats termed target site duplications (TSDs). Class I elements are present in eukaryotic genomes only, thus they're predominant in mammalian and plant species. They comprise 43% of the human genome and have influenced the evolution of eukaryotes, for example by expanding the genome, leading to new mutations, and altering gene expression (Wessler, 2006). Retrotransposons are

further classified into one of two types: long terminal repeat (LTR) and non-LTR. The LTR retrotransposons are so named for the presence of long sequence repeats (in the hundreds of bases) located at the 5' and 3' ends of the element (Figure 1B). The vast majority of LTR retrotransposons are elements referred to as endogenous retroviruses (ERVs), termed for their origination from past germline retroviral infections. Although they tend to be degraded as part of the host lineage genome after integration, ERVs are mostly classified as autonomous elements, meaning they encode all necessary functions to replicate on their own from within a host cell, making use of some host-encoded machinery. They mobilize by infecting a host cell, reverse transcribing their mRNA into DNA. After their DNA has been reverse transcribed, ERVs use integrase to insert themselves into the host (Bannert and Kurth, 2006, Katzourakis et al., 2005, Boeke and Stoye, 1997, Jern and Coffin, 2008).

Non-LTR retrotransposons include the short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) which are retrotransposons whose expressed transcript is reverse transcribed into double-stranded DNA (dsDNA) via the single-stranded RNA (ssRNA) intermediate in a process termed retrotransposition (Carnell, 2003). More specifically, such elements are transcribed from existing copies within the host genome, and the RNA transcripts are then reverse transcribed to dsDNA during insertion into a new genomic location, the full mobilization cycle being termed retrotransposition. The full retrotransposition mechanism produces new insertions that possess a 3' poly(A) tract and are flanked by short TSD sequence repeats. Together these hallmark features make them identifiable within the genome (Kramerov and Vassetzky, 2011). SINEs are short, discrete elements of an average is ~300bp that do not encode any protein functions. As such, they are referred to as "non-autonomous", meaning that they do not have the ability to retrotranspose themselves, and instead are mobilized by exploiting

enzyme functions provided from LINE-encoded proteins, and so being referred to as retrotransposition *in trans* (Beck et al., 2011, Kramerov and Vassetzky, 2011). LINEs are autonomous, encoding the necessary proteins to drive retrotransposition of its own mRNA (retrotransposition *in cis*) or a SINE RNA (in trans). A full length LINE is roughly 5-9 kb in length and has two open reading frames, ORF1 and ORF2, whose proteins ORF1p and ORF2p are responsible for mobilization in the genome (Bannert and Kurth, 2006, Kramerov and Vassetzky, 2011).

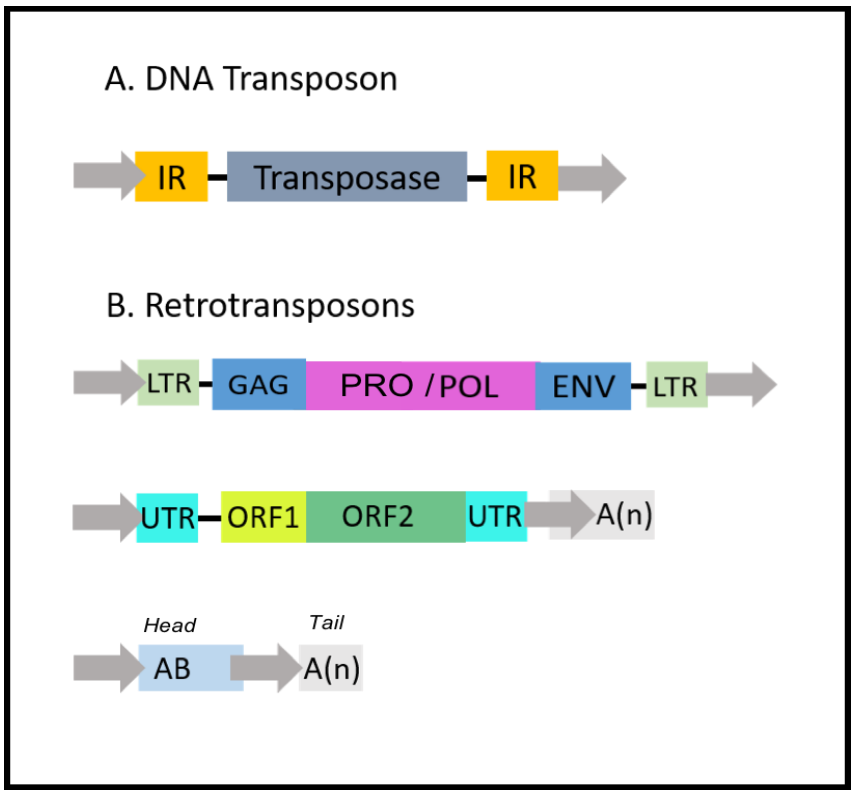


Figure 1: Structures of class I and class II transposons

A) Class II element: DNA transposon containing the transposase gene in between two inverted repeat sequences and like all elements shown, are followed by two target site duplication regions (TSD). B) Class I elements: LTR retrotransposons such as endogenous retroviruses are flanked with two long terminal repeats (LTR) and they contain four coding regions for the enzymatic activities required for their amplifications. Non-LTR retrotransposons such as LINEs and SINEs are depicted. LINEs have untranslated regions (UTR), their two open reading frames (ORF1 and ORF2) responsible for mobilization activities, and their poly(A) tail at their 3' end. SINE structures have a head, body, and tail. The head is where the AB regions that together make up the PolIII promoter, no coding regions in the body and finally, a 3' poly(A) tail.

SINE and LINE Structure

SINEs are typically derived from, but not completely limited to, one of two Pol III transcribed host genes that encode short non-protein coding RNA molecules: 7SL or tRNA. Owing to their derived source, the SINE structurally contains an internal RNA polymerase III promoter region as well as an adenine-rich region referred to as the poly(A) tail (Figure 2A) (Karijokich, 2017). In turn, LINEs contain an internal RNA polymerase II promoter, poly(A) signal, and two open reading frames (ORF1 and ORF2), that encode proteins (ORF1p and ORF2p, respectively).

ORF1p and ORF2p encode with the functions that drive retrotransposition of their own source RNA *in cis* or the RNA of their SINE partner *in trans* (Figure 2B). ORF1p contains an RNA-binding function whereas ORF2p contains reverse transcriptase and endonuclease activities. These proteins are able to recognize the LINE's 3' end *in cis,* which contains the poly(A)-tail. Once the ORF2p RNP has bound to the 3' of the LINE, it is then able to undergo mobilization to the nucleus to undergo target-primed reverse transcription (TPRT), which is necessary for insertion into a new genomic location (detailed further below) (Hancks and Kazazian, 2012).

SINE RNA does not encode functions to mobilize itself and has instead evolved to parasitize proteins encoded from its analogous LINE partner for retrotransposition. It is generally accepted that the SINE accomplishes this by diverting LINE functions at the ribosome, where the SINE transcript is specifically able to compete with the ORF2p protein being translated from LINE messenger RNA *in trans*, thereby utilizing ORF2p functions that mediate translocation of the SINE RNA to the nucleus and its reverse transcription into host DNA. Specifically, in the case of *Alu*, due to its conserved 7SL structure, host proteins from signal recognition particle (SRP9/14) bind the *Alu* RNA intermediate and target it to the ribosome. 7SL derived SINEs are relatively exclusive to mammalian species and identified as either dimeric like the human *Alu* element, or monomeric

as in rodent B1 (Figure 3). Mobilization of 7SL/*Alu* and tRNA SINEs specifically involve a functional LINE ORF2p. Although both ORF1p and ORF2p are necessary for L1 retrotransposition in cis, ORF2p alone is sufficient for the SINE to successfully undergo mobilization *in trans* into different locations within the genome. Specifically, the ORF2p functions responsible for retrotransposing SINE RNA by target-primed reverse transcription are i. endonuclease (nicks the target DNA) and ii. reverse transcription (copies dsDNA from ssRNA template) (see detailed below).

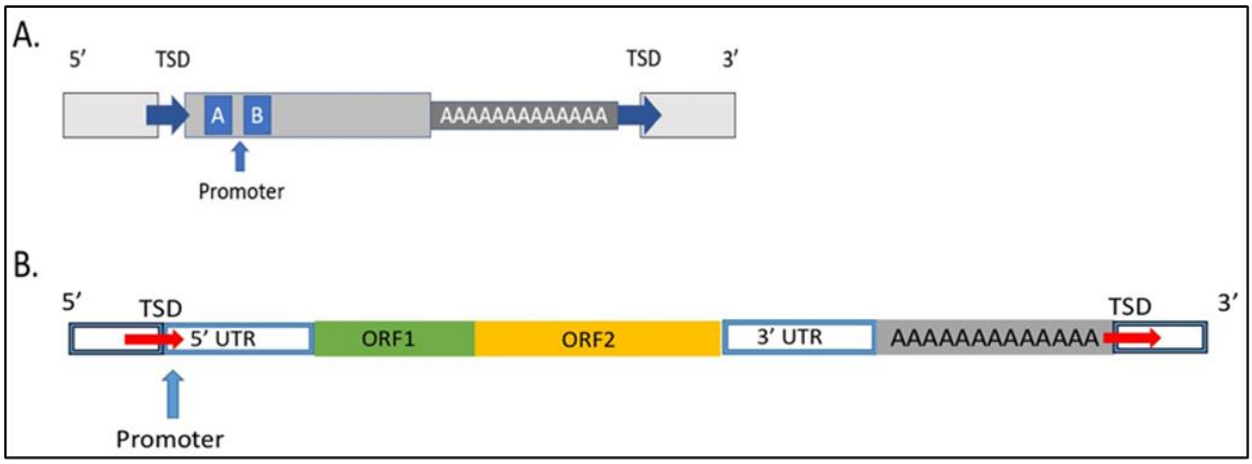


Figure 2: Representation of SINE and LINE structure

A. SINE structure: 5' and 3' regions shown by light gray boxes. The TSD depicted by the blue arrows. Located within the blue A/B boxes is the promoter region. The poly(A) tail shown by dark gray box between the body and 3' TSD region. Abbreviation of TSD: Target Site Duplication.
B. LINE structure: 5' and 3' region flank ends of the structure in blue bordered boxes. TSD indicated by red arrows. The 5' and 3' UTR are flanking ORF1 and ORF2 with blue bordered boxes. Poly(A) tail depicted as dark gray box in between the 3'UTR and 3' TSD. Abbreviations are as follows: TSD, target site duplication, UTR, untranslated region, ORF1, Open reading frame 1, ORF2, open reading frame 2.

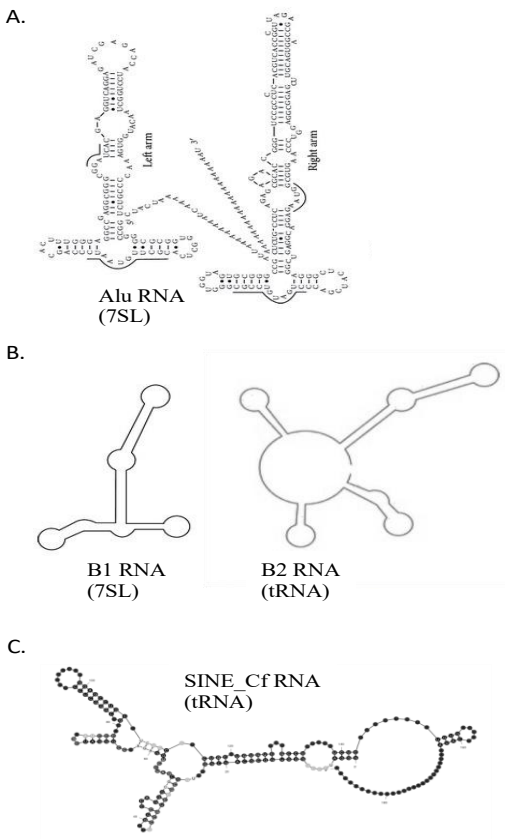


Figure 3: 7SL and tRNA derived SINEs

A. The 7SL-derived dimeric *Alu* element.

B. B1 is 7SL monomeric SINE and B2 is tRNA derived.

C. The tRNA derived SINE_Cf element.

Image modified from (Dridi, 2012).

Retrotransposition

The full process of retrotransposition is shown in Figure 4 and detailed as follows. A full length, 'retrotransposition-competent' LINE is transcribed via its internal RNA polymerase II promoter, and the mRNA intermediate then travels to the cytoplasm to be translated on the ribosome. The translated proteins from the LINE, ORF1p and ORF2p, can associate with the LINE mRNA being translated on the ribosome (termed *cis* interaction) or with a SINE RNA (*trans* interaction) to retrotranspose the transcript a new genomic location. Because the RNA Pol III mRNA 7SL SINE is non-autonomous and has no open reading frames, the encoded SRP portion of the element assists in hybridizing of the SINE RNA to the ribosome, such that it may utilize the newly-translated ORF2p from the LINE, where it (specifically, the 7SL SINE) has been shown to have role in pausing during translation of proteins with signal peptides because of functions as the SRP of the ancestor molecule (Collart, 2020, Ahl et al., 2015). Although the cytoplasm interaction is not fully understood for tRNA SINE mRNA, they are able to utilize the enzymatic machinery of a retrotransposition competent ORF2p with great efficiency.

Target-primed reverse transcription is the term for the mechanism where the endonuclease activity and the reverse transcriptase activity of ORF2 LINE protein act. Endonuclease activity directs a cut in the DNA backbone at a thymine-rich region, the consensus of which is 3'AA/TTTT, where '/' represents the endonuclease target site (Dunker et al., 2017) (Figure 3). The 3' PolyA region of the template RNA can then hybridize with the exposed T-rich track on the target site, where the 3'-OH of the last A serves as the primer for reverse transcriptase. Thus, reverse transcription is target primed. Once the SINE is copied into the DNA, it is reverse complemented on the opposite strand by host DNA polymerase, also generating the flanking TSDs, and the ends sealed by host ligase (Taylor, 2013).

SINEs and LINEs distribute differently throughout the genome and although speculated to be completely random, data has shown otherwise (Weiner, 2002, Elbarbary et al., 2016). It is understood that 30% of the human genome is comprised of these elements and through thorough analysis, it's been shown that SINEs tend to be inserted into more GC-rich regions whereas LINEs have been most commonly found in AT-rich regions (Weiner, 2002, Elbarbary et al., 2016) Although these findings do not prove an insertion preference for either element, it does show the greater potential for SINEs to interfere with gene-rich regions and can lead to gene alterations and mutations in both somatic and germ cell lines (Elbarbary et al., 2016).

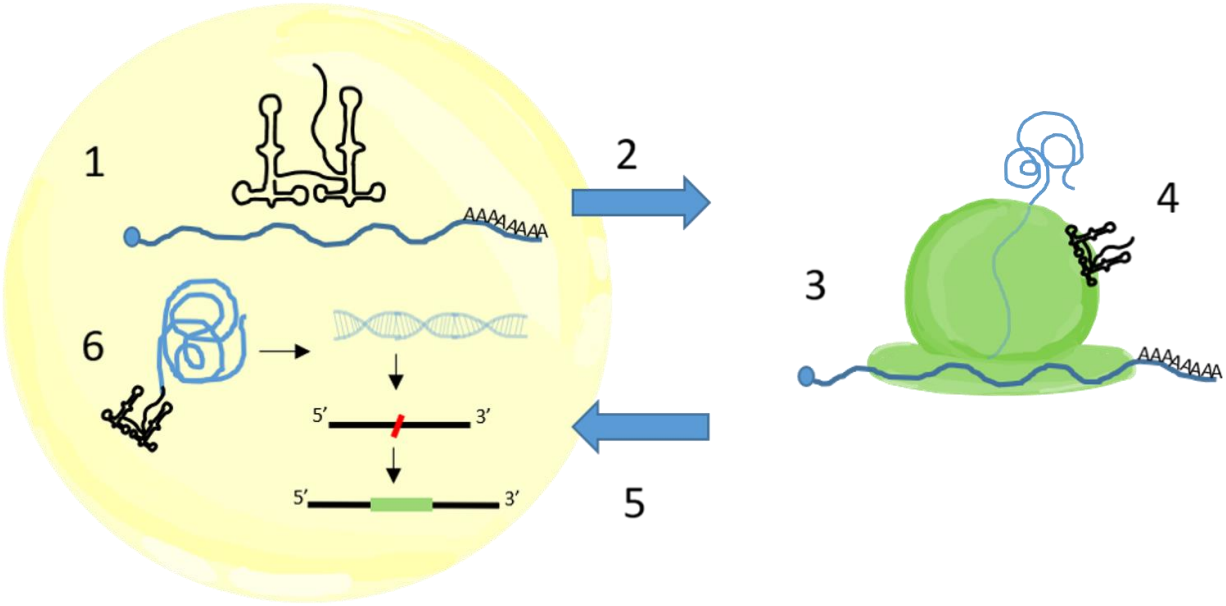


Figure 4: Retrotransposition of SINE and LINE elements

Process detailed above. Steps are defined as followed: 1, SINE and LINE translated in nucleus, 2, cytoplasmic import of SINE and LINE, 3, LINE translated on the ribosome, 4, ORF1p and ORF2p polypeptide chain are able to attach to either SINE or LINE, 5, element returns with ORF1p and ORF2p, 6, target-primed reverse transcription occurs.

## Comparison of Mammalian SINE Presence and Activity

The two main types of SINE are 7SL and tRNA derived, which are both evolutionarily distinct from the other. The human SINE, *Alu*, is 7SL derived and therefore has been studied extensively for understanding of genomic diversity, contributions to disease, *etc*. It is known how the 7SL derived SINEs are able to intercept the mobility machinery of the LINE by ribosomal binding via their structural components SRP9/14, hence their high mobility rates. SINEs derived from tRNA do not share a conserved secondary structure with 7SL RNA as they are not evolutionarily related (Figure 3). In contrast to the 7SL derived SINE, there is a minimal understanding of amplification steps within the cytoplasm thus far for tRNA-derived SINEs, for example structure function relationships, host-interacting proteins, or mechanisms of ribosome access. For this reason, it is still unclear how the tRNA SINE RNA is able to undergo movement from the nucleus through the cytoplasm, associate with a translated L1 ORF2p, and undergo association/recognition of the ribosome in the cytoplasm.

### *Human: Alu*

7SL RNA-derived SINE elements, such as the primate *Alu* embody over 10% of the human genome and *de novo* insertions have been estimated to occur in roughly 1:20 live births (Hancks and Kazazian, 2012, Hancks and Kazazian, 2016). 7SL derived SINEs are relatively exclusive to mammalian species and identified as either dimeric like the human *Alu* element, or monomeric as in rodent B1 (Figure 3). The human genome also contains 1.1 million *Alu* insertions and contribute greatly to genomic diversity, as 75% of genes are known to contain *Alu* insertions (Wang, 2005, Cordaux, 2009). Insertions may occur in either somatic or germline tissues and if inserted near or within a gene may affect the gene's expression. Indeed, a handful of *Alu* insertions have been shown as the causative agent of human diseases resulting from, but

not limited to, premature termination during transcription, alternative splicing of transcripts, and deletions via element-element recombination (Hancks and Kazazian, 2012, Hancks and Kazazian, 2016). In the human genome, *Alu* insertions can be classified as one of three subfamilies; *Alu*Y, *Alu*S, and *Alu*J (the latter are named for their foremost researchers and discoverers of the two lineages: Smith and Jurka). The youngest of these families is *Alu*Y (for 'young'), of which two derivatives, *Alu* Ya5 and *Alu* Yb8, contribute to newest (*i.e.*, *de novo*) *Alu* insertions (Batzer, 2002).

As mentioned, 7SL SINEs such as *Alu* encode a signal recognition particle (SRP), which assists in associating to ribonuclear proteins in the cytoplasm. *Alu*, thus has to utilize the conserved RNA/SRP interactions, specifically SRP9/14 (SRP binding regions are lined in Figure 3 on the RNA structure) as the means to access the ribosome in the host cell cytoplasm, where the L1 ORF2p then recognizes the element for import into the nucleus and subsequent TPRT (also refer to Figure 4). The SRP structure of the *Alu* is mirrored in the 7SL RNA SRP which is a ribonuleoprotein particle (RNP); this interacts with the ribosome and has a role in translation and excretion of proteins (Ahl et al., 2015). The *Alu* domain of the SRP is able to attach to the ribosome between the two subunits where it will also interact with the GTPase center, also referred to as the ribosomal translation elongation factor-binding site, where it can assist with stalling the translation process (Ahl et al., 2015). The SRP is then able to wrap around to expose the S domain to the exit site of the ribosome to access newly-synthesized proteins. In order for the SRP to utilize the protein chains, each amino acid contains a hydrophobic signal sequence which is a marker so that the SRP can scan and search for the membrane proteins (Voorhees, 2015).

This understanding of the SRP RNA is how it is possible that elements such as *Alu* (and other 7SL derived SINEs like B1) are able to hijack the proteins from their counterpart element, the LINE. It has been found that mutations in the SRP binding site have been shown to decrease or even completely cease the mobilization capability of the element, thus supporting the role of SRP binding of the *Alu* RNA for access to the ribosome (Figure 5) (Ahl et al., 2015). Similar findings have been shown for the rodent B1 7SL-derived element (Figure 3B) (Dewannieux M, 2005).

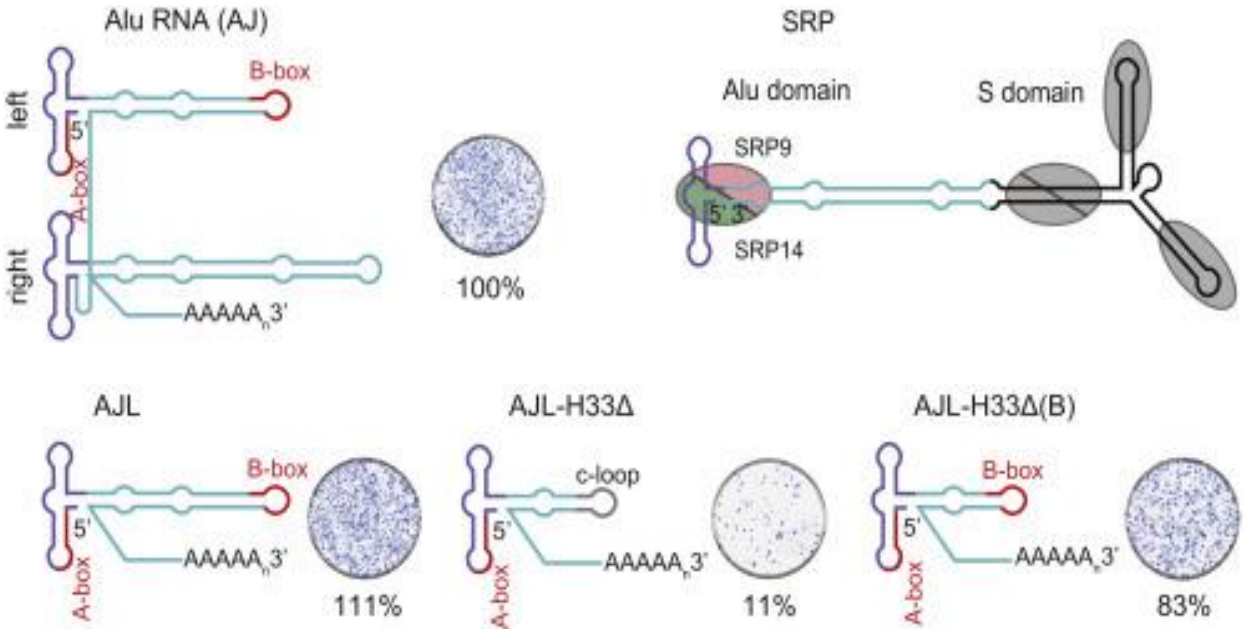


Figure 5: *Alu* SRP amplicons with retrotransposition assay

*Alu* RNA structure (AJ) begins with the two notable monomers, right and left. The left-hand side contains the A and B boxes which contain the internal POL III promoter for transcription. The right-hand side contains the encoded 3' oligo(A) sequence which assists in its mobilization throughout the genome in retrotransposition. The endogenous SRP RNA (shown under SRP) contains two domains, *Alu* and S. The S domain is for location of signal sequence to bind to SRP receptor site of the ribosome. The *Alu* domain is the binding site for the ribosome. AJL is showing an *Alu* with a monomer structure and an A-tract. Its performance with a retrotransposition assay, like AJL-H33Δ(B) whose structure is similar, performed well under the assay. On the other hand, AJL-H33Δ is structurally similar to the others but is missing the B-box portion of the element which has led to poor amplification in the retrotransposition assay. (Source: Ahl, V. et al. 2015)

*Mouse: B1 and B2*

Within the mouse genome, over 30% consist of class I non-LTR retrotransposons: murine L1 (L1-Md), B2, and B1. B2 SINE elements are a prime example of tRNA derived SINEs (Figure 4B, right). The majority of tRNA-related SINE RNA secondary structures differ from the ancestral tRNA structure and sequence. SINEs that are related to tRNA are found abundantly in most vertebrates although focally studied in mammalian species. B2 is derived from $tRNA^{Ser}$ and has been found to preform similarly to that of the human SINE *Alu.* Although it is unclear how the tRNA mobilize within the cytoplasm, which suggests that there may be a structural characteristic that contributes to ribosomal binding. Similar to the human SINE element *Alu*, the monomeric element B1 is 7SL derived and contains SRP9/14 binding sites. However, they do not amplify as well as *Alu,* due to a base change in their SRP9/14 site which has been shown to decrease their potential amplification rate 5-20 fold. Because B1 and B2 are ancestrally derived from 7SL and tRNA, respectively, they make great subjects to illustrate key characteristics of two varying SINE structures within the same genome (Dewannieux M, 2005).

*Canine: SINE_Cf*

SINE_Cf from the canine genome is $tRNA^{Lys}$ derived and (Figure 3C). SINE_Cf is ~170bp in length and roughly 31% of the canine genome is made up of retrotransposons (Eo, 2013). In contrast to *Alu* and 7SL derived SINEs, there is a minimal understanding of amplification steps within the cytoplasm thus far for tRNA-derived SINEs, for example structure function relationships, host-interacting proteins, or mechanisms of ribosome access. Although these processes have been deeply explored in the human *Alu*/L1 element pair, tRNA SINEs have been given far less attention, and it is still unclear how the tRNA SINE RNA is able to undergo

movement from the nucleus through the cytoplasm, associate with a translated L1 ORF2p, and undergo association/recognition of the ribosome in the cytoplasm.

To demonstrate links to phenotypes and orthologs of SINE_Cf, utilizing ~100 whole genome Illumina data and anchored mapping of non-reference domestic and wild breed canines my advisor, Dr. Halo, was able to locate 48 insertions not annotated within the reference genome. With these same insertions, Dr. Halo was able to successfully show mobilization of SINE_Cf with human-L1 in an *ex vivo* tissue culture assay, which from this data it can be inferred that like *Alu,* SINE_Cf may have conserved regions within its structure that contribute to ribosomal binding for access to enzymatic machinery of LINE.

We hypothesize that mutation to bases involved in retrotransposition of the SINE will be identified by correlation to decreased rates of mobilization in *ex vivo* assay. Using a standard ex vivo assay to clone various SINE loci to show efficient amplification, while focusing on the mouse genome, as it harbors both 7SL and tRNA SINE families that are currently mobilized. We anticipate that through selecting multiple insertions that vary in nucleotide sequence diversity and corresponding to the SINEC_Cf, B1 and B2 element families for a comparative mobilization analysis we will identify structural characteristics that have role(s) with mobilization to the ribosome. By comparing the inferred bases and/or regions to the element secondary structure, we should be able to hone in on regions we anticipate to be involved with host interacting functions. Moreover, if such structural findings are to be found, further understanding can be made of TPRT and the evolution and mechanisms of tRNA-derived SINEs.

# CHAPTER 2: MATERIALS AND METHODS

## Genetic Materials

Genetic materials in this project are 3T3 murine cells and HeLa cells from ATCC. Cells from ATCC were kept under 5% carbon dioxide (CO2) and 37 ºC culturing conditions. For complete growth medium for 3T3, 10% bovine calf serum and 2% antibiotic PenStrep in Dulbecco's Modified Eagle's Medium (DMEM). While subculturing, cells at roughly 75-80% confluency to 3 to 5 X $10^3$ cells/$cm^2$ in a 75 $cm^2$ flask in a sterile and air-controlled ventilation hood. Using NucleoSpin gDNA extraction kit (Machery-Nagel), dsDNA from 3T3 cells was utilized for PCR for B1 and B2 insertions. To test the cDNA, GAPDH primers were used to ensure the efficacy of PCR reactions for various B1 and B2 loci. Subculturing HeLa cells requires using complete media that contains 10% fetal bovine calf serum (FBS), 1% Glutamax and 1% PenStrep in Dulbecco's Modified Eagle Medium (DMEM). Subculturing cells at roughly 75-80% confluency in a 175 $cm^2$ flask in a sterile and air-controlled ventilation hood. In this research, HeLa cells were used for the retrotransposition assay. Using the Countess II by Life Technologies, the exact cell concentration of $4\ x\ 10^5$ was achieved in each 6-well plate.

## Primer Design

Designed primers were formulated from whole genome sequences (WGS) that resembled the consensus sequence of the desired element from either CanFam3.1 or GRCm38 within a precise location in the genome. Such primers were designed to flank the predicted 5' TSD region and the 3' TSD regions of the element TE element. To ensure correct chromosomal location of SINE insertions, external primers were designed roughly 50-100bp upstream and downstream of the predicted SINE sequence. Internal primers flanked the 5' and 3' region of the SINE, excluding the poly(A) tail, with the addition of restriction sites and a poly(A) signal site.

Restrictions sites for *Not*I and *Age*I were chosen as they are optimized in the expression vector PCEP4 such that the SINE and the vector will adhere when ligated.

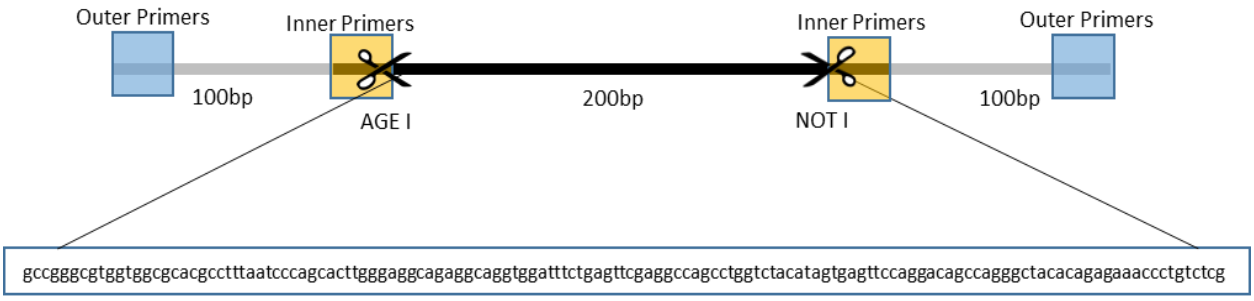


Figure 6: Primer design of SINE amplicons

Details described above. Blue boxes flanking each end of the sequence indicate the outer or external primers. The yellow boxes flanking the 5' and 3' region of the sequence are the inner primers with the restriction sequences for restriction digest and ligation for cloning. The sequence depicted is an example of a B1 SINE from the GRCm38 genome.

DNA Cloning

Amplifying SINE-specific products was completed in a nested PCR strategy, digested those products, and cloned them using a DH5alpha (chemically competent *E. coli*) strategy into an expression vector that was prepared by strategic digestion, gel extractions, and phosphatase treatment. To target the murine B1 and B2 elements, a similar strategy based on the mouse reference genome and 30 insertions per element family (tRNA-derived vs. 7SL-derived). Utilizing NIH 3T3 cells in culture for genomic extraction to assist in the cloning of individual insertions. Roughly 200 SINE_Cf primer sets were designed and amplified utilizing Canine DNA provided by PacBio genome of a breed dog (Zoey; GenBank GCA_005444595.1), following the same protocol as B1 and B2. Invitrogen *Taq* based PCR with 5uL of 10x *Taq* buffer, 4uL of 2mM dNTPs, 2uL of each primer, 0.4uL of 25 mM MgCl2, 0.6uL *Taq* enzyme and water per 50uL reaction. Then, utilizing an Eppendorf Mastercycler Thermocycler, each reaction underwent: 95ºC for 2 minutes, 35 cycles of 95 ºC for 30 seconds, annealing at a range of 58-60 ºC for 30 seconds, then extension at 72 ºC for 60 seconds, lastly final extension 72 ºC for 3 minutes. After successfully completing PCR, 10uL of the reaction was used 1.0% agarose in 1 X TBE for gel electrophoresis to assure the insertion was the appropriate length of 300-500bp, depending on the SINE element. Using Nucleospin Gel and PCR (Machery Nagel) according to the manufacturer's protocols, PCR products were cleaned and nano dropped using a Nanodrop Lite (Thermo Fisher) to confirm concentration in ng/uL.

Using the taq PCR amplicons, a 50uL digest reaction was created using 1uL *Age*I enzyme, 1uL *Not*I enzyme, 5uL Buffer, 30uL of cleaned PCR product and water. Using the same thermocycler, each reaction underwent: 37ºC for 2 hours followed by 65 ºC for 20 minutes. The digested products were then cleaned and read using electrophoresis and stored at -20ºC.

Expression vector (from PCEP4) and digested PCR products are then ligated together and grown in a chemically competent *E.Coli*, DH5a. Once cloned, the transformants were grown overnight on ampLB plates. After a 24 hour period, the colonies grown on the ampLB plates contained the expression vector and SINE clones. These colonies were harvested and grown in ampLB for a 24 hour period to then be purified using MINIplasmid Kit (Machery Nagel). These purified plasmids were then used for transfection into a specific concentration of HeLa cells to undergo *in vitro* assay.

Retrotransposition Assay

In the assay, we seeded $4\ x\ 10^5$ HeLa cells into each well of a 6 well plate and then transfected the cells in a transient transfection over a 24-hour period. Next, we then co-transfected a mixture of plasmids of the SINE clone marked with $neo^R$ and the appropriate L1 using the FuGENE transfection agent in Opti-Mem to each well. Following a 24-hour incubation, the transfected cells were grown in complete media supplemented with Genticin to allow for the selection of $neo^R$ cells. Over a period of 12-14 days, $neo^R$ cells expanded in small cell clusters, or foci, while the remaining cells that did not contain the report cassette marker died off. Therefore, the detection of retrotransposition is reliant upon the $neo$ marker on the SINE; this marker is situated in opposite orientation and contains a self-splicing intron with its own promoter and poly(A) signal. Expression of neomycin occurs when the L1-reporter is transcribed, spliced and reverse-transcribed and inserted into the DNA, thus confirming all steps of retrotransposition have successfully occurred. Mobilization was assessed at the end of the experiment by fixing in paraformaldehyde and staining the cell foci with crystal violet, after which they were quantified by foci counting (Figure 7).

## Bioinformatics Tools

Utilizing the consensus provided to utilize WGS to locate TEs similarly matching the elements studied with varying divergences using a BLAST-like tool (https://genome.ucsc.edu/index.html) to both the mouse (GRCm38.p6) and dog (CanFam3.1) reference genomes. The more variable sequences tested, the more likely it is that we may narrow which locations assist in retrotranspositional machinery. Based on an expanded set of ~200 SINEC_Cf insertions in the dog, strategic primers were designed for amplification and cloning in order to assess their mobilization in tissue culture. After amplification of each primer set design using the listed protocol, the cleaned and sequenced product were aligned in BioEdit Sequence Alignment Editor.

## Gentoyping SINE_Cf

For each reference loci, primers were designed to flank the predicted 5' and 3' regions based on the CanFam3.1 reference genome. Amplification of these primers were designed to be ~20bp in length and have a 59ºC annealing temperature. Each primer set was run with a breed specific DNA sample, collected previously. Cell lines were used as the template in an Invitrogen Taq PCR reaction with 10x buffer, 2.5 uM dNTPSs, 10uM each primer, 2.5uM MgCl2, and *Taq*. Each reaction was performed in an Eppendorf Mastercycler with the following conditions: denaturation at 95 ºC for 2 minutes, then 35 cycles of 95 ºC for 30 seconds, annealing at 59ºC for 30 seconds, and extension of 1:30 at 72 ºC. Finally, extension 72ºC for 3 minutes and holding at 10 ºC. 10uL of this reaction was then used for observation in gel electrophoresis in 1% agarose 1 x TBE. These genotyping PCR reactions are to assist in our visualization of sequence diversity within SINE_cf elements across various canine genomes and to further understand our sequence diversity across various SINE_cf insertions.

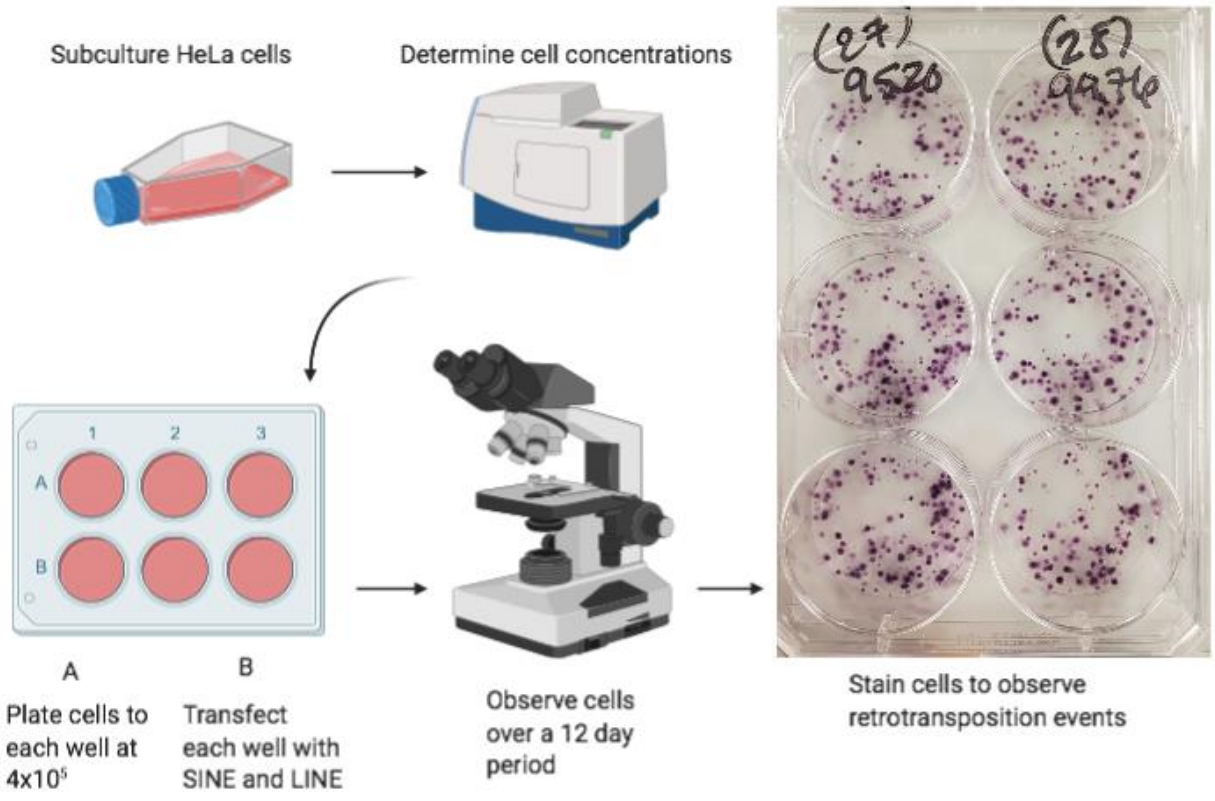


Figure 7: Retrotransposition assay

For successful retrotransposition assay, followed protocol for $4\ x\ 10^5$ subcultured confluent HeLa cells in a 6-well plate for transfection of SINE_Cf and reporter cassette. This two-week process utilized genticin to indicate retrotransposition events which can be visualized by the colonies that are stained using crystal violet.

Table 1: Genotyping whole genome amplification breed dog DNA samples

| Name | Breed | Sample Number |
|---|---|---|
| 5546 | Labrador Retriever | 1 |
| 5793 | Labrador Retriever | 2 |
| 4036 | Poodle | 3 |
| 5548 | Golden Retriever | 4 |
| 1308 | English Springer Spaniel | 5 |
| 4837 | Chinese Shar-Pei | 6 |
| 455 | Boxer | 7 |
| 2397 | Golden Retriever | 8 |
| 4826 | German Shepherd | 9 |
| 8260 | English Springer Spaniel | 10 |
| 1297 | Labrador Retriever | 11 |
| 4296 | Labrador Retriever | 12 |
| 704 | Samoyed | 13 |
| 8420 | Mix | 14 |
| 5921 | Maltese | 15 |
| 8851 | Golden Retriever | 16 |

DNA samples, previously collected, used for genotyping SINE_Cf reference and non-reference insertions. DNA samples were whole genome amplified and tested using GAPDH for PCR efficiency.

# CHAPTER 3: RESULTS

We hypothesize that mutations to bases involved in retrotransposition of the SINE can be identified by correlation to decreased rates of mobilization in our *ex vivo* assay. We will test this hypothesis by identifying and confirming polymorphic SINE_Cf insertions within the canine genome. After confirmation, cloning each SINE_Cf to use a standard retrotransposition assay to show various SINE loci amplification, while focusing on the canine genome. By selecting multiple insertions that vary in nucleotide sequence diversity corresponding to the $tRNA^{Lys}$ derived SINEC_Cf element, a comparative mobilization analysis was conducted to assist in the identification of structural characteristics that have potential role(s) with mobilization to the ribosome. These steps are further detailed through the following goals of my project:
Identification and selection of non-reference and reference insertions

Goal 1. Identify Non-reference (not present in the CanFam3.1 reference genome, however discovered in other dog genomes) and Reference (present in the CanFam3.1 reference genome) insertions. A total of 200 non-reference insertions were identified using the reference SINEC_Cf sequence to BLAT using the Genome Browser, against the most recently updated genome build of the domestic dog, CanFam3.1). Collectively, each SINE is roughly 170-210 bp in length with varying divergences from the consensus sequence (Figure 8). These sequences were used to design primers to confirm SINE insertions and polymorphism across various locations within the canine genome (Figure 6). Confirming polymorphism will assist with our assessment of mobilization of tRNA SINE insertions in later steps.

Goal 2. Confirm polymorphic presence of identified non-reference and reference SINEs. To determine variance between SINE_Cf insertions, we gathered reference SINE insertions and designed primers for 20 locations to compare to the non-reference insertions. We then utilized

previously gathered breed dog DNA samples and used whole genome amplification (WGA) to have working DNA for PCR, given the limited amounts of genomic DNA we have for each sample (Table 1). We then used PCR to show polymorphism between the different breeds and reference/non-reference SINE_Cf insertions (Figure 9).

Goal 3. Obtain sequences for functional experiments. Sequences from the non-reference insertions were looked at for confirmation of polymorphism observed during genotyping. We predicted that those sequences with variance from the consensus will assist in our hypothesis of a structural characteristic that is correlated to cytoplasmic mobilization to the ribosome and insertion efficiency.

Identifying Non-Reference and Reference Insertions

Reference insertions were identified using the BLAT Genome Browser and the consensus SINE_Cf sequence. Each primer set at varying chromosomal locations were designed using the process previously explained. Identifying non-reference insertions utilized similar mapping process but instead looked for sequences that had SINEs absent from their reference genome. The discoveries of non-reference SINEs were previously performed by Dr. Halo from whole sequenced genomes of breed dogs and mapping them against CanFam3.1. Here, I further mapped these insertions identified specifically within a breed dog for which we had genomic DNA as a mechanism to later clone them for functional assays (below). For reference SINEs, I identified roughly 20 within CanFam3.1 that were more or less ‘intact’ and varied slightly from the consensus (*i.e.*, no two reference SINEs were identical at the nucleotide level). Using the predicted SINE insertion location, primers were designed for amplification and use of identifying further polymorphic insertions (Table 2). Following the identification of non-reference and reference insertions (Figure 8), confirmation of working PCR on WGA gDNA was preformed

using GAPDH (Figure 9). Genotyping PCR using the WGA gDNA from various breed dogs was conducted under regular PCR conditions with both non-reference and reference SINEs to locate multiple polymorphic SINE insertions (Figure 10).

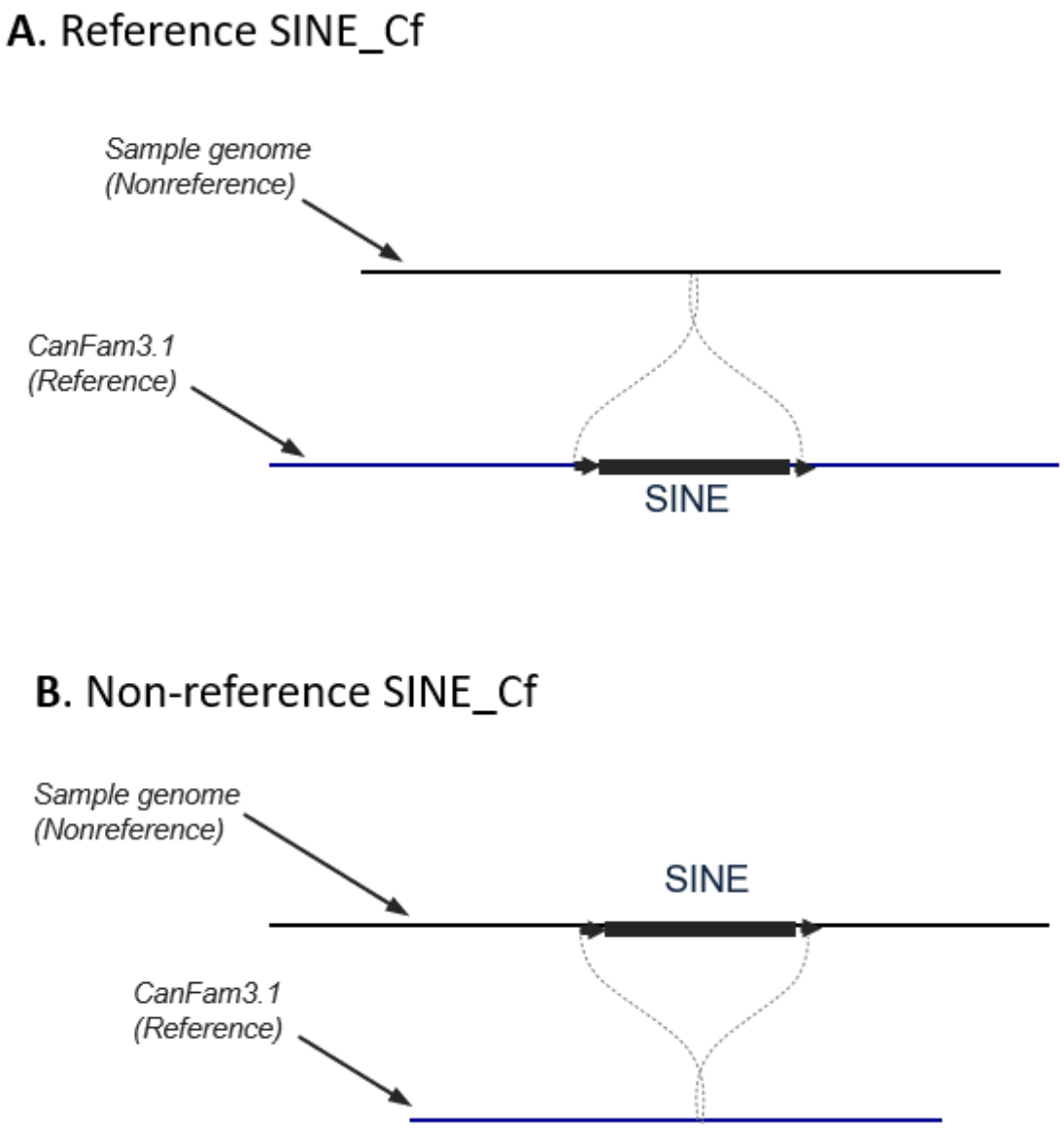


Figure 8: Non-reference and reference insertions

A) Reference SINE_Cf shows a sample genome compared to the reference CanFam3.1 genome known SINE insertion whereas; B) the non-reference depicts an insertion in a sample genome where the reference CanFam3.1 genome does not. Utilizing both non-reference and reference SINE_Cf's will confirm polymorphism within various SINE_Cf insertions.

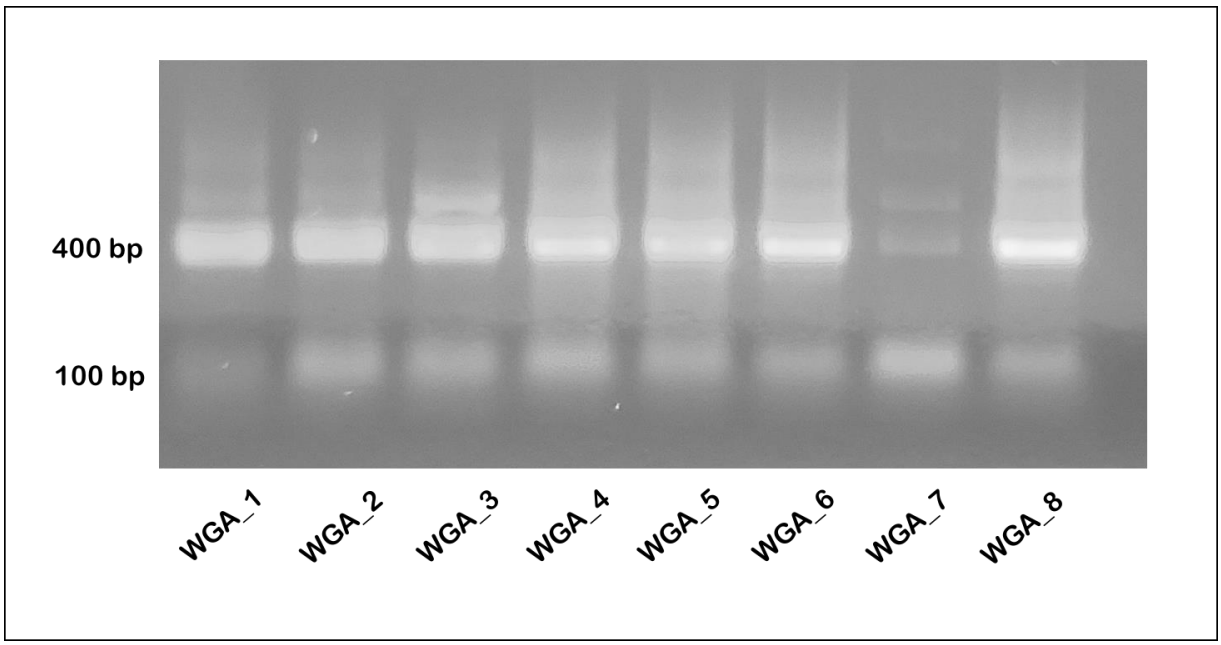


Figure 9: PCR screen of whole genome amplified samples by GAPDH

16 breed dogs and our control DNA from the boxer, Zoey were used for genotyping. This gel depicts WGA 1-8 (breeds detailed in Table 1) from a GAPDH PCR reaction to show that these samples were amplified correctly and can perform under PCR conditions (with the exclusion of WGA_7 as it did not appear very clearly on the electrophoresed gel).

Table 2: Genotyping of non-reference and reference insertions

| Location | Length (bp) | Forward | Reverse | Reference |
|---|---|---|---|---|
| Chr1_19675135 | 913 | TTCTGTTGGTTGTCTCAATG | ACCAGGTGAGTTGTCATAAC | ✓ |
| chr9_47773389 | 630 | CTCCTGCATGTGAGCTCC | AACCATAACCTTATATCTAACCCTG | ✓ |
| chrX_23569384 | 539 | gccatatggacagattacatagaa | gtgtatattgtgaaatggtcgc | ✓ |
| chr10_24575890 | 647 | TGAGGCAGAGATAGGAGAGAG | ATGGAGCCTTGCTTCTCC | ✓ |
| chr3_49837239 | 630 | CACTATCAGTGTTTGAACATGC | CAAACAATATGAATTAATCTCATGA | ✓ |
| chr12_56429765 | 630 | gaccacatgaataggcacagat | gataaagcggtacaaaattggaa | ✓ |
| chr26_24823379 | 628 | tcaagattcacctatgtggtagc | tgcctcgttctctctctctc | ✓ |
| chr37_10762733 | 575 | gatccatatgtttttcttacgg | gagagagaatgagagcaggg | ✓ |
| chr18_40171644 | 483 | cttcccaaccatctgcaa | accttgtatgtaaattacggctg | ✓ |
| chr31_14533602 | 531 | atagagggaagccaagagg | ctgacttttggatttggct | ✓ |
| chr12_37391048 | 561 | atgcaggaagcctgacat | aagatacaaggattggaaagg | ✓ |
| chr9_45172720 | 587 | gccttggaaccagagacag | cagcaagagtgagagagcaga | ✓ |
| chr1_73554738 | 585 | atcaggctcctcccatttt | gttcagtcggttaagtgtccac | ✓ |
| chr16_4722620 | 663 | gagaagatggaggtgggc | caaattggccaaccaactg | ✓ |
| chr21_35732239 | 608 | gaccctcttccctgtttcag | ccaagatgcttccccc | ✓ |
| chr27_35672969 | 592 | CCTTGTGTTTCATTTCAAGC | GCCCACTTCTCTTACAGGATA | ✓ |
| chr13_9471735 | 561 | gtgaggacaccagataacagc | aagccatccatattttcatcac | ✓ |
| chr12_20124012 | 596 | gttgaatacaagtggcaagagtg | gtgaacttagatgcaaaatttgtc | ✓ |
| chr17_46080704 | 670 | tcccagttgggctctatatg | ggtaatttctctacggttatgattc | ✓ |
| chr34_17642968 | 664 | attcaagagagacacacacagag | cattacttagctctgcctatttagaa | ✓ |
| chr1_42663957 | 400 | TCATTTCCCCATGCAAGTTCA | CCTGGGCCTTGGTTTTCTTAC | ✕ |
| chr2_78122938 | 489 | TGTGCATGGATTGGCTTCT | AGGACCTTGAAGAGACCATCA | ✕ |
| chr3_79314029 | 380 | TCTGCACTGAGAAAGCACAT | CTCTGAAACCAAGTGTCTGTC | ✕ |
| chr4_17775700 | 302 | TGTGGTCAAAGCACAGGAAA | AGGGGAATTTCAGGACAGACT | ✕ |
| chr4_44409338 | 691 | GTGCAGAGCTCAAGACATCAG | CCGATTAACCAGCCCTTGTC | ✕ |
| chr5_52186136 | 396 | TCGACGTATACGCAGAGAGGT | TCACTGAAACTGTTTGGCCA | ✕ |
| chr6_16714574 | 487 | TCAGTGATTGGTGTTGTGTGT | CACCCTCAACCTCATCCTCA | ✕ |
| chr6_39333015 | 468 | TGTAGACTTTTGGCGGAGGT | AATTCCAAAGGCCCAGGAGA | ✕ |
| chr8_9826039 | 369 | TTAGAGCAGCACAGACCCTT | TGAAGCAGACACTAAGACAGGT | ✕ |
| chr9_27288596 | 532 | TAGAAGGAAGCTTTGCCTGC | TGCTGGAGTTCATTGTGTGG | ✕ |
| chr10_10355036 | 348 | TGGCTCTTGACATGGGAAAC | GGCTTCTACACCCACCCC | ✕ |
| chr10_23841649 | 398 | TCAGGAAGAAAGGTGGGCAG | GGAGACCCTGCTGCCTCA | ✕ |
| chr11_45513514 | 467 | TGGGACTCATGCACAGGAAT | TGGGTTATTGCTATGACGGGA | ✕ |
| chr12_56638786 | 497 | TCTGGAAGAAAGCCTCGTGT | GGAGAACCCCTACAATGTCAT | ✕ |
| chr12_59717732 | 831 | TATAGGGTTGGCAGGGAAGG | TGCCAGGAATGAAACAAGCA | ✕ |
| chr15_54592932 | 819 | TGAGTTTGTCACCTTCCCAGT | TGGTGTAGGAGGTCAAACAAA | ✕ |
| chr20_20399760 | 486 | GTGCAGCACCCTTTCTTTGG | GCTGGGGTTTTGACTTCAGG | ✕ |
| chr21_11240910 | 500 | TCAAGTCTAATCGATGACTGACA | AGGCAGACTCTTGGGATAGT | ✕ |
| chr22_52052586 | 494 | TGATTGCTTCAGGAAAACAAG | TGCAGGGAGTTGAGAAGTGA | ✕ |
| chr27_6438544 | 484 | TCAGTTTATCACATGCCACCT | GGGGAAACTGAGTGGGGAAA | ✕ |

Using the primer strategy as outlined within the text, non-reference and reference primers were designed to map polymorphism between various SINE insertions.

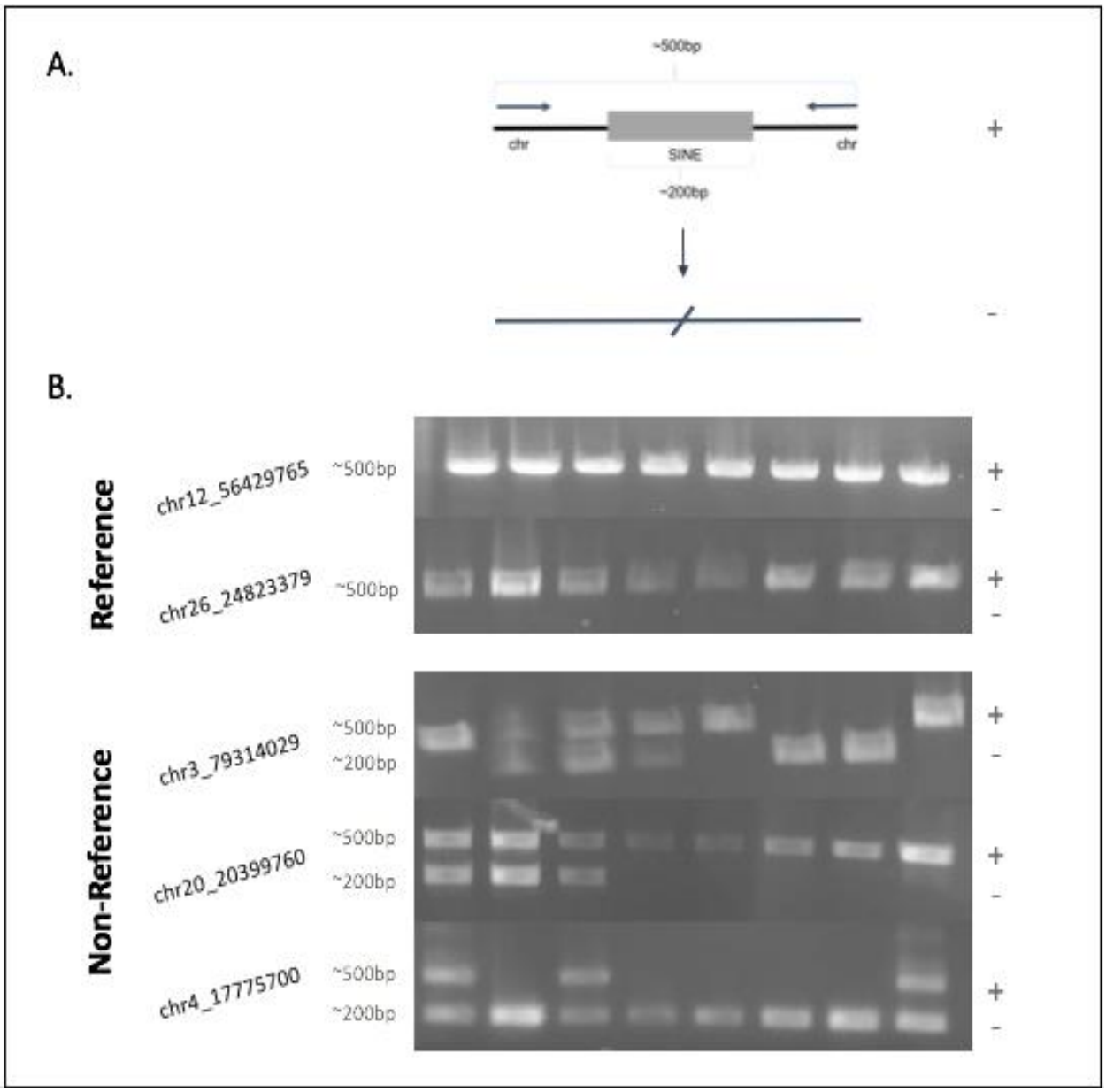


Figure 10: Genotyping visualization using non-reference and reference SINE primers.

A) Showing the reference and non-reference insertions depicted by +/-, refer to Figure 8 for non-reference and reference insertions; B) Electrophoresed PCR products from reference and non-reference primers using WGA gDNA 1-8 excluding 7 with control gDNA from boxer, ‘Zoey’.

Polymorphism Within tRNA Derived SINEs

There were 20 reference SINE insertions utilized to analyze the polymorphism between non-reference insertions and reference insertions. Each reference and non-reference insertion had a different chromosomal location, ensuring variance and accuracy when assessing the non-reference insertions. Using PCR and agarose electrophoresis, the reference SINEs were expected to have a product ~500bp in length whereas the non-reference SINEs, in most cases, seemed to contain a product the same length as the reference as well as a smaller SINE about 200bp (Figure 10). Each reaction was performed with WGA gDNA from breed dogs (Table 1) and as a control, 'Zoey' gDNA. 'Zoey' is gDNA from a boxer which is the baseline canine gDNA used for all amplifications of SINE_Cf.

Following screening for polymorphism, we used sequencing to confirm their identification and polymorphic status assists in the understanding of each subsequent SINE identified. The sequences from non-reference and reference insertions prove varying divergence from the consensus SINE_Cf sequence (Figure 11), which will be useful in the analysis of mobilization of each subsequent SINE identified to hone in on structural motifs related to mobilization in the continuation of this research.


10 20 30 40 50 60 70 80 90 100 110
SINEC_Cf
chr2_78122938
chr3_37219113
chr3_8982357
chr3_89994515
chr4_44409338
chr4_71144107
chr4_71114855
chr4_76648845
chr4_77702161
chr5_50461339
chr5_5200814
chr5_65673675
chr5_86377120
chr6_19839395
chr7_23746260
chr7_55657454
chr7_63789776
chr8_20646394
chr9_27288596
chr9_55650336
chr10_35567571
chr10_39912736
chr10_20635019
chr10_55227532
chr10_60188973
chr11_18429440
chr11_48491691
chr12_9509802
chr13_49246069
chr13_61325819
chr15_63171940
chr15_63693761
chr23_50957059
chr24_24405569
110 120 130 140 150 160 170 180 190 200 210 220 230

Figure 11: Alignment for confirmation of polymorphism within various SINE_Cf insertions SINEs were amplified after using BLAT to design non-reference primers in various SINE_Cf insertions genomic locations within the canine genome. The SINEs were confirmed by electrophoresed PCR and sequenced before analyzing mobilization via retrotransposition assay. Sequences indeed show polymorphism between insertions, confirming our second goal.

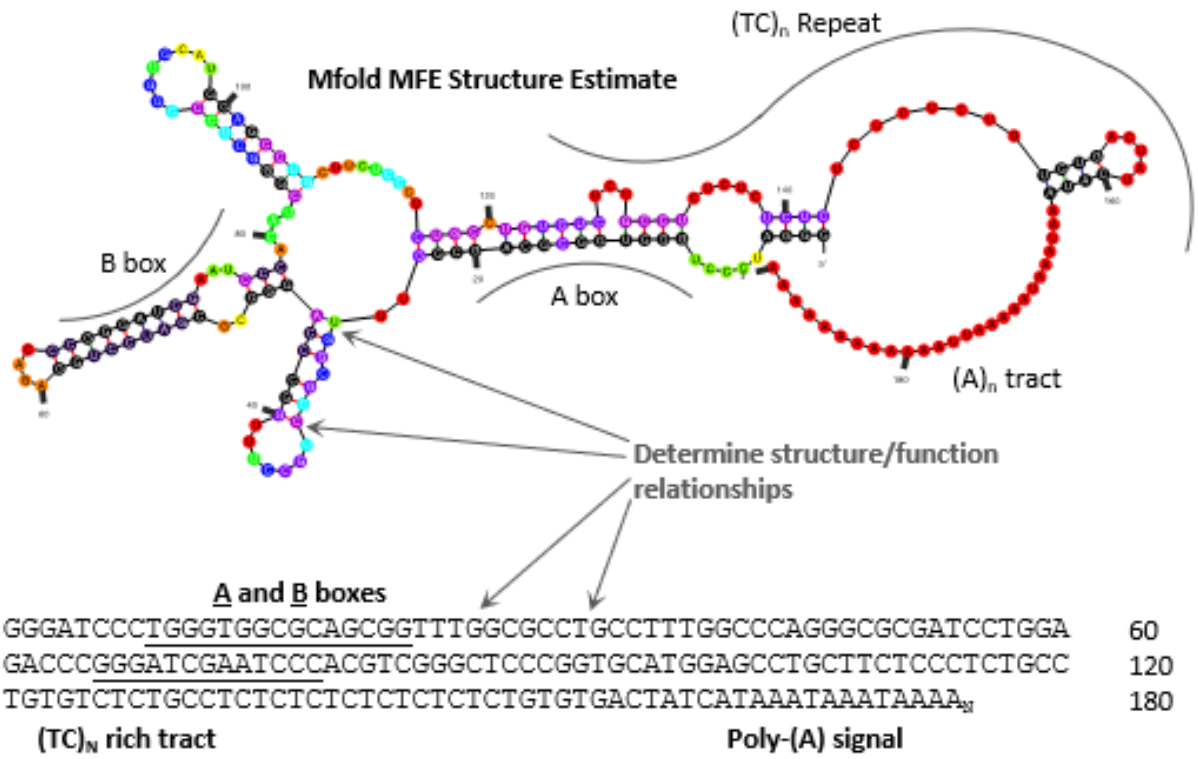


Figure 12: Secondary structure of tRNA SINE 'mapping' of polymorphisms over the structure The consensus sequence of SINE_Cf mapped to determine a secondary structure of the Lys-tRNA SINE. With the utilization of this model it can be used to identify a specific region and potential structural characterization that assists in mobilization to the ribosome.

Testing For Mobility of SINE_Cf Using *ex vivo* Assay

We hypothesized that mutations to bases involved in retrotransposition of the SINE can be identified by correlation to decreased rates of mobilization in our *ex vivo* assay. To adequately test our hypothesis, we have successfully cloned around 200 SINE_Cf elements from different percentages of divergence range for nucleotide and length variance. Nucleotide length ranged from 210-230 bp and divergence from SINEC_Cf ranges from ~2-16%. Of the 200 SINE_Cf non-reference insertions that were amplified and cloned, a total of 35 were tested for mobilization through the retrotransposition assay. Also noteworthy in this regard, this test is the best assay you could run for comparing SINE activities of SINE cores. The core of the SINE, meaning the activity of the RNA intermediate, without anything that could interfere with the activity (e.g., genomic location and nearby genome/chromosomal features).

The ex *vivo* retrotransposition assay is able to test the mobility of a specific SINE by utilizing L1 containing an internal self-splicing report cassette marker, $neo^R$. This 14/15-day assay (see Methods) gives us a more crystalized view of our predictions for the efficiency of retrotransposing SINE_Cf elements. When conducting the *ex vivo* assay, we utilized L1Hs and *Alu* as positive controls. Upon completion and clearly defined foci, we then fixed and stained, and counted foci. Average foci counts for SINE_Cf non-reference range from $3.3 \times 10^{-1}$- $2.7 \times 10^{1}$. The consensus SINE_Cf performed extremely well in comparison, which is to be expected because it is intact and highly mobile within the canine genome (Figure 14). When comparing the consensus to the non-reference SINE_Cf elements commonly, the ones that did not have prominent foci had diverged from the consensus more than those that had more retrotransposition events. Based on this analysis of this subset of 35 SINE elements, these data support our hypothesis; specifically, we observe differences in relative mobilization of distinct

element cores. Expanding these assays to include the additional 165 SINE_Cf, 40 B1 and B2 constructs should shed light on specific regions and provide further support for our hypothesis (see Discussion). This study contributes to this larger analysis from my preliminary work and to planned assays to pinpoint cellular interactors of the tRNA intermediate during retrotransposition.

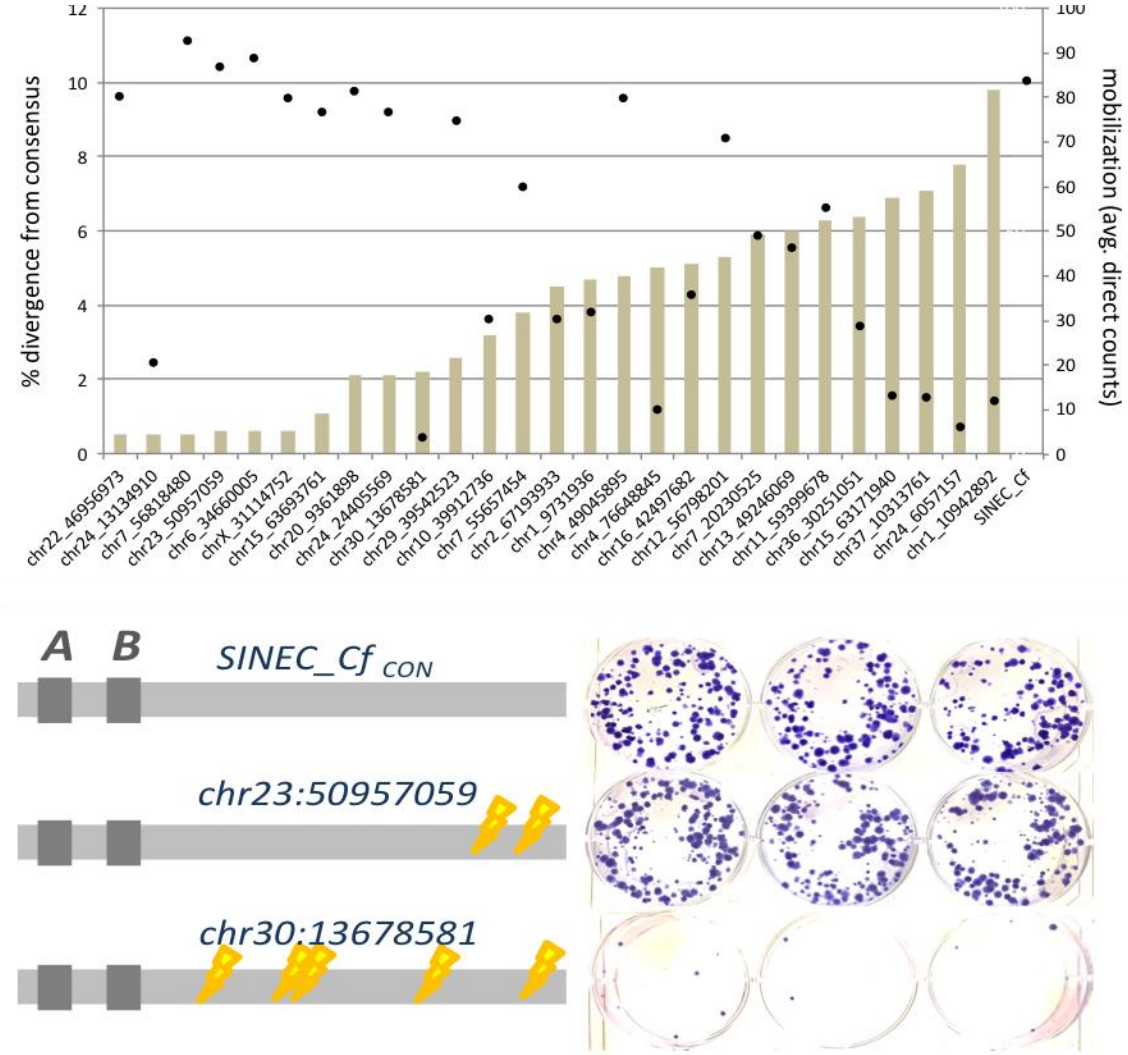


Figure 13: Mapping SINE_Cf insertions against the consensus for mobilization analysis

Utilizing the SINE_Cf consensus in a retrotransposition assay for a mobilization baseline was found to be extremely active. Following the consensus, 26 other SINE_Cf elements were used for retrotransposition assays. In most cases the sequences with some divergence from the consensus, such as chr:50957059, are still highly active. Whereas sequence with larger divergence, such as chr30:13678581, are not mobile under the retrotransposition assay conditions.

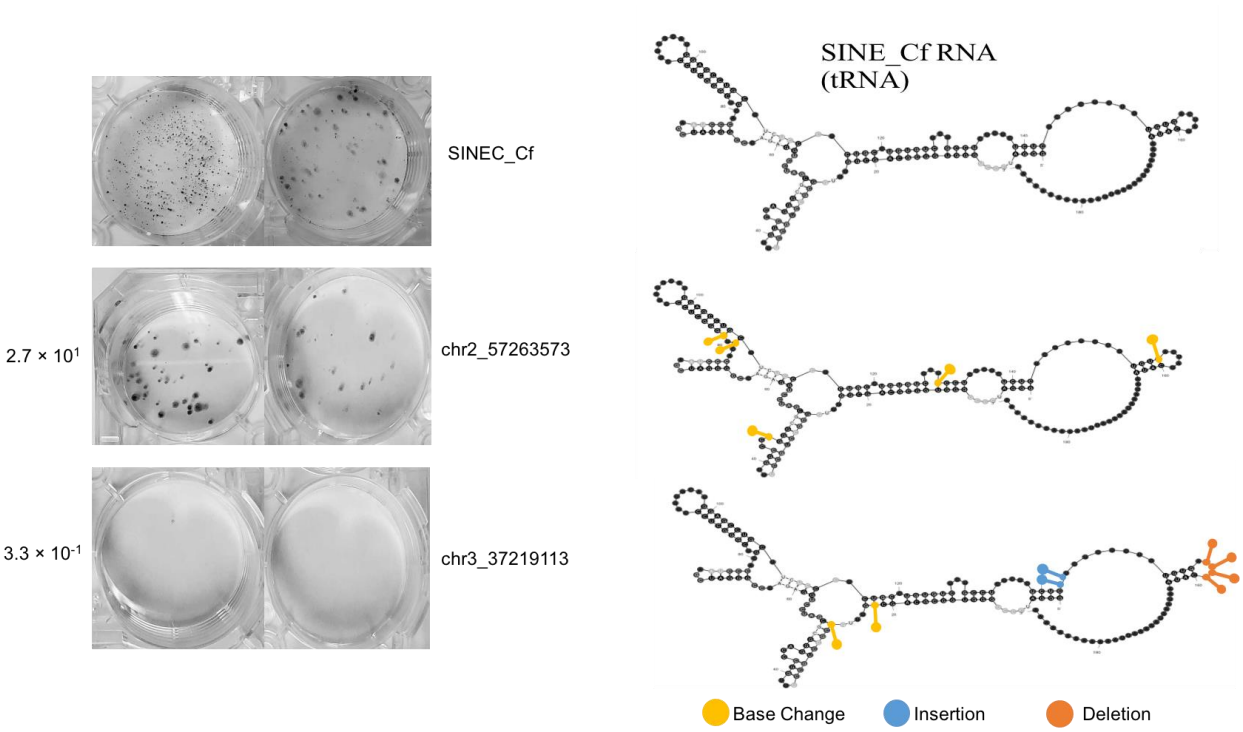


Figure 14: Retrotransposition assays of SINE_Cf and consensus SINE_Cf

Subset of the *ex vivo* mobilization assays preformed on 35 SINE_Cf elements. Utilizing SINEC_Cf as a baseline for optimal retrotransposition, elements such as chr2_57263573 and chr3_37219113 portray how well SINE_Cf with varying divergences mobilize. Figure 11 shows that although they are only slightly different from one another, chr2_57263573 is closer to the consensus as indicated by the insertions and deletions seen in chr3_37219113, explaining the amount of foci compared to the consensus. This is consistent with our previous data (Figure 13).

# CHAPTER 4: DISCUSSION

SINEs have been shown to have a significant role in being templates for driving genome expansion, thus contributing to new mutations, variation, and altering gene expression, as well as affecting phenotypes and natural biology of a species. As mentioned, *Alu* and B1 are 7SL derived SINEs and evolutionarily distinct to the tRNA-derived SINEs such as SINE_Cf and B2. Because of their high insertion rates, owing in part to its presence and ongoing activities in human, *Alu* have been linked to a multitude of human diseases such as Hemophilia A and B, X-Linked Agammaglobulinemia and Breast Cancer (Payer, 2017). Similar to *Alu*, the most recent (and putatively) active canine SINE, SINE_Cf, has insertions remain insertionally polymorphic between dogs and segregate within some breeds, having been shown to cause phenotypic effects. For example, merle coat color patterning seen in border collies and some other breeds is caused by a SINE_Cf insertion in the dog *SILV* gene that has been linked to the pigmentation (Spady, 2008). Characterization of the insertion in trait-mapping of affected dogs led to the identification of the orthologous gene in humans, possibly with role(s) in the auditory–pigmentation genetic condition of Waardenburg syndrome (Clark, 2006). Similar trait-mapping has identified other disease-associated human orthologs by virtue of dog phenotypes and the identification of the causative loci in diseases. For example, SINEC_Cf insertions have linked the hypocretin family of receptors to narcolepsy in Doberman pinschers, the *PTPLA* gene to centronuclear myopathy in retrievers (Pele M., 2005), and the *IGF1* gene with small body size in dogs (Gray MM, 2010). Understanding how tRNA SINEs such as SINE_Cf are able to mobilize at such high rates, thus being able to contribute to multiple phenotypes, is the purpose of this research.

SINE mobilization of *Alu* has been owed to its ancestral structure derived directly from the *Alu* domain of SRP, whose mouse counterpart B1 has been shown to have a decrease in

amplification due to a mutation in the SRP9/14 binding site that remains intact in *Alu*. Once this mutation was identified and reversed, the B1 was able to amplify at a rate comparative to *Alu* and B2 in the same assay background (Dewannieux M, 2005). However, as described in this thesis, the mobility machinery of tRNA derived SINEs have not been given enough attention. These SINEs differ from *Alu* and B1 due to their tRNA source. tRNA-derived SINEs do not contain the SRP9/14 that makes the 7SL-derived SINEs so insertionally successful. In our research we were able to identify around 200 non-reference SINE_Cf, in addition to 20 B1 and 20 B2 insertions. The focus of my work was centered around amplifying, cloning and eventually analyzing the mobilization of SINE_Cf. As I was able to begin the preliminary work with both B1 and B2, our incoming MS student, Madison Baltzly, will be not only continuing the SINE_Cf assays, but also expanding data with B1 and B2 SINEs.

Each of the SINE_Cf's were amplified from genomic DNA and cloned for a functional analysis of their mobilization (Figures 9, 10; Tables 1, 2). We found that the non-reference SINE_Cf had many polymorphic novel insertions in various chromosomal locations within the canine genome. The variance in nucleotide identity/divergence between each insertion is crucial when looking for mobility within the core SINE sequences as we now have a greater chance at locating a structural component that assists in cytoplasmic mobilization and ribosomal association. Thus, testing the direct mobilization ability of individual SINE cores, when compared to the consensus element of the family, is designed to begin to directly test our hypothesis that mutations to bases involved in retrotransposition of the SINE can be identified by correlation to decreased rates of mobilization in our *ex vivo* assay.

Of the 200 SINE_Cf insertions isolated, amplified and cloned; 35 were subjected to retrotransposition assays to assess the mobilization capabilities of each novel insertion. In

running the assays, we predict that as sequences diverge from the consensus SINE_Cf we will witness a lesser mobilization rate. We observed that, as the core SINE sequence diverged from the consensus SINE_Cf sequence, the less mobile the core became, which aligns with our previous data and conclusions (Figure 13). When referring to the sequences of the insertions that underwent mobilization assays (Figure 11), there is clear variance between the elements which will assist in our future assessment and isolation of the SINE_Cf element. For example, when we observe non-reference SINE insertions at chr2_57263573 and chr3_37219113, we saw that there was a drastic difference between retrotransposition events although their sequences varied only slightly. This tells us that these small changes are potentially linked to the mobilization of SINE_Cf and instances such as these are expected as more SINE_Cf, B1 and B2 insertions are analyzed. Because of this, SINE_Cf sequences help us to isolate specific portions that may be associated with successful mobilization also map a secondary structure for SINE_Cf (Figure 12).

The future goals of this research are to complete all retrotransposition assays of SINE_Cf as well as B1 and B2. Once each of these insertions are analyzed for mobilization, using the sequences to make a more complete secondary structure to isolate regions associated with mobilization. Using Ahl et al. (2015) model of isolating the SRP 9/14 from *Alu* to further assess zones of the secondary structure, we will use SINE_Cf consensus secondary structure to map and portion so that when regions that are removed can be analyzed for mobility (Figure 15). If the element is no longer able to retrotranspose after the artificial deletion of a specific region, we can assume that location within the sequence is essential for mobility.

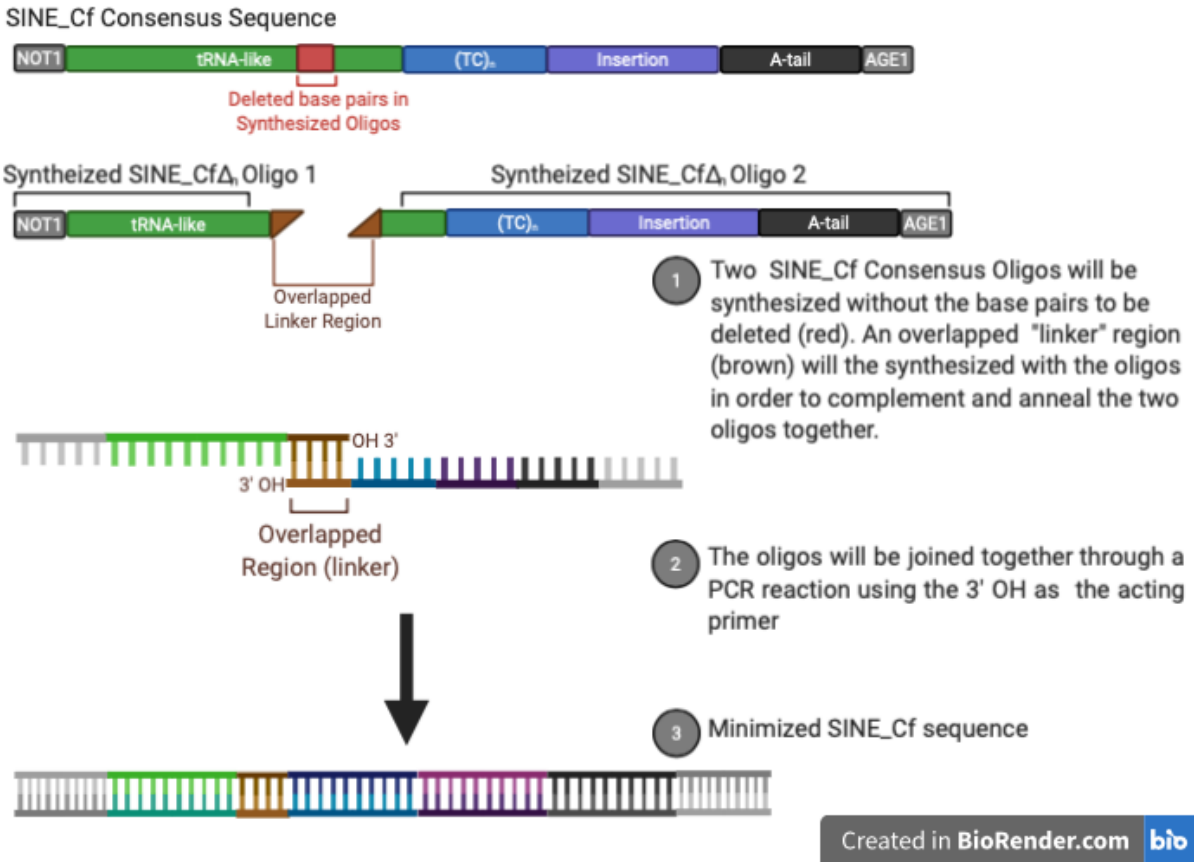


Figure 15: Structural mutagenesis of SINE_Cf

Using the SINEC_Cf secondary structure to identify potentially mobile characteristics by analysis of SINE_Cf retrotransposition assays compared to the sequences for efficiency, we can confirm, and pinpoint areas associated with successful retrotransposition. 1) By deleting portions of the consensus that have been previously analyzed by the *ex vivo* assay to then 2) join the SINE back together without the potentially active characteristic to 3) amplify, clone and assess for mobility. By completing structural mutagenesis of SINEC_Cf, we can accurately depict a structural component that assists in mobilization through the cytoplasm to associate with the ribosome.

(Figure created by Madison Baltzly, BGSU MS Biology Student)

# REFERENCES

Ahl, V., Keller, H., Schmidt, S. & Weichenrieder, O. 2015. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. Mol Cell, 60, 715-727.

Bannert, N. & Kurth, R. 2006. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet, 7, 149-73.

Batzer, M. et al. 2002. Alu repeats and human genomic diversity. Nature Reviews Genetics, 3, 370-379.

Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. 2011. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet, 12, 187-215.

Boeke, J. & Stoye, J. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin, J., Hughes, S. & Varmus, H. (eds.) Retroviruses. New York, NY: CSHL Press.

Campos-Sanchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. & Makova, K. D. 2014. Genomic landscape of human, bat, and ex vivo DNA transposon integrations. Mol Biol Evol, 31, 1816-32.

Carnell, A. et al. 2003. The Long (LINEs) and the Short (SINEs) of it: Altered Methylation as a Precursor to Toxicitiy Toxicological Sciences, 75, 229-235.

Clark, L., Wahl, J.M., Rees, C.A., Murphy, K.E. 2006. Retrotransposon insertions in SILV is responsible for merle pattering of domestic dog PNAS, 103, 1376-1381.

Collart, M. & Weiss, B. 2020. Ribosome pausing, a dangerous necessity for co-translational events Nucleic Acids Res, 48.

Cordaux, R. & Batzer, M.A. 2009. The Imapct of Retrotranspons on Human Genome Evolution.

Nat Rev Genet, 10, 691-703.

Dewannieux, M, & Heidmann, T. 2005. L1-mediated Retrotransposition of Murine B1 and B2 SINEs Recapitulated in Cultured Cells. Journal of Molecular Biology 349, 241-247.

Dridi, S. 2012. Alu Mobile Elements: From Junk DNA to Genomic Gems. Scientifica.

Dunker, W., Zhao, Y., Song, Y. & Karijolich, J. 2017. Recognizing the SINEs of Infection: Regulation of Retrotransposon Expression and Modulation of Host Cell Processes. Viruses, 9.

Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. 2016. Retrotransposons as regulators of gene expression. Science, 351, aac7247.

Eo, J. & Heui-Soo, K. 2013. Repeat polymorphism analysis for examining genetic variably and the implications for specific traits in dog breeds. Genes and Genetic System 88, 273-278.

Gray, M.M., Sutter, N.B., Ostrander, E.A., Wayne, R.K. 2010. The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. BMC, 8.

Hancks, D. C. & Kazazian, H. H., JR. 2012. Active human retrotransposons: variation and disease. Curr Opin Genet Dev, 22, 191-203.

Hancks, D. C. & Kazazian, H. H., JR. 2016. Roles for retrotransposon insertions in human disease. Mob DNA, 7, 9.

Hancks, D. C. & Kazazian, H. H., JR 2012. Active Human Retrotransposons: Variation and Disease. Current Opinion in Genetics and Development, 22, 191-203.

Jern, P. & Coffin, J. M. 2008. Effects of retroviruses on host genome function. Annu Rev Genet, 42, 709-32.

Karijokich, J., Zhao, Y., Alla, R., & Glaunsinger, B, 2017. Genome-wide Mapping of Indection-

Induced SINE RNAs Reveals a Role in Selective mRNA Export. Nucleic Acids Res, 45, 6194-6208.

Katzourakis, A., Rambaut, A. & Pybus, O. G. 2005. The evolutionary dynamics of endogenous retroviruses. Trends Microbiol, 13, 463-8.

Kramerov, D. A. & Vassetky, N. S. 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity (Edinb), 107, 487-95.

Kurth, R. & Bannert, N. 2006. The Evoluationary Dynamics of Human Endogenous Retroviral Families. Genomics and Human Genetics, 7, 149-173.

Lisch, D. & Burns, K. H. 2018. Editorial overview: Genome architecture and expression: Mobile elements at work. Curr Opin Genet Dev, 49, iv-v.

Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R. D., Feschotte, C. & Craig, N. L. 2013. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc Natl Acad Sci U S A, 110, 234-9.

Payer, L. M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D., Liu, C., Boeke, J.D., Avaramopoulos, D., & Burns, K.H. 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. PNAS, 114, E3984-E3992.

Pele M., Tiret, L., Kessler, J.L., Blot, S., Panthier, J.J. 2005. SINE Exonic Insertion in the PTPLA gene leads to multiple splicing defects and segragates with the autosomal recessive centronuclear myopathy in dogs. Human Molecular Genetics, 14, 1417-1427.

Platt, R. N., 2nd, Vandewege, M. W. & Ray, D. A. 2018. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res, 26, 25-43.

Spady, T. C., & Ostrander, E.A. 2008. Canine Behavioral Genetics: Pointing Out the

Phenotypes and Herding up the Genes. Am J Hum Genet., 82, 10-18.

Taylor, M. S., Lacava, J., Mita, P., Molloy, K.R., Huang, C.R., LI, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T., Rout, M.P., Boeke, J.D., & Dal, L. 2013. Affinity Proteomics Reveals Human Host Factors Implicated in Discrete Stages of LINE-1 Retrotranspositon. Cell, 155, 1034-1048.

Voorhees, R. & Hegde, R.S. 2015. Structures of the scanning and engaged states of the mammalian SRP-ribosome complex. Biochemistry and chemical biology.

Wang, W. & Kirkness, E.F. 2005. Short Interspersed Elements (SINEs) are a major source of canine genomic diversity. Genome Res, 15, 1798-1808.

Weiner, A. 2002. SINEs and LINEs: The Art of Biting the Hand That Feeds You. Curr Opin Genet Dev., 14, 343-50.

Wessler, S. R. 2006. Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A, 103, 17600-1.