Interaction of Zinc Finger Proteins 146 & 507 with the Transposable Element LINE-1

A Major Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degrees of Bachelor of Science

in

Biochemistry and

Biology and Biotechnology

by

Eric Larsen

CDR Deadline: April 26, 2018

APPROVED:

Jeanne Lawrence, PhD
Dept. of Neurology
UMass Medical School
MAJOR ADVISOR

Robert Dempski, PhD
Chemistry and Biochemistry
WPI CBC Project Advisor

David Adams, PhD
Biology and Biotechnology
WPI BB Project Advisor

This report represents the work of WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the projects program at WPI, please see http://www.wpi.edu/academics/ugradstudies/project-learning.html
ABSTRACT

The long interspersed nuclear element-1 (LINE-1 or L1) is a transposable element comprising approximately 17.5% of the human genome. Its transposition has been linked to a variety of genetic disorders, resulting from gene disruptions, nucleotide deletions, duplications, and general chromosomal instability. L1 ORF2 protein is toxic even without active transposition. L1 is not strongly expressed in most somatic cells, but is transcribed in some germ, embryonic stem, and cancer cell lines. Zinc finger proteins (ZNFs) bind nucleic acids, and this MQP project investigated two ZNF’s, 146 and 507, initially identified in our lab (unpublished data) as potential L1 ORF-2-binding proteins from Encode Chip-Seq data. The functions of ZNF146 and ZNF507 functions are unknown. In this project, a bioinformatics approach was used to identify potential binding sites on the L1 element for the two ZNFs, and the sites were verified biochemically using pull-down assays. The binding sites for ZNFs 146 & 507 were found to be highly conserved across primate L1 subfamilies, with ZNF507 showing a specific period for loss in binding. Fluorescence microscopy showed that both ZNF146 and ZNF507 are localized to the nucleus. Over-expression of either ZNF through plasmid based transfection was toxic to HeLa cells, and showed possible signs of cell phase arrest in HEK293 cells. Knockdown of either ZNF in HEK293 cells resulted in a morphology similar to senescence. Immunoprecipitation of flag-tagged ZNF507 followed by mass spectroscopy identified arginine methyl transferase PRMT5 as a ZNF507-binding protein. Future experiments will analyze transcriptome alterations in ZNF146 and ZNF507 knockdown cells, and determine associated proteins for ZNF146.
# TABLE OF CONTENTS

Signature Page ........................................................................................................ 1

Abstract .................................................................................................................. 2

Table of Contents .................................................................................................. 3

Acknowledgements ................................................................................................. 4

Background ........................................................................................................... 5

Project Purpose ..................................................................................................... 22

Methods ................................................................................................................. 23

Results .................................................................................................................... 33

Discussion ............................................................................................................. 50

Bibliography ......................................................................................................... 54
ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr. Jeanne Lawrence, the Major Advisor of this project, for allowing me to work in her lab and for guiding me. I would also like to thank Postdoctoral fellow Kevin Creamer of the Lawrence lab for his tremendous input and assistance during the course of the project. Finally, I would like to thank WPI Professors David Adams and Robert Dempski for helping initiate the project, for ongoing advice, and for assistance in editing the report.
BACKGROUND

The Junk Genome

It has been known since the completion of the human genome project that only a small portion of the human genome is comprised of protein-coding sequences. The vast majority of the genome is non-protein-coding (non-coding) and had been relatively understudied for years as “junk DNA” (Palazzo and Gregory, 2014). Research regarding the importance of the non-coding regions of the genome has become more prevalent in recent years, in part due to large-scale efforts to identify functional regions in the human genome. An example is the Encyclopedia of DNA Elements (ENCODE) project (Palazzo and Gregory, 2014), whose authors state that biochemical function can be assigned to an astounding 80% of the genome. Although the percentage is still highly debated, much research focuses on determining how much of the human genome is functionally significant at the organismal level (Nesbo, 2013).

The first use of the term “junk DNA” can be traced back to the 1960s, when researcher Susumu Ohno used the term to describe what are now known as pseudogenes (Ohno, 1972). The application of the term has since expanded to apply to any DNA sequence that does not play a functional role in development, physiology, or some other function at the organismal level (Palazzo and Gregory, 2014). This application of the term is central to the ongoing debates about the quality, or even existence, of “junk DNA” within humans and other organisms. But is any part of the genome really junk? Does every sequence fulfill some function that simply has yet to be identified?
Non-Coding Genome Composition

Although the functions of non-coding portions of the genome are not well understood, the characteristics of these sequences are generally known. The non-coding genome makes up roughly 98% of the genome, and can generally be broken down into four major groups: LINES (21%), Introns (21%), SINES (13%), and Tandem Repeats (12%) (Figure 1) (Alexander et al., 2010; Meisenberg and Simmons, 2017).

![Non-coding genome composition chart](chart.png)

Figure 1. Divisions of the Non-Coding Genome. Note this is only one proposed model, there are many others which account for the range in proportions of the various elements that shift with continuous research. Adapted from Principles of Medical Biochemistry, Figure 7.5.

The groupings are very dynamic, and change from year to year, and source to source. One group is the conserved sequences (not shown in the figure), shared between humans and other mammals, which constitute about 5% of the genome. The conserved sequences are comprised mainly of long non-coding RNAs (lncRNAs), theorized to represent about 0.4% of the genome, and sequences which are bound by transcription factors, representing about 8.5-12% of the genome (Palazzo and Gregory, 2014; Meisenberg and Simmons, 2017). Another group of
sequences are known as \textbf{pseudogenes}, genes in which one copy has been mutated such that it has lost its ability to encode a functional protein. Pseudogenes can also be produced when an RNA is reverse transcribed and then integrated into the genome. It is estimated that there are roughly 12,600 to 19,700 pseudogenes within the human genome, comprising about 6% (Pei et al., 2012). Introns comprise a large portion of non-protein-coding regions, making up roughly 20-40% of the human genome, although this value is likely overestimated due to the fact that the introns often contain elements of the other groups like transposable elements, thereby inflating their contribution (Bulmer 1987; Gibbs, 2003; Palazzo and Gregory, 2014).

A larger fraction, roughly two-thirds, of the genome consists of \textbf{highly repetitive DNA}. These regions are extremely varied among individuals of the same population, because the sequences can expand or contract through processes like unequal crossing over or replication slippage (Koning et al., 2011). Some highly repetitive sequences have been found to play a role in gene regulation, while other sequences play critical roles in chromosomal maintenance (Koning et al., 2011). The most heavily studied repetitive elements are \textbf{transposable elements} (TE). Transposable elements make up a large proportion of the non-coding DNA, and include various well-described retro-elements such as \textbf{Short and Long Interspersed Nuclear Elements} (SINEs and LINEs), endogenous retroviruses, and cut-and-paste DNA transposons (Rodic and Burns, 2013; Palazzo and Gregory, 2014). These elements in total make up roughly 45% of the human genome (Beck et al., 2011). While high in copy number, a vast majority of these elements are inactive in humans due to high degradation by mutations (Gregory, 2005). As a result of the degeneracy, estimates of the TE portion of the human genome widely vary, are thought to make up at least two-thirds of the genome (Gregory, 2005; Koning et al., 2011; Palazzo and Gregory, 2014). While TE’s make a large contribution to the genome, their function is known for only a
few. After factoring in other sequence elements, such as unique elements which do not fall into any of the previous categories or tandem repeats and segmental duplications, it is thought that only 1-2% of the genome contains **protein coding sequences** (Alexander et al., 2010; Palazzo and Gregory, 2014; Meisenberg and Simmons, 2017).

**Long Interspersed Nuclear Element-1 (Line-1 or L1)**

The non-LTR long interspersed nuclear element 1 (LINE-1/L1) is a transposable element that comprises approximately 17.5% of the human genome (Beck et al., 2011 & repeatmasker.org). Due its nature as a transposable element, most L1 elements are 5’ truncated upon transposition and are therefore rendered inactive (Beck et al., 2011). Of the roughly 950,000 fragments of L1 that exist in the genome, only roughly 150 copies remain intact and theoretically can move through transposition. Of that portion, it is estimated that only roughly 60 copies are potentially active in the genome (Brouha et al., 2003; Bao et al., 2015). The generally conserved structure of L1 (**Figure 2**) is comprised of two open reading frames (ORF1 and ORF2), along with 3’ and 5’ UTRs.

![Figure 2. Diagram of the L1 Structure and Domains. Abbreviations: C, cysteine-rich domain; CC, coiled coil; CTD, carboxyl-terminal domain; RRM, RNA recognition motif; EN, endonuclease; RT, reverse transcriptase; SVA, SINE-R/VNTR/Alu; UTR, untranslated region; AAA, poly(A) tail. Adapted from Gregory, 2005.](image)

Studies regarding the function of the two L1 open reading frames have speculated that ORF1 functions to produce a chaperone for other factors required for transposition (Martin, 2006; Beck et al., 2011; Rodic and Burns, 2013). ORF2 on the other hand contains the elements
needed to perform transposition, including domains for both reverse transcriptase and endonuclease activity. Presence of ORF2 alone has been shown to induce cell senescence without the need for full L1 retrotransposition (Wallace et al., 2008). In addition to L1 itself, the ORF2 is required for other repeat elements, like Alu and SVA, to transpose successfully (Beck et al., 2011; Rodic and Burns, 2013). Due the truncation of the 5’ end of the gene, the second open reading frame can exist as an independent fragment, where only the second open reading frame is inserted back in the genome. 3’ truncated fragments also exist by the same logic, although these are less numerous and less studied.

**L1 Retrotransposition**

L1 sequences, although inactive in most somatic cells, are active in the germ line cells, embryonic stem cells, and some cancer lines (Rodic and Burns, 2013). Although the mechanism of retro-transposition is not fully understood, it can be inferred from the two key enzymes of the sequence. The first is integrase, an endonuclease that cleaves at the site of integration to generate a staggered break (Figure 3) (Alberts et al., 2002). The other enzymatic activity comes from the reverse transcriptase.

![Figure 3. L1 Retrotransposition and Life Cycle.](image)

*Figure 3. L1 Retrotransposition and Life Cycle.* Shown on the left is L1 inhibition which occurs within somatic cells. The mechanism of inhibition is currently unknown. Shown on right are cell lines in which L1 expression is active (i.e. neoplastic, cancer, and germ cell lines) along with the theorized mechanism of retrotransposition. Adapted from Rodić & Burns, 2013 and Richardson, et al., 2015.
The first step of retro-transposition involves the transcription of genomic L1 into RNA, which is mediated by RNA polymerase II from an internal L1 promoter. Next, the RNA is translated into the two L1-encoded proteins: ORF1, an RNA-binding protein, and ORF2, a protein with reverse transcriptase and endonuclease activities. These proteins associate with the L1 transcript, and the resulting ribonucleoprotein (RNP) complexes then move to the nucleus. The third step is known as target-primed reverse transcription (TPRT). During TPRT, the ORF2 protein cleaves the target DNA, often at a 5′-TTTTAA-3′ consensus sequence, and uses the 3′ hydroxyl group to prime the reverse transcription reaction (Beck et al., 2010; Richardson et al., 2015). Because the L1 life cycle generates staggered DNA breaks at the target reintegration site, cell host proteins that mediate DNA repair are then likely responsible for integration of the L1 sequence back into the genome (Richardson et al., 2015).

**Proposed L1 Functions**

Several studies have examined the expression and function of L1 on a larger scale. It has been theorized that full-length L1 mRNA is expressed very little, if at all, in somatic tissues, with detection possible mainly in germ line cells (Ergun et al., 2004). However, other studies have noted that L1 expression may occur in both a normal human brain and other cell lines (Richardson et al., 2015). L1 has been shown to be responsible for many genetic disorders stemming transposition, including: gene disruption, nucleotide deletions, duplications, and general chromosomal instability through heterologous recombination (Beck et al., 2010). L1 may also play an important role in genomic functions ranging from regulation of gene expression to being influencing X-inactivation in females (Beck et al., 2011). The L1 reverse transcriptase may also have a role in the creation of processed pseudogenes. L1-associated DNA damage across a
wide spectrum of tissue samples has also suggested that L1 may function as an endogenous mutagen in somatic tissues (Richardson et al., 2015). Other studies have narrowed their focus on the expressional effects of individual open reading frames. ORF2 splice products have been found to lead to genetic damage, mirroring expression of the exogenous full-length L1 (Beck et al., 2011). Due to the truncated nature of ORF1 on the 5’ end, and a lack in regions of homology, the function of the ORF1 protein is much less understood. However, the ORF1 protein is known to be a nucleic acid chaperone that is required for transposition, and also is a high affinity RNA binding protein with L1 RNA, to form a ribonucleoprotein particle (Beck et al., 2011).

**L1 Evolution and Phylogeny**

LINE-1 elements also have a rich and dynamic evolutionary history, extending back more than a hundred million years, characterized by the rise, fall and replacement of subfamilies (Figure 4) (Khan et al., 2006; Waters et al., 2008; Cordaux and Batzer, 2009; Boissinot and Sookdeo, 2016). Most of the data concerning LINE-1 biology and evolution are derived from the human and mouse genomes, which are often assumed to hold true for all placentals and most eutherian. This is due to the fact that the mode of L1 evolution has been conserved since the origin of mammals (Boissinot and Sookdeo, 2016). In mammals, only the most recently evolved group of elements is usually active at one time, so that a single family of progenitors is usually producing novel insertions. In the long term, this will lead to a ladder-shaped phylogeny, stemming from the replacement of one family by a younger one in a cycle (Boissinot and Sookdeo, 2016). This mode of evolution is consistent with the arms race belief, covered in further detail later. Reptiles and fish on the other hand have several highly divergent families concurrently active in the same genome. These active families have coexisted for an extended
period of time, and their divergence may pre-date the origin of vertebrates (Khan et al., 2006; Boissinot and Sookdeo, 2016).

![Figure 4. Phylogenetic Evolution of L1 and Naming Conventions. Shown is the evolutionary progression from ancestral L1 subfamilies still present in the genome to modern human L1 (PA1), along with the naming conventions of the 3'-UTR and 5'-ORF2 sequences. Made using data from Smit et al., 1995 and Khan et al., 2006.](image)

The naming conventions for the L1 subfamilies are based on 5’ truncated or 3’ ORF2 regions (Smit et al., 1995). The 3’ ORF2-derived sequences are labeled as M or P, M for mammalian, or P for primate. Subfamily specifications are shown by a following letter ranging from A to E, A being the youngest, and so on. The subfamilies are then arranged by age in reverse chronological order, with the most recent family numbered as 1, and older families increasing in age until the sequence reaches the point of divergence from a previous subfamily, like primates divergence from mammalian (Figure 4). 5’ truncated sequences are named in a similar manner, M for mammalian or P for primate. However subfamily specification is broader, only containing the reverse chronological numbering. The 3’ and 5’ families can generally be grouped together in terms of age and the evolutionary pathway which leads to the current human L1 subfamily (Figure 4). Furthermore, it is also theorized that primate L1 diverged from modern mammalian L1 at roughly the MA6 subfamily (Smit et al., 1995; Khan et al., 2006; Boissinot and Sookdeo, 2016).
L1 Repression and Evolutionary Pressure

Although L1 expression is known to be toxic and must be repressed in somatic tissues, the exact mechanism of inhibition is also not fully understood at this time. Several reports have identified factors which inhibit L1 expression in some cases, but these do not explain long-term repression of endogenous L1 (Rodic and Burns, 2013; Walter et al., 2016; Guler et al., 2017; Sokolowski et al., 2017). Others report the simultaneous suppression of Alu elements along with L1 elements. A common thread between most findings is that they all share similar modes of repression, either DNA methylation or histone modification, to suppress active L1 sequences (Rodic and Burns, 2013; Paco et al., 2015; Guler et al., 2017). However, what is clear from the phylogeny and evolution of L1 (previously discussed) is that there has been a pattern of L1 reactivation in which L1 bypasses repression, becomes active, and once again must be repressed. This theory is believed to be a type of arms race between the host, which represses L1 transposition, and L1, which evolves to bypass repression, and has gone on throughout the evolution of L1. L1 elements are also known as selfish elements, that persist over time due to their replicative advantage over the host genome (Paco et al., 2015).

This is not to say that L1 expression is without benefits. As previously stated, L1 sequences play an important role in gene expression and other activities. In a broad sense, L1 sequences can arguably be one of the most significant and dynamic forces operating on the mammalian genome, having potentially significant impact of the genesis of genetic diseases like cancer and genome evolution (Cordaux and Batzer, 2009; Pei et al., 2012; Boissinot and Sookdeo, 2016). L1 retro-transposition can cause various effects in the genome like rearrangement of gene promoters, enhancers, and even exons by the transcription of flanking non-L1 sequences (Cordaux and Batzer, 2009). Genes generally are affected when they are a
target site for insertion, and there is evidence to say this can also happen without interfering in the coding region (Cordaux and Batzer, 2009). L1 insertion into the UTR of a gene can affect the regulation of its transcription and translation (Boissinot and Sookdeo, 2016). Even outside of the transposition itself, homologous sequences are created in non-homologous regions, providing a platform for recombination on misaligned chromosomes, potentially resulting in chromosomal rearrangements (Cordaux and Batzer, 2009; Boissinot and Sookdeo, 2016). If L1 was repressed with no possibility of escape, it would not have the same potential to drive evolution.

**Zinc Finger Proteins (ZNFs)**

Zinc finger proteins (ZNFs) are a widely diverse class of proteins which bind to specific motifs within their target structures, either nucleic acid or amino acid sequences. Although they are diverse, all zinc finger proteins generally share a common feature, a coordination with at least one zinc ion (Krishna et al., 2003; Laity et al., 2001). The zinc ion acts to stabilize the integration of the protein itself and therefore is not directly involved in target binding. The fingers are secondary structures held together by the zinc ion. These zinc fingers contain domains that typically serve to interact with other elements, such as binding with DNA, RNA, proteins, or other small molecules. The specific association of these proteins with the zinc ion classifies them as zinc finger domains. Many types of zinc finger structures form in a predictable manner, based predominately on specific sequence features, and these features have been used to identify many new ZNFs (Krishna et al., 2003).
Classes of ZNF-Binding Proteins

Zinc finger proteins are most often classified by the type of fold group of their zinc fingers. Fold group types include a C2H2-like finger, gag knuckle, treble clef finger, zinc ribbon, Zn2/Cys6-like finger, Taz2 domain-like, short zinc binding loops, or metallothionein domains (Krishna et al., 2003). Furthermore, the type of interaction the finger has can often be linked to its fold group (Krishna et al., 2003). Each type of fold group binds zinc and associated proteins through different features within the protein sequence. For example, the treble clef motif consists of one β-hairpin at the N-terminus and one α-helix at the C-terminus, that together contribute two zinc-binding domains. The first two ligand-binding domains derive from the zinc knuckle (a turn caused by interactions with two zinc ions), and the other two ligand-binding domains are donated by the N-terminal turn of the helix. On the other hand, gag knuckle fingers are composed of two short β-strands connected by a turn, one zinc knuckle, followed by a short helix or a loop. The size of the fingers of these two fold groups also differs, as do the sizes of any of the fold groups (Krishna et al., 2003). These differences in size and structure of the zinc fingers allow each of the different classes to bind different targets and specific binding motifs by methods beyond just the sequence of the binding region of the finger itself. These differences allow for zinc finger proteins to specifically bind to not only DNA and RNA sequences, but also amino acid sequences of other proteins, as seen with some zinc ribbon domains (Krishna et al., 2003). Some zinc finger domains can even bind specifically to a molecule outside of the central dogma, such as to a phosphate residue. An example is the Phosphatidylinositol-3-phosphate-binding domain, a subset of the treble cleft fold group (Krishna et al., 2003).
Cys2His2 (C2H2) Zinc Finger Structures and DNA Binding

One of the most abundant zinc fingers in the eukaryotic genome is the Cys2His2 zinc finger (C2H2). This finger type is 20-30 aa long, and often binds to DNA (Krishna et al., 2003). This mode of DNA binding is similar among nearly all DNA-binding zinc fingers. Furthermore, for C2H2 fingers, the finger-to-DNA binding occurs in a very conserved manner (Krishna et al., 2003; Persikov et al., 2008). Nucleic acid-binding C2H2 zinc fingers bind to the major groove of DNA through the N-terminus of the α-helix. C2H2 zinc fingers also contain three conserved hydrophobic amino acids at positions -12, -3 and +4, in addition to the two conserved cysteines and histidines. These seven amino acid residues are necessary and sufficient to fold peptides properly, as shown by using a designed-synthetic peptide to create a C2H2 zinc finger (Luchi, 2001).

Recognition of specific DNA sequences along with increased stability and binding strength is created through tandem C2H2 zinc fingers. Tandem zinc fingers allow for cooperative binding of the α-helices of several C2H2 zinc fingers arranged in tandem to longer and more complex DNA sequences. While a single finger usually specifies binding to three nucleotides, tandem zinc fingers allow for longer binding sequences, which increases the specificity and stability of the binding (Luchi, 2001) (Figure 5).
The C2H2 zinc finger distribution within a protein can also play a role in DNA binding. Based on the distribution of zinc finger domains, ZNFs are divided into three classes: triple, multiple-adjacent, and separated-pairs (Figure 6) (Luchi, 2001). This classification is useful for predicting binding activity. Triplet and multiple-adjacent types bind to DNA at consecutive fingers. However, for separate-pair ZNFs, a minimum of two consecutive zinc fingers are required for DNA-binding (Luchi, 2001). The link between pairs is also flexible and allows for separate pairs to interact and bind DNA. This is not however universal, as separated-pairs of zinc fingers can interact with DNA at different contact points, and they do not necessarily interact directly with each other (Luchi, 2001).
C2H2 fingers are also among the most common DNA-binding motifs found in eukaryotic transcription factors (Krishna et al., 2003). Finger predictions can be made from an amino acid sequence without the need for crystal structure analysis (Persikov et al., 2008). Finger predictions are generally made by detecting cysteines spaced four residues apart at the amino terminal, and histidine residues with the same spacing at the carboxy terminal (Krishna et al., 2003, Persikov et al., 2008a, b). Binding residues are usually deduced by looking only within the finger predictions, although this approach is tentative without the protein crystal structure.

**Zinc Finger Protein 146 (ZNF146)**

This project investigated two ZNF’s: 146 and 507. These two ZNFs were identified in our lab (unpublished data) from Encode Chip-Seq data as potential LINE-1-binding proteins. Zinc finger protein 146 (ZNF146), also referred to as “only zinc fingers” (OZF) (Chalony et al., 1994), is a zinc finger protein composed almost entirely of zinc finger domains (Persikov et al., 2008b, Heil and Noor, 2012, Zerbino et al., 2018). The gene encoding the 292 amino acid protein is located on chromosome 19, band 13.12 (Zerbino et al., 2018). The sequence contains ten equally spaced C2H2 zinc finger domains that constitute most of the protein (**Figure 7**).
However, very little is known about ZNF146 function. ZNF146 over-expression has been linked to certain cancers, like pancreatic and colorectal cancer (Ferbus et al., 1999; Ferbus et al., 2003). Targeted expression of ZNF146 has also been shown to be associated with impaired mammary development in mice (Xie, 1997). ZNF146 has also been found to interact with telomeric protein hRap1 in colon carcinoma (Antoine et al., 2005). Apart from these studies, little else is known about ZNF146 at this time.

**Zinc Finger Protein 507 (ZNF507)**

Similar to ZNF146, not much is known about zinc finger protein 507 (ZNF507). This protein is much larger than ZNF146, with a length of 957 amino acids. The gene is located on chromosome 19, band 13.11 (Zerbino et al., 2018). This protein contains only nine zinc fingers, three of which are grouped at the amino terminus, and the other 6 are grouped at the carboxy terminus (Figure 7) (Persikov et al., 2008b; Zerbino et al., 2018). Two recent genomic findings have been made in regards to ZNF507 function. The first study revealed a locus that includes ZNF507 and confers risk to neurodevelopmental disorders across diagnostic boundaries (Talkowski et al., 2012). The second study found ZNF507 to be one of many genes with increased expression linked to schizophrenia (Curtis, 2015).
**ZNF-Binding Proteins**

Zinc finger proteins have the ability to associate with other proteins. Many of the documented interactions with ZNF proteins occur with proteins that affect gene regulation (Schneider et al., 1997; Schultz et al., 2002; Sun et al., 1996). Examples include binding to regulatory proteins like KAP-1 (a H3K9 methyl-transferase) and Myc (Schneider et al., 1997; Schultz et al., 2002). In these cases, the ZNF proteins act as a guide for the associated proteins to allow for specific DNA sequence targeting by interacting with proteins that would otherwise be unable to bind DNA at that location, or are unable to bind DNA at all by themselves (Sun et al., 1996). The type of gene regulation affected depends on the associated protein, the ZNF mostly acts as a guide to the DNA. For example, the ZNF interaction with KAP-1 leads to gene repression, while the interaction with Myc leads to gene activation (Schneider et al., 1997; Schultz et al., 2002). ZNFs have also been found to interact with other ZNFs. This interaction can occur via Krüppel-type (KRAB) domains. KRAB domains are transcriptional repressors found within some ZNF proteins (Schneider et al., 1997; Schultz et al., 2002; Sun et al., 1996). Alternatively, ZNFs can interact with other ZNFs via non-KRAB domains. KRAB domain-containing ZNFs are of high interest for study because they themselves can directly regulate gene expression.

**Project’s Relation to Ongoing Research in the Lawrence Lab**

This study of the repetitive LINE-1 element and its potential binding proteins is part of a larger project of interest to the UMMS Lawrence lab concerning chromosomal organization and substructure. Similar to the cytoplasmic cytoskeleton, the nucleus also has a skeleton or scaffold.
This nuclear skeleton is composed of nuclear laminins, non-coding RNAs, and protein complexes which help organize and stabilize the DNA in the nucleus. LINE-1 elements, due to their highly repetitive, conserved, and dispersed nature across a majority of the genome, have the potential to play role in this structure. Nuclear scaffold structure is currently a topic of interest to our lab’s postdoctoral fellow Kevin Creamer, whose work analyzes the means by which the DNA is organized and positioned within the nucleus. He theorizes that non-coding RNA’s comprise a major portion of this nuclear skeleton, and his work aims to verify and characterize this structure. Since L1 elements are so abundant in the nucleus, they may play a role in genome organization.
PROJECT PURPOSE

The long interspersed nuclear element-1 (LINE-1 or L1) is a transposable element comprising approximately 17% of the human genome. Due to its transposition, L1 has been shown to be responsible for many genetic disorders resulting from gene disruptions, nucleotide deletions, duplications, and general chromosomal instability. L1 is not strongly expressed in most somatic cells, but is transcribed in germ line cells, embryonic stem cells, and some cancer cells. Zinc finger proteins (ZNFs) bind nucleic acids, and this MQP project investigated two ZNF’s that bind L1 DNA: ZNF146 and ZNF507. These ZNFs were initially identified in our lab (unpublished data) from Encode Chip-Seq data as potential L1-binding proteins. The functions of ZNFs 146 and 507 are unknown, but ZNF146 is over-expressed in pancreatic and colorectal cancers, and is associated with impaired mammary development in mice, while 507 is associated with a locus over-expressed in neurodevelopmental disorders. In this project, a bioinformatics approach will be used to identify potential binding sites on the L1 DNA element for the two ZNF proteins, and if present, to determine their potential conservation throughout evolution. The binding sites will also be verified biochemically using pull-down assays. The cellular location of ZNF146 and ZNF507 will be determined using flag-tagged proteins and fluorescence microscopy. ZNF-binding proteins will be identified using immunoprecipitation assays with flag-tagged ZNFs.
METHODS

Cell Culture

Cell lines used during this project included HeLa (human cervical carcinoma cell line) (ATCC), HEK293 (human embryonic kidney cell line) (ATCC), and human TIG-1 fibroblast cell lines (ATCC). Cultures were fed using 10mL of DMEM+10%FBS+1XP/S, and incubated in T-75 flasks at 37C until confluent. Cultures were split for subculture after aspiration of the DMEM media and a wash with 5mL of Hanks culture medium. Cells were then treated with 2mL of 1X TrypLE Express (ThermoFisher) for five minutes at room temperature to dislodge the cells, then the cells were re-suspended in an additional 2mL of DMEM+FBS+PS. 1mL of this cell suspension was placed into 10mL of fresh DMEM+FBS+PS in a new T-75 flask, and the flask was incubated until confluent, and then either sub-cultured or used for an experiment.

Coverslip Culture and Fixation

In order to visualize cells under the microscope, cells were cultured and fixed onto glass coverslips. Cell suspensions were taken after testing, and cultured on coverslips placed into wells of a 6-well plate. 500uL of cell suspension was cultured with 1.5mL of fresh medium for at least 24 hours prior to fixation at 37C. Following culture, the medium was aspirated from the wells, and the coverslips were then washed twice using hanks balanced salt solution (HBS). For standard fixation, cells were moved into coplin jars on ice containing CSK buffer long enough to rinse. Coverslips were then moved into coplin jars, also on ice, containing a mixture of 9mL CSK, 0.5mL 10X trition, and 0.5mL VRC for 4 minutes. Cells were then fixed to the coverslips by placing them in a third coplin jar containing 4% paraformaldehyde at room temperature, for 10 minutes. Coverslips were then stored in 70% ethanol within a new 6-well plate at 4C until imaged.
Immunofluorescence Assay (IFA)

The cellular locations of plasmid-expressed Flag-tagged ZNFs 146 and 507 were determined by an immunofluorescence microscopy assay (IFA) using a Flag antibody. IFA microscopy was also used to visualize Ki67, laminin-B, and tubulin. The following antibodies were used: anti-mouse Flag M2 (Sigma), anti-rabbit Ki67 (Abcam), anti-mouse laminin B (Santa Cruz), and anti-rabbit tubulin (Lab Vision). Coverslips stored at 4C were moved into a coplin jar containing 1X PBS for at least 10 minutes to rehydrate the cells. During this time the primary antibody stain mixtures were prepared using 500uL of 1X PBS/1% BSA with the desired concentration of antibody for the target diluted within with 1uL of RNase inhibitor added. The coverslips were placed cell side down on top of 70uL of the desired primary antibody stain on parafilm placed on a glass plate. This was then sealed and incubated in a humidified chamber for 1 hour. Following incubation, coverslips were washed in coplin jars containing 1X PBS, then 1X PBS+0.1% triton, and finally 1X PBS, respectively for 10 minutes each on a shaker. Coverslips were then stained with the secondary antibody, 1:500 dilutions of either Alexa or Dylight 488 or 594, that recognized the respective animal host of the primary stain, in the same manner as the primary stain, except covered in tinfoil to keep dark. The coverslips were washed again as described for the primary antibodies, except they were covered to keep the coplin jar in the dark on the shaker. Coverslips were then stained using a DAPI solution for 30 seconds and then washed in 1X PBS. Coverslips were then mounted on microscope slides using one drop of Vectashield mounting medium and sealed using nail polish. Slides were stored at 4C when not being imaged. Coverslips containing cells were visualized under 100X magnification.
Protein Expression Alterations Through Transfections

In order to examine the effects of differential expression of ZNFs 146 and 507, expression plasmids for each ZNF were introduced into cells by transfection. Flag-tagged expression plasmids were used for over-expressions, and shRNA expressing plasmids were used for knock downs. Cells were dislodged from culture flasks when they were about 90% confluent. Approximately 10 million cells per T75 flask were washed with Hanks culture medium and were then dislodged through treatment with 2mL of 1X TrypLE for five minutes at room temperature. 2.5mL of DMEM+10%FBS without antibiotics was then used to deactivate the TrypLE and to re-suspend the cells. For continued culture, 500uL of this suspension was moved into a new T75 flask with 10mL of DMEM+FBS with antibiotics. 2mL of the remaining volume of cell suspension was then spun down for three minutes at roughly 1,000rpm. The medium was aspirated from the cell pellet which was then washed and spun down twice in 5mL of 1X PBS under the same centrifugation parameters. The cell pellet was then re-suspended in 4mL of MEM without antibiotics, and 0.5mL of the suspension and 0.5mL of MEM was added to coverslips which were placed into individual wells of a 6-well plate.

Next, the plasmid solutions were prepared for transfection. 10ug of each flag-tagged over-expression plasmid, or an empty vector control, were added to 500uL of opti-MEM. For each plasmid mixture, 12uL of lipofectamine 2000 was added to 500uL of opti-MEM. After four minutes the lipofectamine mixtures were added to the plasmid mixes to undergo lipofection for twelve minutes. 500uL was added two one well plate for each plasmid mixture, two sets for each plasmid. After the addition of all plasmid mixtures, an additional 1mL of MEM was added to the wells, and the cultures were mixed through shaking of the well plate. Cells were allowed to
transfect for 6 hours at 37C, after which the medium was replaced with fresh MEM. The cells were then cultured for 48 hours at 37C prior to fixation on the cover slips.

**Nuclear Extractions**

In order to isolate or quantitate the ZNFs, nuclear extracts were prepared from transfected cell cultures. Using at least 12, 10-cm dishes of 90% confluent cultures, cell pellets were obtained through treatment with 2mL TrypLE, re-suspended in 2mL of MEM for each dish, pooled, and then centrifuged at 1,000 rpm for 3 minutes. Cell pellets were then washed and spun down in 50mL of 1X PBS twice more under the same parameters.

To lyse the cell pellets, two buffers were prepared and kept on ice. Buffer A was made using 10mM HEPES pH 7.5, 1.5mM MgCl₂, 10mM NaCl, 1mM DTT, and 1 protease inhibitor cocktail tablet. Buffer B was made with 20mM HEPES, 500mM NaCl, 1.5mM MgCl₂, 25% glycerol, and 1 protease inhibitor tablet. Once washed, the cell pellets were re-suspended in 4mL of Buffer A, transferred into 4 Eppendorf tubes, and spun down for 4 minutes at 3,000rpm in a cold environment (20C). The supernatant was removed, and each pellet was re-suspended in 1mL of Buffer A supplemented to 0.075% NP-40 (37.5ul 10%/5ml) and incubated on ice for 5 minutes. The cell suspension was then spun down for 3 minutes at 3,000rpm at 20C once again, and the supernatant was removed. The pellets were then washed once with 1mL Buffer A, by thoroughly re-suspending and again spinning to pellet the nuclei. The nuclear extract was then made by thoroughly re-suspending the cell pellet in 1mL Buffer B. The four tubes were then combined into a single 15ml conical tube, and vortexed for 5 seconds at max speed and placed on ice 15 minutes. The samples were then vortexed once more before spinning 12,000rpm for 5 minutes at 4C to remove cell debris. The supernatants of identical samples were combined into a
new 15ml conical tube. The nuclear extract was then divided into aliquots and frozen on dry ice before storing in -80C for later use.

**Immunoprecipitations**

In order to verify the presence of and to purify the specific ZNF proteins within the nuclear extract, target proteins were isolated from the samples using immune-precipitation. Once thawed 20uL of extract was taken and placed on ice for use as an input. Next, 400uL of extract was taken from the sample and added to 1mL of IP dilution buffer for an extract from 4x10cm plates and placed on ice. The extracts were pre-cleared with IgG beads, and then incubated with anti-flag conjugated magnetic beads overnight at 4C on a rotator. The beads were then capture, washed with 500uL of TBS+0.1% Triton for 5 minutes on a rotator, and recaptured for a total of three washes. After the second wash in the cycle the beads were transferred into a new tube. The same cycle of washes were then used to wash the beads in just TBS buffer. Following the washes, the beads were then incubated with 50uL of 3X flag peptide for 30 minutes at 4C on the rotator. Beads were then capture and elutions were taken. The beads were then washed and incubated with flag peptide for a total of three elutions. Elutes were pooled following the final elution. The elutions were then treated with trichloroacetic acid (TCA) and incubated for 10 minutes at 4C. Samples were then spun down at 14,000rpm, also at 4C, with the supernatant being removed leaving only the protein pellet. The pellets were washed with 500uL of cold 95% acetone and spun down again. The process was done for a total of two acetone washes. Following the washes 20uL of 2X sample buffer was added to each sample, each sample was mixed with minute traces of NaOH, and each sample was boiled for 10 minutes on the 80C heat block.
**ZNF Biotinylated Oligo Pulldown Assay**

In order to test whether the predicted LINE-1 DNA ZNF-binding motifs are functional, nuclear extracts containing ZNFs were used in biotinylated DNA pulldown assays. First, biotin-labeled double stranded (ds) DNA oligos containing the predicted LINE-1 binding motif (shown below) were purchased commercially (IDT). Then, streptavidin beads were bound to the DNA oligos. To prepare the beads for the assay, the lyophilized oligos were dissolved to a concentration of 1mM (1 nmol/uL) in water. The oligos were then diluted to 20uM in water. 100 uL of complementary oligos with 100uL NEB buffer 2 and 700uL water were then added to the mix to total 1mL. These mixtures and 200uL of salmon sperm DNA were incubated at 95C for 4 minutes, then allowed to cool to room temperature. The oligo-bound beads were now primed to bind proteins in the extracts.

For each reaction, 15uL of magnetic beads was added to 100uL of Wash Buffer A, made with 20mM Tris pH 7.5, 150mM NaCl, 1.5mM MgCl2, 10% glycerol, and optionally 10nM Zinc Acetate. The beads were captured on a magnet and re-suspended in 250ul/reaction in Buffer A. The beads were captured, washed, and recaptured a total of three times. After the third capture, the beads were re-suspended in 150uL of Buffer A per sample. 100uL of this suspension was used for binding to oligo, and 50uL was used for pre-clearing.

For the beads used for oligo binding, 25uL of annealed biotin-dsDNA and 2.5uL of BSA (20mg/mL) were added to 0 100uL of Buffer A, and this was added to the 100uL suspensions. This mixture was incubated for 30 minutes on the shaker and flicked every couple of minutes. The beads were than captured and re-suspended in 500uL of Buffer A, and recaptured for a total
of three washes. After the final wash, the beads were re-suspended in 20uL of Buffer A per sample.

Next, the pre-clearing samples were prepared to capture DNA/protein complexes. For each sample tested, a reaction mix was made containing 150uL of 2X dilution buffer, 28uL of water, 15uL of 20X protease inhibitors, 2uL BSA (20mg/mL), 2uL of annealed salmon sperm, 100uL of nuclear extract, and the 50uL sample of washed beads (without biotin-dsDNA annealed). The 2X dilution buffer consisted of 40mM Tris pH 7.5 (2ml 1M/50ml), 20% glycerol (10ml/50ml), 0.1% NP-40 (0.5ml 10%/50ml), 2mM MgCl$_2$ (100ul/50ml), and optionally 10nM Zinc Acetate. The reaction mixtures were incubated for 15 minutes on a shaker at room temperature, and were inverted every few minutes. The beads were then captured and the supernatant was moved into a new tube for each sample. 20uL of annealed beads in Buffer A was then added to the supernatant and incubated for 30 minutes on a shaker at room temperature, again flicking every few minutes. The beads were then captured and washed once in 500uL of Buffer A, capturing again immediately after suspension. The beads were then washed again with 500uL of Buffer A, but were incubated for 5 minutes on the shaker before capture. The beads were then washed in 500uL of wash buffer B under the same parameters as the previous wash. Buffer B consisted of 10mM Tris pH 7.5, 75nM NaCl, 1.5mM MgCl$_2$, 10% glycerol, and optionally 10nM Zinc Acetate. Finally the beads were washed one final time with 500uL Buffer B, then immediately re-suspended, inverted, and recaptured. The beads were then eluted with 35uL of 1X Laemmli sample buffer with BME added. These samples were placed in the 80C heating block for 15 minutes and then spun down for 1 minute at 1,000rpm. The supernatant was then transferred to a new tube for each sample and stored at -80C until analyzed by SDS-PAGE.
Predicted LINE-1 ZnF146 binding motif- TGG AAT ACT ATG CAG CCA TAA AAA AGG ATG AG
Forward strand consensus- CATATAACCATGGAATACTATGCGCCACAATAAATGATGAGTT CATATCCTTGTAGG
Complement- CCTACAAAGGATAGAATCTCATTTTTTTATGGCTGCAATGATATTCATGCTGTA
Reverse strand consensus- CCTACAAAGGATAGAATCTCATTTTTTTATGGCTGCAATGATATTCATGCTGTA
Complement- ACACCATGGAATACTATGCGCCATAAATAATGATGAGTTATCATATCCTTGTAGG

Predicted LINE-1 ZnF507 binding motif- CAA ATT CAA ATT TAA ACA TAA TAA TAT TA
Forward strand consensus- CATCATAATGACAGGATCAAATTCACACATAACAATATTAACTTTAAATATAAATGGACT
Complement- AGTCCATTTATATTAAAGTTAATTTTATGCTGATCCGATTATGATG
Reverse strand consensus- AGTCCATTTATATTAAAGTTAATTTTATGCTGATCCGATTATGATG
Complement- ATAATGACAGGATCAAATTCACACATAACAATATTAACTTTAAATATAAATGGACT

Quantitation of Proteins by BCA assay

Protein concentrations for samples collected from protein purifications and nuclear extracts were quantified using a Pierce BCA assay kit. Protein absorbance levels were measured and compared to a series of BSA dilution standards.

Western Blots

The levels of ZNF proteins 146 and 507 were determined by Western blots. Protein samples were loaded onto pre-cast 26-well PAGE gels (4-20% gradient) (Criterion TGX). Samples were electrophoresed for 150 volts for roughly an hour, until the bromphenol blue bands had reached the bottom of the gel. The separated proteins were transferred onto DUDF membrane film by transblotting at 100amp electrophoresis for 2 hours using transfer buffer (10% methanol, 0.04% SDS). The membrane was then blocked by incubating with PBS+0.05% Tween+5% blotting-grade blocker for 1 hour. The membrane was then washed in PBS+0.05% Tween (PBST), and stored in fresh PBST at 4C.

The following primary antibodies were used: anti-rabbit ZNF146 (Novus), anti-rabbit ZNF507 (Thermo), anti-mouse Flag M2 (sigma). The membranes were trimmed around the areas of interest, and then incubated in primary antibody diluted 1:400 in PBST overnight at 4C.
on a shaker. The membranes were then washed three times in PBST for 5 minutes each, and then treated with HRP conjugated secondary antibody. The following antibodies were used; anti-mouse and ant-rabbit HRP (Abcam). The secondary antibody (for the specific animal of the primary) was diluted 1:5000 in 10mL of PBST+3% blotting-grade blocker. The membrane was incubated with the secondary antibody on the shaker for 1 hour, rinsed in PBST, and then washed three times in PBST. Washed membranes were treated with the buffers (Pierce ECL reagents A+B) of the protein imaging to stimulate chemiluminescece, and then imaged under high sensitivity parameters at increasing exposure times until the bands were sufficiently exposed. After imaging, the membranes were rinsed and stored in fresh PBST at 4C.

**Mass-Spec Sample Prep**

In order to detect other proteins interacting with the ZNFs, Flag-tagged ZNFs over-expressed in transfected cells were immuno-precipitated from cell lysates. The IP extracts were then electrophoresed on a gel, and stained with coomassie blue overnight to visualize the protein contents. The gel was then de-stained, and the lanes of interest were cut out and sent to the Mass Spectrometry Facility within UMass Medical School for analysis.

**Zinc Finger Predictions**

LINE-1 DNA sequences predicted to bind ZNFs 146 and 507 were determined using Princeton servers (http://compbio.cs.princeton.edu/zf/). Amino acid sequences for the ZNF proteins were obtained from the Uniprot Database and used as input for the server to generate predicted zinc finger sequences and their associated DNA-binding domains. The Princeton serves were also used to predict whether the predicted zinc fingers for the two proteins showed significant binding specificity to a 40nt region around the predicted binding motifs.
DNA and Protein Sequence Alignments

LINE-1 DNA sequence alignments were obtained through Clustal-Omega alignments of L1 subfamily sequences from the dfam and repbase databases (http://www.dfam.org/ & http://www.girinst.org/repbase/). These alignments were then imported into Jalview (http://www.jalview.org/) to generate conservation scores of the tentative binding region within the alignment. This process was performed to generate L1 alignments and determine the conservation of sequences for the binding region of each ZNF protein across several species.

Alignments of the amino acid sequences for the zinc fingers of ZNF 146 and 507 were obtained in the same manner using amino acid sequences found on Uniprot, and aligning the sequences using Clustal-Omega. These alignments were imported into Jalview to generate conservation scores. Phylogenetic trees were generated through importing the alignments into wasabi (http://wasabiapp.org/).
RESULTS

The retro-transposition of L1 elements has been found to result in a variety of genetic disorders, resulting from the induced gene disruptions, nucleotide deletions, duplications, and general chromosomal instability. Furthermore, the expression of L1 ORF2 itself is toxic and can induce senescence. Thus, the cellular repression of L1 transposition or expression is important, yet the understanding of the proteins involved, and mechanisms that contribute to, L1 repression are incomplete. Zinc finger proteins (ZNFs) bind both DNA sites and other regulatory proteins that affect gene expression, so this MQP sought to identify ZNFs that bind L1, and that could regulate L1 expression.

Identification of Potential L1-Binding ZNFs

We began by examining the ENCODE database for ChIP enrichment peaks of various zinc finger proteins against repetitive elements. The data (Figure 8) showed that ZNF146 and ZNF507 were strongly and uniquely enriched on L1 DNA relative to 153 other ZNF protein scrambled background peaks. Compared to the enrichment levels of 153 other ZNFs, ZNF146 had 12.1-fold enrichment, while ZNF507 had 7.6-fold enrichment for L1 elements. Both ZNFs had notably stronger enrichment than any other ZNF on L1.
Figure 8. ChIP Enrichment of Zinc Finger Proteins on Various Repetitive Elements. The analysis of ENCODE ChIP data sets showed that ZNF146 and ZNF507 (black circle) have much stronger enrichment levels on L1 elements compared to 153 other ZNFs. The two proteins also showed no similar enrichment level across nine other common repetitive elements.

Further refinement of the ZNF146 and ZNF507 enrichment peaks (Figure 9) was performed to determine which types of repetitive elements the ZNFs bind. The data showed that 81% of ZNF146 peaks occurred in annotated L1 sequences, while 78% of the repetitive elements bound to ZNF507 peaks were in L1 sequences. ZNF146 had roughly 37,000 ChIP peaks on L1, while ZNF507 had roughly 12,000 (Figure 9). However curiously, despite being enriched on the same element, only 116 of the peaks for the two ZNF proteins overlapped.

Figure 9. ChIP Enrichment Binding of ZNF146 and ZNF507 to Various Repetitive Elements. Note that 81% of the repetitive elements bound to ZNF146 were to L1, while 78% of the elements bound to ZNF507 were to L1. Overlap of the two analyses occurred in only 116 of the 48,712 identified peaks.
The enrichment peaks for the two ZNF proteins on L1 were then further analyzed to identify the potential binding sites on the repetitive elements. ZNF146 enrichment peaks strongly localized to the 3’ end of L1 ORF2 (blue in the diagram), while ZNF507 enrichment peaks (red in diagram) localized to the 5’ end of L1 ORF2 (Figure 10). Homer prediction software was then used to construct a binding consensus for each ZNF (Figure 10). The binding consensus for the most recent subfamily of L1 elements were found to be GAATACTATGCAG and TAAATATAAATG, for ZNF146 and ZNF507, respectively.

![Figure 10. Potential Binding Site Locations of ZNF146 and ZNF507 on L1 DNA Elements. Shown is the ChIP enrichment data for ZNF146 and ZNF507 for the ORF2 of the most recent subfamily of L1 DNA elements.](image)

**Evolutionary Conservation of the Predicted Binding Domains**

The conservation of the predicted binding regions for the two ZNF proteins on L1 sequences was characterized through a clustal alignment of the L1 subfamily sequences. The analysis included an area approximately 10nt surrounding the predicted binding motif. These domains were imported into Jalview to generate a conservation score and arrange the subfamilies by age. The conservation of ZNF-binding sequences was determined for several L1 DNA
families, whose nomenclature is as follows: PA denotes primate-specific L1 sequences, and MA denotes mammalian. PA L1 sequences diverged from mammalian L1’s at MA6. MA sequences, however, continued to evolve within the human genome. The sequences are also divided into sub-families. Evolutionarily the youngest (most recent) L1 sub-families have the lowest numbers, so L1 PA1 denotes modern human L1, and PA4 is an older sub-family. The oldest sequences L1 sub-families within the human genome are named in descending order, with MA being the youngest, and ME being the oldest. Due to a 5’ truncation, the L1 sub-families are divided into two categories based on which end of the sequence was analyzed. However, these sequences are synonymous with one another, so 5’ L1 P1 is equivalent to 3’ L1 PA2-PA3 (Figure 4).

The binding region of ZNF146 (Figure 11) was found to be highly conserved for primate L1 subfamilies, only encountering a T-C substitution at PA15. A comparison of primate to mammalian L1 revealed a prevalent G-C substitution in primates within the binding motif when compared to present mammals, among several other substitutions (Figure 11). Furthermore, the area surrounding the predicted binding motif showed similar high levels of conservation across both primate and mammalian sub-families.
Figure 11. Conservation of the Predicted Binding Regions for ZNF146 Across L1 Subfamilies. DNA sequences within 10 bp surrounding the predicted ZNF146-binding motifs of various L1 elements were analyzed for conservation using a clustal alignment. These domains were imported into Jalview to generate a conservation score and arrange the subfamilies by age. The sequences represent, from top to bottom: L1PAs (primate specific L1 subfamilies, including human), L1Mas (mammalian specific L1 subfamilies, including human), L1MM (youngest mouse L1 subfamily). The colors denote specific nucleic acids: red (guanine), yellow (cytosine), blue (thymine), and green (adenine). Marked in red above the alignment and below the conservation scores is the predicted ZNF binding motif.

The data for the conservation of the binding region for ZNF507 (Figure 12) shows an A-to-G substitution in primates at the P4 and PA13-17 sub-families. Interestingly, the youngest subfamily PA1, showed a specific G-A substitution in the middle of a palindromic stretch of the sequence. Apart from these differences, the predicted binding motif was found to be highly conserved across all primate subfamilies tested, but was less conserved between mammalian and primate subfamilies.
Figure 12. Conservation of the Predicted Binding Regions for ZNF507 Across L1 Subfamilies. DNA sequences within approximately 10 bp surrounding the predicted ZNF507-binding motifs of various L1 subfamilies were analyzed for conservation using a clustal alignment. These domains were imported into Jalview to generate a conservation score and arrange the subfamilies by age. Due to a truncation the nomenclature is altered from the previous ZNF146 figure as follows: P1 is equivalent to PA2-PA3; P2 is equivalent to PA4-PA6, and so on as illustrated in Figure 4. The sequence codes are as indicated in the previous figure, except for the following: n (any nucleotide), r (adenine or guanine), and m (cytosine or adenine).

A second alignment of the predicted binding regions for the two ZNFs on L1 across several species using the same programs implemented previously (Figure 13) revealed similar data: a high conservation of the ZNF146 and ZNF507-binding sequences, and lower levels of conservation flanking the binding sequences. The blue bars within these figures denote the predicted binding regions of the ZNFs on L1 as predicted by the Princeton software as of December 2017.
Figure 13. A Second Analysis of the Species Conservation of the Predicted Binding Regions for ZNF146 and ZNF507 on L1 Sequences. The alignment was performed as described in the previous figure, and the sequence codes are the same except for the following: the blue bar above the sequences denotes a predicted binding site found by the Princeton program, and the black bar denotes the oligo used during our pull-down procedures.

Evolutionary Conservation of the Zinc Finger Domain Sequences

The zinc finger domain sequences of both ZNF146 and ZNF507 were predicted using Princeton prediction software, and aligned in the same manner as the previous figures to examine the evolutionary conservation of the binding regions of the proteins. The various species selected for the alignment contained both ZNF146 and ZNF507 amino acid sequence data. The phylogenetic trees shown to the left of each figure panel denote the general evolutionary divergence of the ZNF proteins between the species. The zinc finger sequences for ZNF146 (Figure 14) were found to be completely conserved in 7 out of 10 zinc fingers, with 2 showing only a single amino acid substitution in a single species. All the substitutions were within the same class of amino acids.
The alignment for ZNF507 zinc fingers (Figure 15) showed less conservation. While most amino acid substitutions were within the same amino acid group, two of the fingers showed species-specific substitutions that were not within the same group of amino acids. The second zinc finger of ZNF507 had an E (Glu) (charged) to Q (Gln) (neutral) substitution in mice and rats. However, primates showed a different N (Asn) to H (His) substitution in the fifth zinc finger that was highly conserved and was not seen in any other species tested (Figure 15).
Present-Day Conservation of L1 ZNF Binding Sites Using ChIP-Seq Enrichment

The conservation of ZNF 146 and 507 binding retentions was analyzed across several present-day L1 subfamilies through ChIP-seq dataset mapping. ZNF146 was found to have conserved binding locations across all L1 primate subfamilies (Figure 16, top panels). The reduced binding signal observed for subfamily PA1-PA4 likely resulted from the low unique mapability of these L1 subfamily sequences (Figure 16, bottom panels). On the other hand, the L1 binding capacity for ZNF507 was lost from subfamily P4/PA10 onward (Figure 17). This loss was also seen in the nucleotide alignments as the A-T nucleotide substitution in the binding motif (Figure 12). Thus, this substitution may contribute to ZNF507 acquiring the ability to bind L1 repeats.

Figure 16. Present-Day Conservation of L1 Binding Sites for ZNF146 Using ChIP-Seq Enrichment.
Verification of ZNF Binding Sites Using Antibody Pull-Down Assays with Biotinylated DNA Oligos

The binding of ZNF146 and ZNF507 to their respective binding consensus was verified biochemically using pull-down assays (Figure 18). Biotinylated DNA oligos representing each binding consensus were mixed with HeLa cell nuclear lysates containing ZNFs. Biotinylated oligo-ZNF complexes were purified by ZNF-binding precipitations using oligo bound beads containing the specific binding site for each ZNF. Then the intensity of the ZNF protein concentration was assessed by a western blot using antibodies against the specific ZNFs. When ZNF proteins were purified using the predicted binding site oligos bound to beads, the pulldown of ZNF146 showed a strong pulldown efficiency. In the same sample, a much weaker signal was observed when pulling-down with the ZNF507 binding consensus, and no signal was observed for the ZNF507-purified sample. Lowering the salinity or adding Zn^{2+} into the washes yielded similar results (data not shown). Thus, ZNF146 was able to bind the predicted ZNF146
consensus. No signals were observed for the sample immuno-precipitated with the ZNF507 antibody, so the binding of ZNF507 to its consensus has not been validated.

Figure 18. Biochemical Verification of the Binding of ZNF146 to its Predicted L1 Consensus.
Biotinylated DNA oligos representing the predicted ZNF146 and ZNF507 binding sites were mixed with nuclear lysates containing ZNFs, and the samples were immuno-precipitated using streptavidin coated beads bound to oligos containing for either the ZNF146 (upper panel) or ZNF507 (lower panel) binding motif. The samples were electrophoresed, blotted to membrane, and the signal of biotinylated oligo determined by exposure to a secondary antibody to ZN146 or ZNF507. Lanes denote the type of biotinylated oligo used in each pulldown. Poly TA was used as a negative control. Further testing was done using flag-tagged ZNF transfections, which were assessed using mouse anti-flag and anti-mouse HRP antibodies. Alterations in the salinity content and adding Zn²⁺ during washes yielded identical results. Antibody designations on the right denote the antibody used for detection of ZNF146 or ZNF507 in the western blot of purified samples.

Cellular Localization of ZNF146 and ZNF507

The cellular locations of ZNF146 and ZNF507 were determined by immunofluorescence microscopy (IF) on Flag-tagged ZNFs. HeLa cells were transfected with plasmids encoding Flag-tagged ZNFs, and IF was performed at 48-hours post-transfection using anti-Flag antibodies (Figure 19). The data showed that both ZNF146 and ZNF407 localized to the nucleus in HeLa cells, as expected for DNA binding proteins. The ZNF proteins were excluded from the nucleolus. Slight cytoplasmic localization was also observed in ZNF507 transfected cells. These findings were confirmed in both HEK293 and TIG-1 cells (data not shown).
Figure 19. Cellular Localization of ZNF146 and ZNF507. The locations of flag-tagged ZNFs 146 and 507 in transfected HeLa cells were detected using anti-flag antibodies. Both ZNFs show a nuclear localization, and nucleolar exclusion, as expected for DNA-binding proteins. Green denotes flag-tag signal (ZNF146 or ZNF507), blue denotes DAPI staining for nuclear morphology.

We noticed that some of the transfected HeLa cells showed signs of toxicity, so this was tested further (Figure 20). Stable flag-tagged ZNF over-expression lines were generated, but after a 3-week period only a few HeLa cells expressing the flag-tagged ZNF proteins remained (Figure 20), suggesting a selection against cells over-expressing these ZNFs. The few cells that still over-expressed the ZNFs at 3-weeks (as tested by IF) showed a very unhealthy nuclear morphology. Thus, over-expression of the ZNFs long-term may have a deleterious effect on HeLa cells.
Over-Expression of ZNFs in HEK293 Show Possible Cell Phase Arrest

Transfection of the ZNF expression plasmids into HEK293 cells (Figure 21) showed nuclei that appear to be enlarged relative to non-transfected cells. This phenotype may result from blockage at the G2 cell phase, so Ki-67 IFA staining was performed (Figure 21). This analysis showed that the ZNFs (detected by flag antibody, green), were found exclusively in Ki-67 G2 cells. These results suggest that over-expression of the ZNFs could result in cell phase arrest at G2, thereby causing the ZNF-containing cells to be outcompeted, as seen in the long term HeLa cell transfections.
Effect of ZNF Knockdown in HEK Cells

The potential effects of lowering the expression of ZNF146 and ZNF507 in HEK293 cells were studied using knockdowns induced by sh-RNA expressing plasmids (Figure 22). Although we have not yet verified a lowering of the ZNF signal in these sh-plasmid-treated cells, the data showed that knockdown of either ZNF protein appears to cause two distinct types of cell senescence. The nuclei of HEK293 cells treated with ZNF146 sh-plasmid appeared to have significantly enlarged nuclei and a severely weakened DAPI signal (Figure 22). Additional detection by laminin B antibody revealed that the nuclear boundary shown by DAPI staining was indeed accurate. The cytoplasm also appears to have enlarged as seen by tubulin detection (Figure 22). This phenotype is indicative of induced cell senescence (Leontieva et al., 2011; Micco et al., 2011).

On the other hand, knockdown of ZNF507 appeared to result in a different type of cell senescence. In this case, the nuclei retained a relatively normal size and shape, but the DAPI detection took on a dotted, almost speckled appearance, known as compacted punctate DAPI (Figure 22). Punctuated DAPI appearance is known to commonly occur in senescent cells (Zhang et al., 2007).
Identification of ZNF-Associated Proteins

As discussed in the Background, ZNF proteins often bind other proteins that regulate gene expression. In these cases, the ZNF confers DNA-binding ability to the regulator. In order to identify proteins that interact with ZNF146 and ZNF507 (and that could affect L1 gene expression), HEK293 cells over-expressing flag-tagged ZNF146 or ZNF507 were lysed, and the ZNF and any associated proteins were purified by immunoprecipitation (IP) with anti-flag antibody (Figure 23). The initial IP experiment (Figure 23A) confirmed that flag-tagged ZNF507 could in fact be purified from a nuclear extract. The subsequent IP (Figure 23B) showed that flag-tagged ZNF146 and ZNF507 could be isolated.
Figure 23. Immunoprecipitation of Flag-tagged ZNF507 and Associated Proteins. HEK293 cells were transfected with plasmid encoding flag-tagged ZNF507 or 146, lysed, and then nuclear extracts were prepared. Nuclear extracts were immuno-precipitated with anti-flag antibody to pull down ZNF507 or ZNF146 and any associated proteins. A) The initial IP experiment analyzed the ZNF507 IP sample relative to input sample to confirm successful purification compared to input. B) Analysis of the ZNF507 and ZNF146 IP samples to verify successful purification of the ZNFs.

Potential ZNF-interacting proteins were identified from the purified IP samples using mass spectrometry (Figure 24). Purified IP samples were briefly run through a gel, stained, cut out, and then sent for mass spectrometry analysis at UMMS. The amount of protein isolated from the ZNF146 sample (middle lane in the figure) was insufficient for analysis, but the ZNF507 IP sample (right lane) contained a robust ZNF507 band and faint other lower MW bands (Figure 24). Mass spectrometry results of the ZNF507 IP sample (Table 1) revealed several proteins with gene regulatory function, including PRMT5, PPM1B, WDR77, and two RBBP proteins.
Figure 24. Isolation of Protein Present in the ZNF507 IP Sample Sent for Mass Spec Analysis. The over-expressed ZNF507 band (red circle) was the most prevalent band in the 507 IP sample (right lane), and other lower-MW proteins are also present. The middle lane between the ladder and ZNF507 shows the ZNF146 IP sample, with a faint protein band near the bottom of the gel likely representing 146.

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>MW (kDa)</th>
<th>Peptide Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ZNF507</td>
<td>106</td>
<td>100% (283)</td>
</tr>
<tr>
<td>2</td>
<td>PRMT5</td>
<td>73</td>
<td>100% (45)</td>
</tr>
<tr>
<td>3</td>
<td>VCP</td>
<td>89</td>
<td>100% (38)</td>
</tr>
<tr>
<td>4</td>
<td>PPM1B</td>
<td>53</td>
<td>100% (35)</td>
</tr>
<tr>
<td>5</td>
<td>PRKDC</td>
<td>469</td>
<td>100% (33)</td>
</tr>
<tr>
<td>6</td>
<td>WDR77</td>
<td>37</td>
<td>100% (16)</td>
</tr>
<tr>
<td>7</td>
<td>YWHAE</td>
<td>29</td>
<td>100% (23)</td>
</tr>
<tr>
<td>8</td>
<td>EEF1A1</td>
<td>50</td>
<td>100% (22)</td>
</tr>
<tr>
<td>9</td>
<td>ELF4B</td>
<td>69</td>
<td>100% (20)</td>
</tr>
<tr>
<td>10</td>
<td>ELF3B</td>
<td>99</td>
<td>100% (16)</td>
</tr>
<tr>
<td>11</td>
<td>RRM4E</td>
<td>103</td>
<td>100% (20)</td>
</tr>
<tr>
<td>12</td>
<td>EEF1G</td>
<td>50</td>
<td>100% (19)</td>
</tr>
<tr>
<td>13</td>
<td>STK88</td>
<td>54</td>
<td>100% (21)</td>
</tr>
<tr>
<td>14</td>
<td>SLC25A5</td>
<td>33</td>
<td>100% (22)</td>
</tr>
<tr>
<td>15</td>
<td>GAPDH</td>
<td>36</td>
<td>100% (18)</td>
</tr>
<tr>
<td>16</td>
<td>ROCK1</td>
<td>158</td>
<td>100% (19)</td>
</tr>
<tr>
<td>17</td>
<td>RBBP4</td>
<td>46</td>
<td>100% (14)</td>
</tr>
<tr>
<td>18</td>
<td>RBBP7</td>
<td>48</td>
<td>100% (9)</td>
</tr>
<tr>
<td>19</td>
<td>TGM3</td>
<td>77</td>
<td>100% (15)</td>
</tr>
<tr>
<td>20</td>
<td>KCTD17</td>
<td>33</td>
<td>100% (16)</td>
</tr>
</tbody>
</table>

Table 1. Mass Spec Analysis of the ZNF507 IP Sample. The list shows proteins identified by mass spec analysis of the ZNF507-IP sample. Shown in yellow is ZNF507, and in green are several regulatory proteins of interest.
Non-coding DNA sequence elements make up roughly 98% of the human genome. LINE-1 (L1) repetitive elements are currently thought to make up 17.5% of the genome, excluding those found in intronic sequences. Despite comprising such a large proportion of the genome, the function and regulation of L1 elements remains unknown. Studies have shown that L1 expression induces retrotransposition and can be toxic to cells. L1 ORF2 contains the main machinery for retro-transposition. Due to the innate danger that large-scale L1 transposition would have on nearby genes, this process is normally strongly repressed in somatic cells. However, L1 elements sometimes escape repression, giving rise to new L1 subfamilies that require new a means of repression. This cycle of L1 expression, retrotransposition, and silencing can be thought of as an evolutionary arms race between the emergence of new L1 subfamilies and the adaptation of new repressive machinery.

Zinc finger proteins (ZNFs) bind DNA, and their direct binding to regulatory proteins confers sequence-specific recruitment to DNA. This MQP project investigated whether ZNFs bound to L1 elements might provide a mechanism for L1 repression or activation. A bioinformatics approach showed that ZNF146 and ZNF507 proteins potentially have strong binding affinity to L1 elements at opposite ends of ORF2, and this binding was verified by oligo pull-down assays. The ZNF146 and ZNF507 protein zinc finger sequences and the L1 binding regions showed strong evolutionary conservation. Furthermore, ZNF binding across L1 subfamilies was also found to be strongly retained for ZNF146. On the other hand, ZNF507 showed a specific period of binding loss past a certain subfamily age that appeared to be linked to a specific point mutation in the predicted L1 binding region. Immunoprecipitation of
ZNF507-protein complexes followed by Mass spectrometry identified several ZNF507 interacting regulatory proteins (purple in the figure) that potentially use ZNF507 as a guide for targeted L1 repression (Figure 24). PRMT5, the protein found with the highest peptide count, is a histone arginine methyltransferase. Histone arginine methylation is often associated with gene repression (Girardot et al., 2014; Litt et al., 2009).

The effects of over-expression and under-expression of the two ZNFs was also studied. Over-expression of either of the two ZNF proteins appeared to be toxic to cells. Such cells displayed Ki-67 patterns seen only in the G2 cell phase (Figure 21) (Solovjeva et al., 2012), so these cells may be arrested in the G2 phase due to ZNF over-expression. If the two ZNF proteins in fact bind strongly to L1, which constitutes 17.5% of the human genome, the binding of the overexpressed proteins to additional targets may mis-regulate the genome or specific cell cycle genes. Under-expression on the other hand led to DAPI stain patterns typical of cells in senescence (Figure 22) (Micco et al., 2011). Given that expression of the L1 ORF2 alone has been shown to induce cell senescence, this finding suggests that the ZNFs may help silence L1 ORF2, and the ZNF knockdown allowed ORF2 expression.
These studies would have benefitted from validation of ORF2 expression levels under different ZNF expression conditions. Unfortunately, as of the writing of this report, the only commercially available antibody against ORF2 showed no signal, even when tested against a cell line over-expressing ORF2. Because L1 has approximately 950,000 fragments in the genome, assays such as Northern blots or RT-PCR would have yielded too many false positives to be conclusive. Without a confident means for accurately detecting full length L1 or ORF2 transcription, a direct correlation between the presence of the two ZNF proteins and a lowered L1 gene expression could not be established. Future studies would hope to characterize an indirect correlation between ZNF 146 and 507 levels and L1 transcription.

We have verified a specific binding of ZNF146 to the predicted L1 DNA consensus domain in vitro. The unsuccessful oligo pulldown of the ZNF507 binding site was likely hindered by our lack of knowledge of the ZNF507 tertiary structure. The predicted binding motif was found by analyzing its strong stretch of five consecutive zinc fingers, while the other non-contiguous zinc fingers are separated further away on the protein and could play a role in further binding upstream on L1. The lack of inclusion of this upstream consensus sequence in our biotinylated oligo could have resulted in unsuccessful ZNF507 pulldown. Alternatively, a key cofactor could have been lost which was required for ZNF507 binding.

With respect to the Mass spec analysis of ZNF-associated proteins, the approach worked well with ZNF507 (to identify regulatory protein PRMT5), but not with ZNF146. This may simply have resulted from an insufficient amount of protein in the ZNF146 IP sample, possibly due to elevated cell toxicity, so this could be repeated using larger amounts of input material or using optimized conditions. If some of the ZNF146-binding proteins are found to be regulatory in nature, combined with the findings documented in this report, this would support the theory
that both ZNF146 and ZNF507 play a role in L1 repression in somatic cells. Very few proteins have been identified as binding to transposable elements and regulating their expression. This MQP study goes a long way towards characterizing two ZNF proteins with relatively no known function as L1 regulatory proteins. These proteins have been implicated as misregulated in cancer and neurological disorders, and these findings may impact the understanding of disease.
BIBLIOGRAPHY


