Molecular Mechanisms of LINE-1 Retrotransposition Inhibition

by

Peter A. Larson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in the University of Michigan
2017

Doctoral Committee:

Professor John V. Moran, Chair
Associate Professor Scott Barolo
Associate Professor Sundeep Kalantry
Assistant Professor Jeffrey M. Kidd
Professor Miriam H. Meisler

Peter A. Larson

peterlar@umich.edu

ORCID iD: 0000-0002-1490-150X

To the Larson family.

Acknowledgements

First and foremost I would like to thank my mentor, Dr. John V. Moran. John's guidance and leadership were indispensible to my development as a competent scientist. He challenged me to develop hypotheses and rigorously test them. John's enthusiasm and persistence is passed on to his students, and his ability to effectively push just beyond my limits was crucial for me to realize my full potential.

I also thank my committee members, Dr. Miriam Meisler, Dr. Scott Barolo, Dr. Sundeep Kalantry, and Dr. Jeffrey Kidd. My interactions with them were always positive. They continually challenged me and demanded rigor, intellectual reasoning, and consistency. I am indebted to the seriousness in which they took their roles in helping me to become and effective and thoughtful experimentalist.

I would also like to thank all of the members of the Moran laboratory, both past and present. Their drive, enthusiasm, and supportive attitude helped me persevere. They created an enjoyable and positive lab environment. I am grateful for their continued support and genuine interest in helping me succeed. I would like to specifically thank Ms. Madeleine VandenBrink. She was an illuminating undergraduate that I had the pleasure of mentoring and she made the final year and a half of my graduate career more enjoyable, and productive.

Finally, I would like to thank all of my family and my friends. Their unwavering support was essential and I cannot stress enough how much they mean to me. It really does take a village to earn a Ph.D, and I am indebted to the city that helped me earn mine.

Table of Contents

List of Figures

List of Appendices

Abstract

Molecular mechanisms of LINE-1 retrotransposition inhibition

By

Peter A. Larson

Chair: John. V. Moran

Long INterspersed Element-1 (LINE-1 or L1) retrotransposons are an ancient family of repeated DNA sequences present in all inspected mammalian genomes. L1s are the only active autonomous mobile element in the human genome and are present at over 500,000 copies, representing ~17% of genomic DNA. A full-length human L1 is ~6 kb in length and contains an internal RNA polymerase II promoter within its 5' untranslated region (UTR). Following the 5'UTR are two open reading frames (ORF1 and ORF2) that encode two functional proteins (ORF1p and ORF2p) that are required for mobilization (*i.e.*, retrotransposition). L1s end with a 3'UTR and a poly(A) tail. The evolutionary success of L1 relies on the reiterative retrotransposition of full-length L1 RNAs. The vast majority (>99.9%) of genomic L1s are inactive; however, on average, ~80-100 L1s per diploid genome are capable of retrotransposition. Since, L1 retrotransposition by its nature is mutagenic, it is likely that cellular host-factors have evolved to inhibit or restrict unabated L1 retrotransposition.

Previous studies identified functional splice donor, splice acceptor, and polyadenylation sequences in full-length L1 RNA. Here, I demonstrate that retrotransposition of intra-5'UTR or 5'UTR/ORF1 spliced L1 RNAs leads to the generation of Spliced Integrated Retrotransposed Elements (SpIREs). Additionally, I uncovered a new intra-5'UTR SpIRE that is approximately ten times more abundant than previously identified SpIREs. Using biochemical and genetic approaches, I definitively demonstrate that intra-5'UTR SpIREs lack *cis*-acting transcription factor

binding sites, resulting in reduced 5'UTR promoter activity compared to a full-length 5'UTR. Moreover, I demonstrate that 5'UTR/ORF1 SpIREs lack *cis*-acting sequences required for L1 transcription and produce non-functional ORF1p variants. These results establish that SpIREs are evolutionary "dead ends," which are unlikely to contribute to additional rounds of L1 retrotransposition. Finally, in agreement with previous publications, I demonstrate that a subset of splicing factors may repress L1 expression and/or retrotransposition.

Previous experiments in embryonic human carcinoma-derived cells (hECs) revealed that an engineered L1 tagged with a retrotransposition indicator cassette (L1-reporter) successfully retrotransposed into hEC genomic DNA. However, either during or immediately after genomic integration, expression of the L1-reporter is silenced in hECs. Previous experiments demonstrated that hEC cells treated with histone deacetylase (HDAC) inhibitors swiftly reverse L1-reporter silencing. I sought to identify host proteins that may be involved in L1-reporter silencing by developing a forward genetic screen using CRISPR/Cas9-based genome editing technology. Using the PA-1 hEC cell line that permits L1-reporter retrotransposition, but subsequently silences L1-reporter expression, I identified potential candidate genes that may play a role in L1-reporter mediated gene silencing. Future work will test whether the candidate genes identified in this screen directly or indirectly inhibit L1-reporter mediated gene silencing.

A continued understanding of the interplay between L1 and host-factors is critical to understanding human genomic variation. Additionally, as the L1-host interaction largely recapitulates a traditional host-parasite "arms-race," new insights gained from these studies may contribute to our understanding of how L1 retrotransposition continues to influence human health and disease.

Chapter 1
Introduction

Thesis Overview

This thesis asks the question, “What cellular mechanisms inhibit unabated human L1 retrotransposition?” Herein, I describe my findings within the context of an evolutionary “arms-race” that occurs between L1 and its host, human cells. Chapter one begins with a general review of transposable element biology, and then focuses on human L1 biology, placing a particular emphasis on L1 evolution and the interaction of L1 with inhibitory host proteins. Chapter two describes the identification and characterization of Spliced Integrated Retrotransposed Elements (SpIREs) and demonstrates how splicing of L1 RNA potently inhibits retrotransposition via disruptions of the 5’UTR promoter sequence and/or translation of full-length ORF1p. Chapter three details my progress toward developing an unbiased forward genetic screen using CRISPR/Cas9-mediated genome editing to identify candidate cellular factors that may contribute to L1-reporter gene silencing. Chapter four provides a summary of my findings as they pertain to the hypothesis that L1 continues to amplify in the human genome despite the presence of a “host-parasite arms-race” within our cells, and discusses areas for future study.

An Introduction to Transposable Elements

Transposable elements (TEs), sometimes called “jumping genes,” are genomic DNA sequences that can physically move to new genomic locations within their host genome. TEs are present in all studied eukaryotic organisms and represent a diverse group of sequences that share a distinct feature, the potential for genomic mobilization (Craig et al., 2015; Fedoroff, 2012; Feschotte and Pritham, 2007). TEs utilize a variety of mechanisms to mobilize within genomes (Levin and Moran, 2011). Ultimately, the ability of any given TE to mobilize is dependent on the cellular environment in which it resides, an environment that can change over large spans of evolutionary time (millions of years (MY)), as well as within the lifetime of the organism (Levin and Moran, 2011).

Barbara McClintock’s discovery that “mutable-loci” in maize are the result of mobile DNA sequence activity demonstrated that the genome was not simply a static entity (McClintock, 1950, 1951). Classically considered “junk DNA,” TEs were assumed to be repeat DNA sequences that did not contribute phenotypically to the organism (*i.e.*, “contributing little or no functional significance”) (Ohno, 1972). Though still sometimes referred to as “junk DNA,” the status of TEs was upgraded slightly by a pair of papers published in 1980 that referred to TEs as “selfish DNA” (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). The two tenets of the “selfish DNA” hypothesis, as outlined by Orgel and Crick, are: 1) “It arises when a DNA sequence spreads by forming additional copies of itself within a genome,” and 2) “It makes no specific contribution to the phenotype.” By strict definition, this second point importantly means that, unlike protein coding genes, TE activity is not evolutionary selected to contribute phenotypically to the organism.

Investigations into the mechanism of TE mobilization have yielded a tremendous amount of biological information since Barbara McClintock’s initial studies. In the context of human genetics, one of the most striking findings was the discovery of actively amplifying (*i.e.*, retrotransposing) human Long INterspersed Element 1 (LINE-1 or L1) sequences (Kazazian et al., 1988). Since that discovery, investigations into L1 biology have yielded important insights into how L1s contribute to intra- and inter-

individual human genetic variation and human disease (Beck et al., 2010; Cordaux and Batzer, 2009; Hancks et al., 2011; Richardson et al., 2015).

The remainder of this introductory chapter focuses on the two main classes of transposable elements [Class I (retrotransposons) and Class II (DNA Transposons)]. It then focuses on a review of human L1 biology, placing a particular emphasis on L1 evolution and the interaction between L1 and its human host.

The Human Genome is Replete with Transposable Elements

The protein-coding portion of the human genome constitutes a mere 1.5% (Consortium, 2012; Lander et al., 2001) of genomic DNA and may encode for as few as ~19,000 genes (Ezkurdia et al., 2014) (Figure 1.1). Surprisingly, at least 45% of genomic DNA is derived from TEs (Lander et al., 2001). Indeed, bioinformatics analyses performed during the last six years suggest that as much as 70% of human genomic DNA may be derived from TEs (de Koning et al., 2011) (Figure 1.1). TE-derived sequences are separated into two classes: Class I retrotransposons mobilize via an RNA-intermediate using a “copy and paste” mechanism called retrotransposition. Class II DNA transposons generally mobilize via a DNA intermediate using a “cut and paste” mechanism called transposition.

DNA Transposons

Class II DNA transposons are among the oldest type of recognizable transposable element, as they are present in prokaryotes (Kleckner, 1981), and all five eukaryote super groups (Feschotte and Pritham, 2007; Krupovic and Koonin, 2015). Although still active in many non-mammalian organisms, as a whole, DNA transposons have been less successful and contribute less genomic DNA content than other TEs (Feschotte and Pritham, 2007). DNA transposons comprise ~3% of the human genome (Lander et al., 2001) (Figure 1.1), but have been inactive in the lineage leading to humans for the past 50 MY. Indeed, it was thought that DNA transposons had become inactive in all mammalian lineages, as no evidence of their activity was present in recently sequenced genomes of other mammals (Feschotte and Pritham, 2007). Thus, it was surprising that the genome of the little brown bat, *Myotis lucifigus,* harbors active DNA transposons (Ray et al., 2008; Ray et al., 2007).

Autonomous DNA transposons primarily move via a “cut and paste” mechanism (*i.e.*, transposition) that is mediated by their encoded transposase gene (Craig et al., 2015; Feschotte and Pritham, 2007). In the genome, DNA transposons are typically flanked by terminal inverted repeats (TIRs), which act as recognition sequences for transposase binding (Craig et al., 2015; Munoz-Lopez and Garcia-Perez, 2010) (Figure 1.2). Transposase mediates both the excision of the DNA transposon from its initial genomic location as well as its integration into a new genomic target site (Craig et al., 2015). The process of DNA transposition generally results in the generation of short 4-6 bp target-site-duplications (TSDs) that flank the TIRs (Figure 1.2) (Craig et al., 2015; Munoz-Lopez and Garcia-Perez, 2010).

The counterparts to autonomous DNA transposons are non-autonomous DNA transposons. Non-autonomous DNA transposons contain TIRs, but do not encode a functional transposase gene (Craig et al., 2015; Munoz-Lopez and Garcia-Perez, 2010; Slotkin and Martienssen, 2007). Instead, non-autonomous DNA transposition is dependent on ‘hijacking’ (*trans*-complementation) the transposase protein expressed from an autonomous DNA transposon (Craig et al., 2015; Munoz-Lopez and Garcia-Perez, 2010; Slotkin and Martienssen, 2007). Indeed, the TIRs are critical for the transposition of both autonomous, and non-autonomous DNA transposons (Craig et al., 2015; Slotkin and Martienssen, 2007). Thus, the interaction between autonomous and non-autonomous TEs necessarily results in competition for recruitment of the catalytic transposase protein that is responsible for TE mobility. Strikingly, Barbara McClintock’s discovery of transposable elements in maize included both the autonomous *Activator (Ac)* DNA transposon, as well as its non-autonomous partner, *Dissociation (Ds)* (McClintock, 1950).

By strict definition, “cut and paste” DNA transposition is non-replicative. Thus, DNA transposon copy number expansion in a given genome often depends upon genomic duplications that encompass an already resident DNA transposon (Craig et al., 2015). Interestingly, it has been shown that, on occasion, the transposase genes of DNA transposons have been “co-opted” by the host for functional purposes (Slotkin and Martienssen, 2007). For example the V(D)J recombination proteins RAG1 and RAG2 were likely domesticated from an ancient transposon (Huang et al., 2016; Kapitonov

and Jurka, 2005; Lander et al., 2001). RAG1 and RAG2 initiate V(D)J recombination allowing B- and T-cells generation of a diverse array of antigen binding receptors (Notarangelo et al., 2016). V(D)J recombination is considered to be a primary driving force in the evolution of adaptive immunity (Notarangelo et al., 2016). Likewise, the acquisition of adaptive immunity is considered to be a primary driving force in the evolution of jawed vertebrates (Huang et al., 2016; Notarangelo et al., 2016). Research suggests that the centromeric binding protein CENP-B is also a domesticated DNA transposon (Casola et al., 2008). CENP-B binds to a repeated centromere sequence and functions to organize the centromere in interphase as well as in mitotic chromosomes (Ando et al., 2002; Casola et al., 2008; McKinley and Cheeseman, 2016).

Experimentalists have co-opted DNA transposons and their encoded transposase genes as tools for use in molecular biology. DNA transposon-based mutagenesis screens have been instrumental in uncovering new cancer genes and have been used as vector systems for gene therapy (Dupuy et al., 2006; Ivics et al., 1997; Kawakami, 2005, 2007; Klinakis et al., 2000; Munoz-Lopez and Garcia-Perez, 2010). For example the *Sleeping Beauty* (SB) DNA transposon has been engineered as an efficient gene delivery system in a variety of cell types (Ivics et al., 1997; Munoz-Lopez and Garcia-Perez, 2010). The SB system relies on a two-plasmid delivery mechanism. One plasmid (*i.e.*, the transposon vector) contains the DNA sequence of the gene of interest, flanked by TIRs, that is to be transposed into genomic DNA (Munoz-Lopez and Garcia-Perez, 2010). The other plasmid (*i.e.*, the transposase expression vector) contains the SB transposase gene, but lacks TIRs (Munoz-Lopez and Garcia-Perez, 2010). The SB transposase acts in *trans* to bind the TIRs of the transposon vector to mediate its integration into a genomic DNA target site (Munoz-Lopez and Garcia-Perez, 2010). Thus, transposase can mobilize a non-autonomous element to a new genomic location as long as it contains TIRs.

Other transposon mutagenesis and/or gene delivery systems have been developed using a similar logic to the one outlined above. For example, the *piggyBac* DNA transposon was first identified in insects and *piggyBac* insertions tend to favor transcriptional units, thus making it ideal for gene mutagenesis studies (Cary et al., 1989; Munoz-Lopez and Garcia-Perez, 2010). Similarly, the *Tol2* DNA transposon, first

identified in fish, is capable of efficiently transposing over 10kb of DNA, making it ideal for delivering large transgenes (Kawakami, 2007; Munoz-Lopez and Garcia-Perez, 2010).

Recently, a DNA capture and deep sequencing technique (ATAC-Seq) was developed using the bacteria derived *Tn5* DNA transposon (Buenrostro et al., 2013). The *Tn5* transposase preferentially targets regions of open chromatin (Buenrostro et al., 2013). *Tn5* transposase simultaneously cuts exposed genomic DNA and ligates adapter sequences to those exposed DNA fragments (Buenrostro et al., 2013). The resultant fragments can subsequently be PCR amplified and used in deep sequencing experiments to identify regions of open chromatin (Buenrostro et al., 2013). In sum, despite their apparent inactivity in the human genome, DNA transposons have become important tools used to study human biology.

Retrotransposons

Class I transposable elements, also known as retrotransposons, account for an impressive ~42% of genomic DNA (Cordaux and Batzer, 2009; Lander et al., 2001) (Figure 1.1). Retrotransposons move via a "copy and paste" mechanism, called retrotransposition, where an RNA molecule(s) is the retrotransposition intermediate (Boeke et al., 1985; Craig et al., 2015). Retrotransposons can generally be subdivided into two classes: long terminal repeat (LTR) and non-LTR retrotransposons (Figure 1.1). Importantly, because retrotransposons utilize an RNA intermediate their mobility relies on reverse transcriptase (RT) activity. Autonomous retrotransposons typically encode their own RT, whereas non-autonomous retrotransposons must 'hijack' an RT-containing protein from an autonomous retrotransposon (Richardson et al., 2015).

LTR retrotransposons

In humans LTR retrotransposons comprise ~8% of human genomic DNA (Figure 1.1) (Lander et al., 2001). Autonomous LTR retrotransposons are closely related to retroviruses and are characterized by the presence of Long Terminal Repeats (LTRs) at their 5' and 3' termini (Figure 1.2). LTRs are regulatory sequences that play important roles in the expression of the retrotransposon RNA, cDNA synthesis, and cDNA integration (Beauregard et al., 2008; Craig et al., 2015). Endogenous retroviruses

(ERVs) are the most relevant autonomous LTR retrotransposons with respect to human biology (Craig et al., 2015). Like retroviruses, complete ERVs encode *gag, pol, and env* genes (Craig et al., 2015) (Figure 1.2). The majority of ERVs in humans are incomplete; they harbor deletions or nonsense mutations in one or more of the *gag, pol,* and *env* genes (Craig et al., 2015; Hohn et al., 2013; Weiss, 2016).

Non-autonomous ERVs contain LTRs, but generally do not encode functional genes (Craig et al., 2015; Havecker et al., 2004). Instead, non-autonomous ERV retrotransposition is dependent on 'hijacking' proteins expressed from an autonomous ERV in *trans* (Craig et al., 2015; Havecker et al., 2004). As with DNA transposons, the interaction between autonomous and non-autonomous ERVs necessarily results in competition for recruitment of catalytic proteins (Craig et al., 2015; Havecker et al., 2004).

ERV retrotransposition relies on *gag* and the multi-functional *pol* gene. After translation, the gag and pol proteins form a virus like particle (VLP) around the ERV mRNA (Craig et al., 2015; Gerdes et al., 2016). The *pol* gene contains protease, reverse transcriptase, integrase, and ribonuclease H activity (Craig et al., 2015; Gerdes et al., 2016; Nelson et al., 2003). The ERV cDNA is synthesized via *pol* RT activity within the VLP (Craig et al., 2015; Gerdes et al., 2016). The ERV integrase binds the LTRs of the mature ERV cDNA and mediates integration of ERV cDNA into the genome (Craig et al., 2015; Gerdes et al., 2016). ERV cDNA integration delivers the entire sequence of the newly replicated ERV and, like DNA transposition, results in short, 4-6 bp TSDs (Craig et al., 2015; Gerdes et al., 2016).

ERVs are not currently amplifying in the modern human genome and largely ceased retrotransposing in the lineage that gave rise to humans ~40 MYA (Cordonnier et al., 1995; Craig et al., 2015; Lander et al., 2001; Smit, 1993). There are, however, a handful of polymorphic human ERVs (HERV-K) insertions in the human population (Hohn et al., 2013; Macfarlane and Badge, 2015; Medstrand and Mager, 1998; Shin et al., 2013). A recent study identified an additional 36 polymorphic HERV-K insertions in data obtained from the thousand genomes project (Wildschutte et al., 2016). Thus,

these data suggest that after the human-chimpanzee divergence, some HERV-Ks retained activity, but have since ceased amplifying in the human genome.

Despite their inability to retrotranspose, ERVs influence various cellular and genomic processes (Craig et al., 2015). Some ERVs contribute to gene regulation during human embryonic development (Grow et al., 2015; Lim and Knowles, 2015; Rebollo et al., 2012; Samuelson et al., 1996), while others contain sequences that are critical for the expression of certain interferon genes (Chuong et al., 2016). Thus, as presciently predicted by Britten and Davidson, in some instances, sequences derived from ERVs have become fodder for the development and acquisition of novel *cis*-acting gene regulatory circuits (Britten and Davidson, 1969; Chuong et al., 2016). Additionally, expression of gag and pol proteins has been associated with some human cancers (Bannert and Kurth, 2006; Conrad et al., 1997; Galli et al., 2005; Hohn et al., 2013). Finally, ERV *env* gene(s) were domesticated in some mammalian lineages, leading to the formation of present day *Syncitin* genes, which play critical roles in placental development (Dupressoir et al., 2012).

Non-LTR retrotransposons

Non-LTR retrotransposons are similar to LTR retrotransposons in that they require an RT-containing protein to mediate their retrotransposition (Luan et al., 1993; Moran et al., 1996). As their name implies, non-LTR retrotransposons lack long terminal repeat sequences; however, they are flanked by variable length TSDs that generally range in size from ~7-20 bp (Gilbert et al., 2005; Lander et al., 2001). Non-LTR retrotransposons comprise 35% of human genomic DNA and are the only active TEs in the human genome (Cordaux and Batzer, 2009; Lander et al., 2001) (Figure 1.1). Human autonomous non-LTR retrotransposons are called Long INterspersed Elements (LINEs) and are discussed in greater detail below.

Non-autonomous non-LTR retrotransposons comprise ~11% of genomic DNA (Figure 1.1, 1.2). Human non-autonomous non-LTR retrotransposons do not encode an RT-containing protein and must ‘hijack’ an RT in *trans* from an autonomous non-LTR retrotransposon (Dewannieux et al., 2003). In humans, non-autonomous non-LTR retrotransposons include Mammalian Interspersed Repeats (MIRs) (Lander et al., 2001;

Smit, 1999), Short INterspersed Elements (SINEs; *e.g.,* Alu) (Deininger et al., 1981; Dewannieux et al., 2003), and SINE-R/VNTR/Alu (SVA) elements (Hancks et al., 2011; Ostertag et al., 2003; Raiz et al., 2012b). Though not termed SINEs, other cellular RNAs [*e.g.*, U6 small nuclear RNA (U6 snRNA) and U3 small nucleolar RNA (U3 snoRNA)] can also 'hijack' proteins encoded by autonomous non-LTR retrotransposons (Buzdin et al., 2002; Garcia-Perez et al., 2007a; Gilbert et al., 2005; Wei et al., 2001). Similarly, retrotransposition of mRNAs can result in the formation of processed pseudogenes (Esnault et al., 2000; Wei et al., 2001) (Figure 1.1, 1.2).

LINEs in the Human Genome

The only active autonomous retrotransposon in the human genome is termed Long INterspersed Element-1 (LINE-1 or L1). Sequences derived from L1s account for ~17% of human genomic DNA and are present at greater than 500,000 copies in the haploid genome (Lander et al., 2001) (Figure 1.1). The vast majority of L1 sequences (>99.9%) have been rendered inactive via pre-mature 5' truncation, point mutations that inactivate the L1-encoded proteins, or structural rearrangements (Grimaldi et al., 1984; Kazazian and Moran, 1998; Lander et al., 2001; Ostertag and Kazazian, 2001). The average human diploid genome, however, contains 80-100 retrotransposition-competent L1s (RC-L1) (Brouha et al., 2003; Sassaman et al., 1997). The mechanism of L1 retrotransposition will be discussed in greater detail later in the Introduction. L1 sequences arose and underwent a monophyletic expansion ~150 MYA, prior to the eutherian-marsupial split (Lander et al., 2001; Smit, 1996) (Figure 1.3). LINE-2 (L2) and LINE-3 (L3) elements are even more ancient and will briefly be described next.

Evolution of LINEs

"We are survival machines – robot vehicles blindly programmed to preserve the selfish molecules known as genes." (Dawkins, 1976).

Homo sapiens, like all organisms, are ultimately vehicles that transmit genetic material to the next generation. Yet within our genomes, repositories that contain all the information that makes us, lies a selfish genetic element that is truly a molecular robot. LINEs sole function is replication and propagation, preferably in germline cells, to ensure that new copies invade the next generation. LINEs follow a simple single rule: "If

they can propagate, they will." In principle, LINEs can retrotranspose in any cell type. Any cellular consequences of LINE retrotransposition (*i.e.*, generating beneficial, detrimental, or neutral mutations) are nothing more than downstream outcomes of the successful replication of a "molecular robot." LINEs have been supremely successful at populating the human genome and comprise the single largest fraction of our genomes (Lander et al., 2001). The function of LINEs is conceptually simple, though the retrotransposition mechanism and resultant affects on the genome are rather complex. Indeed, LINEs represent a major evolutionary force that drives both genome and molecular evolution (Cordaux and Batzer, 2009; Kazazian, 2004; Kazazian and Moran, 1998; Beck et al., 2011; Richardson et al., 2015).

Ancient LINEs

Recognizable LINEs in the human genome include L1s, L2s, and L3s. L2 and L3 sequences comprise ~6% of the human genome (Figure 1.1). Both L2 and L3 are a part of the ancient chicken repeat 1 (CR1) clade of retrotransposons and are present in mammals, marsupials, birds, reptiles, fish, and insects (Jurka, 1998, 2000; Kapitonov and Jurka, 2003; Meyers, 2007). Present at less than a thousand copies in the human genome, all L3s are severely mutated and none are full-length (Jurka, 1998; Kapitonov and Jurka, 2003).

L2 sequences are not necessarily younger than L3s, but apparently were active for a longer period of time in the mammalian lineage (Meyers, 2007) (Figure 1.3). L2 amplification in the mammalian lineage occurred largely before the mammalian radiation (~65 MYA) (Jurka, 1998; Lander et al., 2001; Smit, 1996). In the lineage that gave rise to modern humans, L2 activity likely ceased ~80-100 MYA (Jurka, 1998; Lander et al., 2001; Smit, 1996) (Figure 1.3). L2s were extremely successful, as there are ~300,000 L2 sequences in the extant human diploid genome (Jurka, 1998; Lander et al., 2001). Just like present day SINEs parasitize the RT-containing protein of L1, ancient mammalian interspersed repeats (MIRs) apparently parasitized the RT-containing protein of L2 (Jurka, 1998; Lander et al., 2001; Smit, 1996). The L2 3'UTR contains a 50 bp sequence required for L2 RT binding and subsequent L2 retrotransposition; MIRs also harbored that 50 bp sequence at their 3' end (Lander et al., 2001; Smit, 1996,

1999). These data suggest that MIR elements were able to hijack the L2 EN and RT activities to mediate their retrotransposition to new genomic locations (Smit, 1996).

LINE-1s

Roughly 150 MYA, before the eutherian/metatherian split, L1s began amplifying in the genomes of early mammalian species (Lander et al., 2001; Smit, 1996) (Figure 1.3). L1s amplified concomitantly with L2 elements for millions of years until L2s ceased to proliferate ~80-100 MYA (Lander et al., 2001) (Figure 1.3). The reason why L1 activity supplanted L2 activity is not entirely clear. It is possible that MIRs evolved a 3'UTR that was more efficient in recruiting the L2 containing RT than L2 itself. The ability of MIRs to effectively out compete L2s for RT binding may, in turn, have slowed L2 retrotransposition, allowing genetic drift and host repressive mechanisms to mutate full-length L2 copies, leading to their demise.

In contrast to L2s, L1s employ a *cis*-preference mechanism to recruit RT. Recent data indicate that nascent ORF2p co-translationally binds to the L1 poly(A) tail and suggest ORF2p binding to the L1 poly(A) tail may inhibit further ORF2p synthesis (Ahl et al., 2015; Alisch et al., 2006; Doucet et al., 2015; Esnault et al., 2000; Wei et al., 2001). Recent data also suggest that Alu elements, which are RNA polymerase III transcripts, associate with translating ribosomes, effectively allowing their encoded poly(A) tract to compete with the L1 RNA poly(A) tail for nascent ORF2p binding (Ahl et al., 2015; Doucet et al., 2015; Doudna and Rath, 2002). Indeed, we propose that the ability of L1 ORF2p to bind either the L1 poly(A) tail or Alu poly(A) tract does not necessarily hinder L1 retrotransposition, allowing both elements to co-amplify in current human genomes.

L1s have undergone bursts of rapid expansion followed by periods of lower activity (Lander et al., 2001), a phenomenon known as subfamily succession (Boissinot et al., 2000; Boissinot et al., 2004a; Boissinot and Furano, 2005). L1 subfamily succession can be modeled in terms of a host-parasite arms race—periods of expansion would typify a scenario where L1s are amplifying at high rates, whereas periods of lower activity could reflect times when host repressive processes are limiting unabated L1 retrotransposition.

L1 evolution in mammalian genomes has been well studied and has allowed the reconstruction of consensus sequences of L1 subfamilies (Boissinot et al., 2000; Boissinot et al., 2004a; Boissinot and Furano, 2005; Boissinot et al., 2004b; Khan et al., 2006; Smit et al., 1995). There are two major subdivisions of L1s: those that are present in all mammals (L1M) and those that are only present in primates (L1P) (Smit et al., 1995) (Figure 1.3). Further divisions (called subfamilies) add an additional letter (*e.g*. L1MA-MD and L1PA-PB), and subsequently a number (*e.g*., L1PA15).

Phylogenetic studies indicate that in the earliest primate genomes, both the L1M and L1P subfamilies were evolving simultaneously, but as independent lineages (Khan et al., 2006; Smit et al., 1995) (Figure 1.3). The major difference between various subfamilies amplifying within and between those groups is their respective 5'UTR sequence (Khan et al., 2006) (Figure 1.4). Acquisition of new 5'UTR sequences and mutations within 5'UTRs is likely a primary mechanism driving L1 evolution (Khan et al., 2006) (Figure 1.4). In the lineage leading to modern humans, the L1PB1, L1PA11, and perhaps L1MA1 subfamilies apparently were co-amplifying until ~55 MYA (Khan et al., 2006). After that time, it appears that L1PA11 sequences acquired a new 5'UTR yielding the L1PA10 subfamily (Khan et al., 2006) (Figure 1.3, 1.4). After L1PA10, L1 evolution followed a curious path of subfamily succession where one L1 subfamily was predominantly active in the genome at any given time (Khan et al., 2006) (Figure 1.3, 1.4). This evolutionary trajectory led to a monophyletic lineage of L1 amplification over the last 40MY (*e.g.*, L1PA10 sequences gave rise to L1PA8 (there is no L1PA9), which gave rise to L1PA7, L1PA6, L1PA5….*etc.*, ultimately leading to the L1Hs lineage, which is currently active in the human genome) (Figure 1.4). It is likely that numerous short-lived L1 sequences branched out from these main subfamilies (*e.g.*, L1PA8A), but only few were active for long enough periods of time to become fixed in the human genomic record.

Inspecting non-human primate genomes and comparative genomic studies permits an estimation of the age and success of L1 subfamilies (Bannert and Kurth, 2006; Boissinot and Furano, 2005; Khan et al., 2006; Smit, 1996, 1999; Smit et al., 1995). L1 became wildly successful in primate genomes. The fossil record suggests the earliest primates appeared ~55-65 MYA, although molecular data suggests an earlier

date of primate emergence of ~75-85 MYA (Goodman et al., 1998; Tavare et al., 2002). Subfamily succession likely occurred only after the primate lineage was established (Boissinot and Furano, 2005; Khan et al., 2006). Primate specific L1s also underwent bursts of activity followed by periods of lower activity (Boissinot and Furano, 2005; Goodman et al., 1998; Khan et al., 2006; Tavare et al., 2002) (Figure 1.3). For example, the L1PA12 family (amplifying ~60MYA) is represented by just under 900 insertions in the human genome reference (HGR) (Khan et al., 2006) (Figure 1.3). In contrast, subfamilies L1PA7, 5, 4, and 3 (amplifying 30, 21, 18, 12 MYA, respectively) are each represented by more than 8000 copies in the HGR (Khan et al., 2006) (Figure 1.3). This burst of retrotransposition occurred shortly after the Old World/New World primate divergence and continued up to the lineage giving rise to modern humans and apes ~25 MYA (Khan et al., 2006) (Figure 1.3).

Since the expansion of the L1PA5, 4, and 3 subfamilies in the ape lineage, L1 expansion appears to have slowed down (Khan et al., 2006). The L1PA2 subfamily, which is common to chimps and humans, amplified ~8 MYA and is represented by half the number of sequences as L1PA3 in the HGR (Khan et al., 2006). Since the lineage giving rise to modern humans diverged from the chimpanzee lineage ~6 MYA, only ~1100 the human specific L1 (L1PA1 or L1Hs) sequences have accumulated in the HGR (Khan et al., 2006; Mills et al., 2006). Interestingly, the decrease in human specific L1 sequences is not born out in the chimpanzee lineage. The *Pan trogolodytes* specific L1 sequence (L1Pt) is at least twice as numerous in the chimpanzee reference genome than L1Hs in the HGR (Hormozdiari et al., 2013).

L1 evolution is primarily driven by mutation and large-scale changes to the 5'UTR (Khan et al., 2006) (Figure 1.4). Of the L1PA family of elements, the longest 5'UTR is in L1PA12 (2460 bp); the shortest is in L1Hs (910 bp). Aside from a stretch of guanine nucleotides and a conserved Yin-Yang 1 (YY1) transcription factor binding-site at the 5' end of the 5'UTR, there is little conservation in the 5'UTR amongst L1PA subfamilies. The YY1 site is critical in the positioning of L1 transcription at or near the 1st nucleotide of the element (Athanikar et al., 2004). The final 5'UTR substitution occurred in L1PA10 giving rise to L1PA8 (~40MYA) (Khan et al., 2006) (Figure 1.3, 1.4). Subsequently, through monophyletic subfamily succession, sequences in the 5'UTR were both lost and

acquired, and led to an overall shortening of the 5' UTR from 1338 bp in L1PA8, to its present size, ~910 bp in L1Hs (Khan et al., 2006).

Rapid evolution of the 5'UTR could be due to two non-mutually exclusive processes: 1) competition between co-amplifying L1 subfamilies for limiting host factors essential for retrotransposition; and 2) escape from host factor repression. Host proteins have evolved to target L1 5'UTR sequences and repress L1 transcription. L1 sequences must, in turn, evolve to 'escape' the transcriptional repressive effects of these host proteins. The evolutionary relationship of host repression followed by L1 response exemplifies a "Red Queen" scenario, where active L1s must retrotranspose new full-length L1 copies as quickly as possible to evade the repressive effects of host proteins and genetic drift host (Carroll et al., 2004; Vanvalen, 1973).

A beautiful example of a "Red Queen" scenario was recently demonstrated in a study from the Haussler lab. Cell culture based experiments demonstrated that a zinc-finger protein, ZNF93, inhibits L1 expression and retrotransposition, but in a subfamily specific manner (Jacobs et al., 2014). The ZNF93 protein binds to a 129 bp sequence within the 5'UTR of L1PA3 subfamily members, leading to transcriptional repression (Jacobs et al., 2014). The loss of the ZNF93 binding from the 5'UTR of L1PA2 subfamily members allowed them to escape ZNF93-mediated transcriptional repression and undergo subsequent amplification in the lineage leading to humans (Jacobs et al., 2014).

Additional observations suggest that other repressive factors may have led to the extinction of older L1 subfamilies. For example, recent studies demonstrated that the zinc-finger protein, KAP1, binds to the 5'UTR and represses transcription of older L1s (L1PA6-L1PA2), but not the currently active L1Hs (Castro-Diaz et al., 2014).

In sum, using anthropomorphic reasoning, the evolutionary "objective" of L1 is simple: it retrotransposes throughout the genome as much as the host can tolerate. The host, in turn, evolves restriction factors that inhibit unabated L1 retrotransposition. L1 then subsequently evolves to escape those restrictive factors—and the cycle continues. Thus, L1 amplification meets the definition of a successful selfish genetic element that,

in most mammalian genomes, has not succumbed to repressive processes during the last 150 MY.

Structure of an Active Human LINE-1

Active human L1s are ~6 kilobases (6 kb) in length (Dombroski et al., 1991; Scott et al., 1987) and contain a 5'UTR that harbors both sense (Swergold, 1990) and anti-sense (Speek, 2001) RNA polymerase II promoters (Figure 1.2B). An open reading frame (ORF0) with unknown function resides in the anti-sense orientation within the 5'UTR (Denli et al., 2015). Following the 5'UTR are two open reading frames (ORF1 and ORF2) that are separated by a 63 base pair (bp) inter-ORF spacer that contains two in-frame stop codons (Alisch et al., 2006; Dombroski et al., 1991). L1 ends with a 3'UTR positioned between the ORF2 stop codon and a variable-length poly adenosine tract (Dombroski et al., 1991; Grimaldi et al., 1984; Scott et al., 1987). Genomic L1s are flanked by variable-length (~7-20 bp) TSDs (Gilbert et al., 2005; Gilbert et al., 2002; Richardson et al., 2015; Symer et al., 2002).

5'UTR

The L1 5'UTR is ~910bp and contains an internal RNA polymerase II promoter(s) that direct the transcription of both sense (Athanikar et al., 2004; Becker et al., 1993; Swergold, 1990) and anti-sense (Speek, 2001) L1 RNAs. The centrally located anti-sense promoter is weaker than the sense promoter, but has been shown to drive transcription of adjacent genes residing upstream of a genomic L1 in a tissue-specific manner (Macia et al., 2011; Matlik et al., 2006; Nigumann et al., 2002; Speek, 2001).

Experimental evidence suggests that anti-sense driven promoter transcripts may bind to sense L1 promoter transcripts to create double stranded RNAs that are subsequently processed into siRNAs by Dicer to repress L1 retrotransposition (Yang and Kazazian, 2006). It has also been suggested that the microprocessor (Drosha-DGCR8) binds to and catalytically processes anti-sense transcripts emanating from 5'UTR; however, the role of these processed transcripts in L1 retrotransposition, if any, remains unclear (Heras et al., 2013). Some studies have suggested that the aforementioned processes may represent self-regulatory mechanism that acts as a "rheostat" to dampen L1 retrotransposition (Yang and Kazazian, 2006). However, these

findings violate a key tenet of a selfish genetic element; it is more likely that the L1 anti-sense promoter somehow benefits L1 expression. Indeed, more than 10% of human protein coding genes contains both sense and anti-sense promoters (Bratthauer and Fanning, 1993; Core et al., 2012; Thayer et al., 1993; Trinklein et al., 2004). Some evidence from yeast suggests that anti-sense transcription from gene promoters is associated with alterations in local chromatin that may facilitate sense transcription (Murray et al., 2015). However, the authors of Murray et al., 2015, observed only a slight increase in promoter activity when compared to promoters that lack anti-sense transcription. Rigorous dissection of the L1 anti-sense promoter is necessary to determine any effect it may have on L1 transcription and retrotransposition.

At least five defined transcription factor binding sites reside within the L1 5'UTR. The YY1 transcription factor binds at +13 and directs transcriptional initiation at or near the first nucleotide of the L1 5'UTR (Athanikar et al., 2004; Becker et al., 1993). Two RUNX3 transcription factor-binding sites have been identified (Yang et al., 2003). One of these, at +90, strongly enhances sense L1 transcription and retrotransposition in a cell cultured based assay (Yang et al., 2003). The other RUNX3 site, at +510, may regulate anti-sense promoter activity (Yang et al., 2003). Two SRY-like sites at +472 and +572 are responsive to SOX11 and SOX3 overexpression in cell culture retrotransposition assays and drive L1 promoter activity (Tchenio et al., 2000). Conversely, studies in transgenic mice neuronal cells suggest SOX2 may repress L1 transcription (Muotri et al., 2005).

ORF1

The first L1 open reading frame (ORF1) immediately follows the 5'UTR. ORF1 encodes a ~40 kiloDalton (kDa) protein, ORF1p, containing three defined domains (Hohjoh and Singer, 1996, 1997b; Holmes et al., 1992; Martin, 1991) (Figure 1.2). The amino terminus of ORF1p consists of a coiled-coil domain that is required for ORF1p trimerization (Khazina et al., 2011; Khazina and Weichenrieder, 2009; Martin et al., 2003). Mutations disrupting ORF1p trimerization disrupt L1 retrotransposition (Basame et al., 2006; Doucet et al., 2010; Khazina et al., 2011; Khazina and Weichenrieder, 2009).

An RNA Recognition Motif (RRM) is centrally located in ORF1p (Khazina et al., 2011; Khazina and Weichenrieder, 2009), which is followed by a carboxy-terminal domain (CTD) that is rich in positively charged amino acids (Moran et al., 1996). ORF1p has been demonstrated to bind to L1 RNA forming a ribonucleoprotein particle (RNP) (Doucet et al., 2010; Hohjoh and Singer, 1996, 1997b; Kulpa and Moran, 2005; Martin, 1991). Evidence suggests that the RRM and carboxyl-terminal domains interact and are critical for ORF1p nucleic acid binding (Hohjoh and Singer, 1996, 1997b; Holmes et al., 1992; Januszyk et al., 2007; Khazina et al., 2011; Khazina and Weichenrieder, 2009; Martin, 1991). The RRM and CTD domains may also be important for ORF1p nucleic acid chaperone activity (Khazina and Weichenrieder, 2009; Martin and Bushman, 2001). ORF1p has been demonstrated to bind both ssDNA, and with less affinity dsDNA (Callahan et al., 2012; Khazina and Weichenrieder, 2009; Martin and Bushman, 2001; Martin et al., 2005). ORF1p chaperone activity is hypothesized to facilitate L1 integration into the genome (Martin and Bushman, 2001). Additionally, three short blocks of conserved amino acids in the CTD were demonstrated to be crucial for L1 retrotransposition in a cell-culture based assay (Moran et al., 1996).

ORF2

The second ORF encodes a ~150 kDa, ORF2p, which mediates L1 cDNA insertion into the genome (Figure 1.2) (Doucet et al., 2010; Ergun et al., 2004; Feng et al., 1996; Goodier et al., 2010; Moran et al., 1996). ORF2p contains at least three functional domains. The N-terminal endonuclease (EN) domain was originally identified in a *Trypanosom cruzi* non-LTR retrotransposon L1Tc, (Martin et al., 1995) and resembles an apurinic/apyrimidinic endonuclease (APE) (Feng et al., 1996). The L1 EN domain creates a single-strand endonucleolytic nick in both super-coiled and relaxed DNA (Feng et al., 1996), but unlike traditional APEs (Mol et al., 1995), does not appear to have a preference for cleaving DNA at abasic sites. The L1 EN domain makes a single-strand endonucleolytic cleavage at a degenerate consensus sequence [*e.g.*, 5'-TTTT/A-3' where the (/) represents the location of the scissile phosphate] (Feng et al., 1996). The consequence of endonuclease activity is the liberation of a free 5'-monophosphate and 3'-hydroxyl group (Feng et al., 1996).

The reverse transcriptase (RT) domain of ORF2p follows 3' of the EN domain. The core RT domain is homologous across all sequences that encode an RT, including group II introns, telomerase, LTR-retrotransposons, and retroviruses (Craig et al., 2015; Eickbush, 1997; Eickbush et al., 1997; Hattori et al., 1986; Malik et al., 1999; Xiong and Eickbush, 1990). Experiments in yeast expressing Ty1/ORF2 fusion proteins first demonstrated L1 ORF2p contained RT activity (Dombroski et al., 1994; Mathias et al., 1991). Later work indicated that the RT domain of ORF2p mediates both RNA- and DNA-dependent DNA polymerization (Cost et al., 2002; Piskareva et al., 2003; Piskareva and Schmatchenko, 2006). Like ORF1p, ORF2p is present in ribonucleoprotein particles (RNPs) (Doucet et al., 2010; Kulpa and Moran, 2005, 2006). *In vitro* studies have demonstrated that RNP-associated ORF2p RT activity could utilize L1 RNA as a template for reverse transcription (Doucet et al., 2010; Kulpa and Moran, 2005, 2006).

The final domain of ORF2p is a cysteine-rich (C) domain (Fanning and Singer, 1987; Moran et al., 1996). The function of the C domain is unclear, but mutations in the C domain reduce RT activity, ORF2p localization to RNPs, and retrotransposition of L1 RNAs in human cells (Doucet et al., 2010; Moran et al., 1996) and RT activity in yeast (Clements and Singer, 1998). Recent experiments performed in a cell culture based retrotransposition assay suggest that a ccarboxyl-terminal tyrosine (amino acid 1180) residue is dispensable for L1 retrotransposition, but may aid in ORF2p-mediated retrotransposition of Alu RNAs (Christian et al., 2017). Additional experimentation is required to further elucidate the function of the C domain.

3'UTR

L1s end with a ~200 bp 3'UTR situated between the ORF2 stop codon and a variable length poly(A) sequence at the L1 3' terminus (Dombroski et al., 1991; Grimaldi and Singer, 1983; Grimaldi et al., 1984). A poly-guanosine tract present proximal to the beginning of the 3'UTR is a general feature of mammalian 3'UTRs and may form a stable G-quadruplex (Howell and Usdin, 1997; Sahakyan et al., 2017; Usdin and Furano, 1989). Despite this conservation, the 3'UTR is dispensable for L1 retrotransposition, at least in cultured human cells (Moran et al., 1999; Moran et al.,

1996). Though the function of the guanosine tract remains unclear, its conservation suggests that it may play a role in some aspect of L1 RNA biology. The 3'UTR ends in a weak poly(A) signal that is often bypassed in favor of adjacent poly(A) signals present in downstream genomic DNA (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996). Read through transcripts that bypass the canonical L1 poly(A) signal allows 3' genomic sequences to be appended to the L1 RNA; retrotransposition of these RNAs results in L1-mediated 3' transductions (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996).

LINE-1 Retrotransposition Cycle

Continued success of L1 requires the faithful retrotransposition of a full-length RNA. Genomic L1s are generally 5' truncated or contain internal deletions and/or inversions (Grimaldi et al., 1984; Lander et al., 2001). However, ~1/3 of endogenous L1Hs sequences are full-length (Lander et al., 2001). Similarly, L1 insertions recovered from engineered retrotransposition events in HeLa cells demonstrated that 6/100 were full-length insertions and an additional four contained ~6 kb of L1 sequence that was either internally rearranged or contained short 5' deletions (Gilbert et al., 2005). Despite the low percentage of full-length insertions, L1 genomic retrotransposition has been extremely successful over evolutionary time.

A round of L1 retrotransposition begins with the transcription of a genomic, retrotransposition-competent L1 (RC-L1) (Figure 1.5). RNA polymerase II initiates transcription at or near the first nucleotide of the 5'UTR (Athanikar et al., 2004; Becker et al., 1993; Swergold, 1990). The bicistronic, polyaddenylated L1 RNA is exported to the cytoplasm where ORF1 is translated in a cap-dependent manner (Dmitriev et al., 2007; Leibold et al., 1990; McMillan and Singer, 1993) (Figure 1.5). Translation initiation of the second open reading frame (ORF2) is incompletely understood. It has been suggested that ORF2p translation may involve an unconventional ribosomal termination-reinitiation mechanism (Alisch et al., 2006; Dmitriev et al., 2007; McMillan and Singer, 1993). Notably, ORF2 translation can initiate in an AUG-independent manner and ORF2 is efficiently translated regardless of the composition of the upstream ORF (Alisch et al., 2006). One study suggests that mouse ORF2 could also be

translated via ribosome termination-reinitiation (Alisch et al., 2006). Another study suggests that mouse ORF2 is translated with the aid of an internal ribosome entry site (IRES) at the end of ORF1p (Li et al., 2006), although deletion of this sequence in human L1 does not prevent ORF2 translation (Alisch et al., 2006). Additional experimentation is necessary to determine if an IRES is important for mouse L1 retrotransposition. Importantly both ORF1p and ORF2p are translated independently as no fusion ORF1/2p proteins have been identified (Alisch et al., 2006; Leibold et al., 1990).

Following translation, both ORF1p (Hohjoh and Singer, 1996, 1997a, b) and ORF2p (Kulpa and Moran, 2005, 2006) bind L1 RNA by a mechanism known as *cis*-preference (Doucet et al., 2015; Esnault et al., 2000; Kulpa and Moran, 2006; Wei et al., 2001) forming a ribonucleoprotein particle (RNP). RNP formation is necessary, but not sufficient, for LINE-1 retrotransposition (Kulpa and Moran, 2005) (Figure 1.5). In addition to biochemical assays, immunofluorescence of epitope-tagged L1 proteins demonstrated an intimate cytoplasmic co-localization of L1 RNA, ORF1p, and ORF2p (Doucet et al., 2010; Goodier et al., 2010). Recent work has shown that the poly(A) tail of L1 RNA is crucial for retrotransposition and likely acts to recruit ORF2p to L1 mRNA (Doucet et al., 2015). The L1 RNP likely consists of other cellular proteins and RNAs in addition to ORF1p, ORF2p, and L1 RNA (Goodier et al., 2013; Moldovan and Moran, 2015; Taylor et al., 2013). Determining what other constituents distinguish functional *vs*. nonfunctional RNPs, will be crucial to better understanding the mechanism of L1 retrotransposition.

Following RNP formation, components of the L1 RNP enter back into the nucleus by a process that does not strictly require cell division (Kubo et al., 2006) (Figure 1.5). L1 integration occurs by a mechanism termed target-site primed reverse transcription (TPRT) and minimally requires ORF2p and L1 RNA (Cost et al., 2002; Feng et al., 1996; Kulpa and Moran, 2006; Luan et al., 1993) (Figure 1.5). TPRT initiates with a single-strand endonucleolytic nick at a genomic target DNA sequence mediated by ORF2p EN (Cost et al., 2002; Feng et al., 1996; Morrish et al., 2002, Luan et al., 1993). The single-strand nick occurs in a thymidine-rich sequence (*e.g.,* 5'-TTTT/A-3') where the (/) denotes the siscile phosphate (Cost and Boeke, 1998; Feng et al., 1996; Morrish

et al., 2002). The result of the endonucleolytic nick is exposure of a reactive 3' hydroxyl group, which ORF2p RT activity uses as a primer to initiate minus (-) strand L1 cDNA synthesis (Cost and Boeke, 1998; Feng et al., 1996) (Figure 1.5). It is unclear how second strand cDNA synthesis is completed. Top (+) strand cleavage likely occurs downstream of the initial nick, but whether this is performed by ORF2p or another cellular enzyme requires elucidation. By analogy to the R2 retrotransposon of *Bombyx mori,* it is likely that ORF2p cleaves the top strand and polymerizes second strand L1 cDNA (Christensen and Eickbush, 2005; Luan et al., 1993). The completion of TPRT results in the insertion of a new L1 cDNA molecule flanked by variable length/sequence target site duplications (TSDs) (Figure 1.5).

Despite *cis*-preference, other cellular RNAs are occasionally mobilized by associating with L1 ORF1p and/or ORF2p. In a process termed *trans*-complementation, RNAs 'hijack' L1 proteins that mediate their retrotransposition (Ahl et al., 2015; Doucet et al., 2015; Wei et al., 2001) (Figure 1.5). The Small INterspersed Element (SINE) family, including Alu (Deininger et al., 1981; Dewannieux et al., 2003) and SINE-R/VNTR/Alu (SVA) elements (Hancks et al., 2011; Ostertag et al., 2003; Raiz et al., 2012b), can parasitize the L1 encoded protein(s) (Ahl et al., 2015; Doucet et al., 2015). Indeed, Alus have been more retrotranspositionally successful than L1s, amplifying to over one million copies in the human genome (Lander et al., 2001).

The active human Alu is derived from the 7SL RNA component of the signal recognition particle (SRP) (Bennett et al., 2008; Kriegs et al., 2007; Okada et al., 1997; Ullu and Weiner, 1985; Weichenrieder et al., 2000). Alu retrotransposition requires L1 ORF2p (Bennett et al., 2008; Dewannieux et al., 2003) and a poly(A) tract (Dewannieux et al., 2003; Dewannieux and Heidmann, 2005). Experiments have demonstrated that Alu RNA binds to the signal recognition particle (SRP) 9/14 heterodimer (Weichenrieder et al., 2000) forming an Alu RNP that can interact with the ribosome (Ahl et al., 2015). It has been postulated that active translation of L1 RNAs prevents Alu from efficiently interacting with the ribosome, whereas stalled or slowed ribosomes are permissive for Alu RNP interaction (Ahl et al., 2015). The Alu RNP-ribosome interaction may arrest the L1 translating ribosome, which could favor Alu RNA 'hijacking' of L1 ORF2p (Ahl et al., 2015; Doucet et al., 2015; Doudna and Rath, 2002). The other active human SINE,

SVA, is much less abundant, present at less than 3,000 copies in the human genome (Ostertag et al., 2003; Wang et al., 2005) and unlike Alu, SVAs may require both ORF1p and ORF2p for their retrotransposition (Hancks et al., 2011; Raiz et al., 2012a).

In addition to SINEs, L1 ORF2p can mobilize cellular mRNAs leading to the formation of processed pseudogenes, which lack introns (Esnault et al., 2000; Wei et al., 2001) (Figure 1.2). Additionally, small uracil-rich nuclear RNAs (U6 snRNA) (Buzdin et al., 2002; Garcia-Perez et al., 2007a; Gilbert et al., 2005) and small nucleolar RNAs (U3 snoRNA) (Weber, 2006) can be mobilized in *trans* by L1 ORF2p. The panoply of RNAs that are mobilized by the L1-encoded protein(s) is exceptional. ORF2p enzymatic activity, through the mobilization of L1 or other RNAs, has singularly been responsible for over one-third, or one billion base pairs, of sequence in the human genome (Figure 1.1) (Cordaux and Batzer, 2009; Lander et al., 2001).

Effects of LINE-1 Retrotransposition

L1 biologists frequently get asked the question, “What do L1s do?” L1 function is simple; they selfishly create copies and spread throughout the genome. However, the effect of L1 amplification and insertion varies widely in scale and impact on genomic integrity (Beck et al., 2011; Richardson et al., 2015). As repeat sequences, genomic L1s can serve as scaffolds for non-allelic homologous recombination resulting in reciprocal chromosomal insertions and deletions. L1 insertion via TPRT can also mediate alterations of native genomic DNA. L1-mediated insertions have been implicated in causing diseases, including cancer, while also serving as drivers of genomic evolution by providing fodder for the evolution of gene regulatory systems (Chuong et al., 2017) and *de novo* gene creation (Moran et al., 1999; Nisole et al., 2004; Sayah et al., 2004).

LINE-1 in disease

A major breakthrough in understanding the effects of L1 retrotransposition on the human genome occurred in 1988, when Haig Kazazian discovered de *novo* L1 insertions in the *Factor VIII* gene of two unrelated boys afflicted with hemophilia A (Kazazian et al., 1988). Until that observation, L1s were thought to be inactive. Currently, 124 documented cases of L1-mediated retrotransposition events have resulted in human disease (Hancks and Kazazian, 2016). Of these, 76 are the result of

Alu insertions, 30 are due to L1 insertions, 13 are due to SVA insertions, four are due to poly(A) insertions, and one is due to a processed pseudogene insertion (Hancks and Kazazian, 2016). Furthermore, it has been suggested that 1/250 genetic diseases are the result of an L1-mediated retrotransposition event (Wimmer et al., 2011).

The effects of *de novo* retrotransposition-mediated mutagenesis are diverse. Both of the L1 insertions first identified in the *Factor VIII* gene inserted into exon 14 (Kazazian et al., 1988). Six L1 retrotransposition events into the *DMD* gene resulted in five cases of Duchenne Muscular Dystrophy and one case of X-linked Dilated Cardiomyopathy (Hancks and Kazazian, 2016). Of those events, four retrotransposed into exons resulting in exon skipping and/or the creation of pre-mature stop codons (Awano et al., 2010; Holmes et al., 1994; Musova et al., 2006; Narita et al., 1993), whereas one retrotransposed into the 5'UTR likely affecting *DMD* transcription (Yoshida et al., 1998). The effect of the remaining insertion remains unknown. Intronic L1 retrotransposition events resulting in exon skipping have also been observed in the *FKTN* gene resulting in Fukuyama congenital muscular dystrophy (Kondo-Iida et al., 1999) and the *SLCO1B3* gene resulting in Rotor syndrome (Kagawa et al., 2015). Finally, a L1 insertion resulted in a ~47 kb deletion within the *PDHX* gene resulting in Pyruvate Dehydrogenase Complex deficiency (Mine et al., 2007).

L1 mediated retrotransposition has also been implicated in cancer. The major question remains, "Does L1 mediated retrotransposition *directly drive* the cancer phenotype?" Currently, there are only a handful of examples that suggest L1 retrotransposition resulted in cancer. Somatic L1 retrotransposition into the exons of the *APC* gene (Miki et al., 1992; Scott et al., 2016), the *PTEN* gene (Helman et al., 2014), and within an intron of the *RB1* gene, resulting in cryptic splicing (Rodriguez-Martin et al., 2016), were all shown to drive cancer progression. Notably, *APC, PTEN, and RB1* are tumor suppressor genes. L1 mediated gene disruption could act as either the first, or second, 'hit' of a tumor suppressor gene to drive cancer progression (Knudson, 1971). Recently, a report demonstrated that an SVA insertion in the *CASP8* gene is associated with basal cell carcinoma and breast cancer, but curiously confers protection against prostate cancer (Stacey et al., 2016). Finally, another report suggests an L1 retrotransposition event into the *ST18* gene disrupted a transcriptional repressor,

leading to *ST18* oncogenic activation in a patient with hepatocellular carcinoma (Shukla et al., 2013).

Many laboratories have reported *de novo* L1 insertions in somatic tumor tissues (Doucet-O'Hare et al., 2015; Ewing et al., 2015; Helman et al., 2014; Iskow et al., 2010; Lee et al., 2012; Solyom et al., 2012; Tubio et al., 2014). The Kazazian laboratory has demonstrated that *de novo* L1 insertions can be found clonally in tumor tissues, while other insertions are found in tumor as well as surrounding healthy tissues (Doucet-O'Hare et al., 2015; Ewing et al., 2015). Recent evidence suggests that overexpression of L1 ORF2p may lead to genomic damage, which could, in principle, contribute to tumorigenesis (Gasior et al., 2007; Sciamanna et al., 2013; Sciamanna et al., 2014). The result of increased L1 retrotransposition in some tumors is unclear and continued efforts are necessary to determine the potency of retrotransposition-mediated mutagenesis and cancer progression.

L1 retrotransposition-mediated structural variation

L1-mediated retrotransposition events can contribute to various target-site alterations. Cell culture assays demonstrated that two bp to >71 kb of genomic DNA can be deleted upon retrotransposition (Gilbert et al., 2005; Gilbert et al., 2002; Symer et al., 2002). Genomic deletions that occur during TPRT often manifest as intrachromosomal deletions and could be formed by DNA recombination mechanisms that include non-homologous end joining, single-strand annealing, or synthesis dependent strand annealing (Gilbert et al., 2005; Gilbert et al., 2002; Symer et al., 2002).

L1 retrotransposition can also result in addition of genomic DNA sequence at the genomic target-site via a process called transduction. L1-mediated 5' and 3' transduced sequences alone account for ~1% of genomic DNA (Goodier et al., 2000; Pickeral et al., 2000). A 5' transduction occurs if a promoter upstream of the L1 5'UTR initiates transcription, resulting in an L1 RNA containing sequences transcribed from upstream genomic DNA (Evrony et al., 2012; Lander et al., 2001; Symer et al., 2002; Wei et al., 2001). L1-mediated 5' transductions are rare, but have been observed in the human genome reference sequence, in experiments conducted in human cells, and within brain neurons (Evrony et al., 2012; Lander et al., 2001; Symer et al., 2002; Wei et al., 2001).

L1-mediated 3' transductions occur when the canonical poly(A) sequence of L1 is bypassed and a downstream poly(A) sequence is utilized for transcription termination (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996). L1-mediated 3' transduction events are common due to the weak poly(A) signal present in the L1 sequence (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996) and are associated with ~20% percent of genomic L1s (Goodier et al., 2000; Beck et al., 2010; Kidd et al., 2010; Pickeral et al., 2000; Tubio et al., 2014). Transduced sequences also serve as useful molecular tags to identify actively retrotransposing L1s (Goodier et al., 2000; Badge et al., 2003; Beck et al., 2010; Brouha et al., 2002; Holmes et al., 1994; Macfarlane et al., 2013; Tica et al., 2016; Tubio et al., 2014).

Post-insertion structural variation

The plethora of L1 and Alu sequences in the human genome provides ample fodder for chromosomal alterations. For example, an intrachromosomal deletion between an L1 and Alu sequence, potentially by non-homologous end joining, resulted in ~430kb deletion within the *Dystrophin* gene (Suminaga et al., 2000). Additionally, non-allelic homologous recombination (NAHR) events between homeologous L1s or homeologous Alus has been demonstrated to lead to genomic deletions (Bailey et al., 2003; Fitch et al., 1991; Han et al., 2008; Han et al., 2005; Shen et al., 1981; Startek et al., 2015). NAHR between L1s has resulted in genomic deletions in the β-subunit of *phosphorylase kinase* (Burwinkel and Kilimann, 1998), deletion of the collagen genes *COL4A5 and COL4A6* (Burwinkel and Kilimann, 1998), and a deletion encompassing the *EVC, EVC2, C4orf6,* and *STK32B* genes (Temtamy et al., 2008). Similarly, NAHR between Alus resulted in a ~5 kb deletion of exons within the *Low-density Lipoprotein Receptor* gene (Lehrman et al., 1985) as well as numerous other interchromosomal deletions associated with disease (Konkel and Batzer, 2010). Finally, reports suggest that accumulation of L1 sequences surrounding *Abp* gene families in mice and rats drives rapid gene duplication via NAHR (Janousek et al., 2013; Janousek et al., 2016).

Retrotransposition-mediated evolution of new genes

In principle, L1 sequences can be co-opted for use in regulatory control of genes. An L1 5'UTR located upstream of the *Apolipoprotein A* (*Apo A)* gene was likely co-opted

as an enhancer to augment expression of *Apo A* (Yang et al., 1998). Additionally, L1 is replete with poly(A) signals (Perepelitsa-Belancio and Deininger, 2003). If an L1 retrotransposes within a gene, a cryptic L1 poly(A) sequence(s) could be utilized for pre-mature polyadenylation of the transcript, resulting in a 3' truncated genic mRNA (Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003).

L1-mediated retrotransposition of pseudogenes are responsible for the creation of two independent TRIM/Cyp gene fusion events in two different primate species. Owl monkeys contain a fusion protein that resulted from a processed *Cyclophilin A* (CypA) pseudogene retrotransposing into an intron of the *Trim5* gene (Nisole et al., 2004; Sayah et al., 2004). In addition, macaques harbor a TRIMCyp fusion resulting from retrotransposition of the CypA gene into the 3'UTR of the macaque *TRIM5* gene (Liao et al., 2007; Virgen et al., 2008). TRIMCyp fusion genes are potent anti-virals against some lentiviruses (Malfavon-Borja et al., 2013). Another proposed mechanism of novel gene creation is L1-mediated exon shuffling (Moran et al., 1999). This mechanism could occur either by 3' transduction of genomic sequences to new locations (Moran et al., 1999) or by mobilizing processed pseudogenes in *trans* as mentioned above for the TRIMCyp fusions. The identification of three human *AMAC* gene duplications mediated by SVA 3' transductions suggested that 3' transductions could result in the formation of new genes (Xing et al., 2006)

Host Defense to LINE-1 Retrotransposition

Addressed above are the effects of L1 retrotransposition on genomic integrity. L1 retrotransposition is inherently mutagenic; thus, it stands to reason that the cell has evolved numerous mechanisms to inhibit L1 retrotransposition. Indeed, at virtually every stage of the retrotransposition cycle, the host has developed mechanisms to inhibit unabated L1 retrotransposition (Figure 1.6) (Goodier, 2016).

DNA methylation and other transcriptional regulators

DNA methylation is frequently associated with gene silencing (Jones, 2012). Though DNA methylation affects numerous genes, the majority of methylated residues in the genome reside within TE sequences (Goll and Bestor, 2005; Yoder et al., 1997). Altered methylation of the L1 promoter was first identified in cancer cell lines (Thayer et

al., 1993). Additional investigation suggested that the 5'UTR of some younger, L1Hs subfamily members was de-methylated in some cancers (Alves et al., 1996) and resulted in increased expression of ORF1p (Bratthauer and Fanning, 1993; Thayer et al., 1993) (Figure 1.6). Subsequent in *vitro* and in *vivo* experimental evidence demonstrated methylation of the L1 CpG island within the human L1 5'UTR (Hata and Sakaki, 1997). Additional investigation in tumor cell lines revealed a correlation between L1 hypomethylation and increased L1 expression and retrotransposition (Iskow et al., 2010; Suter et al., 2004; Tubio et al., 2014). In brain tissue (Coufal et al., 2009), and in some hESC and iPSC derived cells lines (Klawitter et al., 2016; Wissing et al., 2012), a correlation between L1 5'UTR hypomethylation and L1 expression were also observed.

More recently, proteins hypothesized to be involved in L1 and non-LTR DNA methylation have been reported. Loss of *DNA-Methyltransferase Like 3 L* protein, *DNMT3L,* in knockout mice led to male meiotic catastrophe and sterility (Bourc'his and Bestor, 2004). Additionally, the testes of these mice expressed increased L1 and the LTR-retrotransposon, IAP, RNA (Bourc'his and Bestor, 2004). Identification and knockout of the catalytic methyltransferase, *DNMT3C,* resulted in a male germline phenotype similar to *DNMT3L* knockout mice (Barau et al., 2016). A triple knockdown of the *de novo* methyltransferases (*DNMT3A and DNMT3B*) as well as a maintenance methylase, *DNMT1,* in hESCs resulted in increased expression of young L1s (Castro-Diaz et al., 2014). Overexpression of the Methyl CpG binding protein 2 (MECP2) in HeLa cells repressed retrotransposition of an engineered L1 (Yu et al., 2001). *MECP2* knockout in mice and rats exhibited increased L1 retrotransposition in neuronal tissue (Muotri et al., 2010) as well as in cell lines derived from patients with Rett Syndrome, a neuron development disorder typically caused by MECP2 mutations (Muotri et al., 2010). In sum, these data suggest that methylation represses L1 transcription in both germ and somatic cells.

Evidence suggests histone modifications also are implicated in repressing L1 activity (Figure 1.6). One study, using a human embryonic carcinoma cell line, demonstrated that a reporter cassette delivered by an engineered L1 construct is subject to transcriptional silencing either during or shortly after integration (Garcia-Perez et al., 2010). Addition of pan-histone deacetylase inhibitors activated expression of the

integrated L1 reporter cassette, suggesting that histone deacetylases may target newly integrated L1s to restrict their expression (Garcia-Perez et al., 2010).

Alu sequences in HeLa cells were demonstrated to bind the histone methyltransferase, Suppressor of Variegation 39 H1, SUV39H1 (Varshney et al., 2015). After treatment of HeLa cells with a pan H3K9 inhibitor, SUV39H1 occupancy at Alus was diminished and a concomitant increase of Alu RNA was detected (Varshney et al., 2015). Recent evidence suggests knockout of the H3K9 di-methyltransferase, *G9A,* and *MILI* in mouse testis lead to increased expression of endogenous L1 and ORF1p (Di Giacomo et al., 2014). The authors hypothesize that piRNAs and histone methylation actively co-repress endogenous L1s in mouse germ cells (Di Giacomo et al., 2014). Though there is growing evidence that histone modifications may repress genomic L1 expression, a concerted effort is required to discern what role histone modifications may play in modulating L1 expression and retrotransposition.

Zinc finger (ZNF) transcription factors have also been shown to negatively regulate L1 transcription (Figure 1.6). Krüppel-associated box (KRAB) motif ZNF transcription factors (KZNF) represent a fast evolving set of transcription factors present only in tetrapods (Birtle and Ponting, 2006). The expansion KZNFs, in part, is likely a response to mobile element activity and it appears some KZNFs specifically inhibit different L1 subfamily members (Birtle and Ponting, 2006; Thomas and Schneider, 2011). For example, in hESCs, knockdown of TRIM28 (KAP1), a KZNF protein, resulted in increased L1 RNA expression, but only in older (L1PA7-L1PA3) subfamilies (Castro-Diaz et al., 2014). Similarly in hESCs, the KZNF protein, ZNF93, only repressed L1 expression in the older L1PA6-L1PA3 (Jacobs et al., 2014). These data highlight the evolutionary “arms-race” occurring within our cells and it may be that a KZNF protein(s) has not yet evolved a response to the active L1Hs subfamily.

Small RNA based interference

Small RNA based interference (RNAi) mechanisms that regulate gene expression are relatively recent discoveries (Fire et al., 1998). Several lines of evidence suggest L1 retrotransposition may also be regulated by RNAi based mechanisms. RNAi relies on short RNA sequences that are loaded onto Argonaute proteins that form an

RNA-induced silencing complex (RISC). RISC and enzymatic proteins associated with the piRNA pathway can subsequently target complementary RNAs for degradation and/or translation inhibition (Wilson and Doudna, 2013).

The P-element induced wimpy testis (PIWI) proteins are Argonaute family members specific to the germline and were first identified in fruit fly (Cox et al., 2000). PIWI proteins bind with PIWI-interacting RNAs (piRNAs) that are ~ 28-31 nucleotides in length (Ishizu et al., 2011; Siomi et al., 2011). The piRNA induced silencing complex (piRISC) can both degrade TE mRNA as well as target TE genomic loci to silence transcription (Aravin et al., 2007; Ishizu et al., 2011; Siomi et al., 2011) (Figure 1.6). Recent observations in fruit fly suggest components of the piRNA pathway recruit proteins that actively enforce transcriptional silencing (Yu et al., 2015).

In mouse testis, loss of the mouse PIWI protein, MILI (Piwi-Like RNA-Mediated Gene Silencing 2), results in an increase of endogenous L1 and IAP expression (Aravin et al., 2007), and a concomitant de-methylation of L1 promoter sequences (Aravin et al., 2008; Aravin et al., 2007) (Aravin et al., 2007) (Figure 1.6). Ablation of MILI in mouse testis leads to spermatogenic arrest and an increase in the expression of L1 ORF1p (Di Giacomo et al., 2013). Recent reports suggest cooperation between piRNA-, DNA methylation-, and histone methylation-mediated repression of endogenous L1 expression in mouse spermatogonia (Di Giacomo et al., 2013; Di Giacomo et al., 2014).

Small interfering RNAs (siRNAs) are ~21 nucleotides in length and represent another class of small RNAs (Wilson and Doudna, 2013). Dicer-dependent siRNAs have been demonstrated to silence plant, invertebrate, and perhaps mouse transposons (Castel and Martienssen, 2013). In mice, there is evidence that siRNAs in oocytes may be derived from L1 and LTR-retrotransposons (Watanabe et al., 2008). Additionally, siRNAs originating from antisense transcription of the L1 5'UTR are postulated to target L1 transcripts in human cultured cells (Yang and Kazazian, 2006). Intriguingly, L1 hypo-methylation and increased L1 expression was observed in a breast cancer cell line upon depletion of Dicer (Chen et al., 2012). Additional experimentation is required to understand how RNAi mechanistically modulates L1 retrotransposition in humans;

however, accumulated evidence suggest a possible role for small RNA mediated L1 repression.

Splicing and Pre-mature polyadenylation

Post-transcriptional processing events including, splicing and/or polyadenylation, are canonical cellular processes that occur on almost every protein coding RNA. As splicing results in the removal of RNA sequence, it was surprising to find that some L1 RNAs undergo splicing (Belancio, 2011; Belancio et al., 2006; Belancio et al., 2008) (Figure 1.6 and Chapter 2). A subset of the potential splice donor (SD), and splice acceptor (SA) sites in L1 RNA can sometimes be activated (Belancio et al., 2006; Belancio et al., 2008). Loss of any sequence within the L1 is hypothesized to be detrimental to L1 mobilization (Chapter 2), and surprisingly it appears that some of these SD and SA sequences are highly conserved in L1 (Chapter 2).

The 3' end of L1 RNAs harbor a weak poly(A) signal, but L1 RNAs also contain numerous pre-mature poly(A) signals throughout the body of L1 RNA (Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003). Pre-mature polyadenylation is hypothesized to be detrimental to L1 retrotransposition (Perepelitsa-Belancio and Deininger, 2003). Mouse and human L1s transfected into mouse NIH 3T3 cells were subject to pre-mature polyadenylation as observed by Northern blot (Perepelitsa-Belancio and Deininger, 2003). However, NIH 3T3 cells transfected with human L1s containing mutations that ablated poly(A) signals resulted in an increase in full-length L1 RNAs, and a decrease in pre-maturely polyaddenylated RNAs (Perepelitsa-Belancio and Deininger, 2003). There are no instances of pre-maturely retrotransposed polyadenylated L1s in the human genome; however, premature polyadenylated L1s were identified in expressed sequence tags (ESTs) databases (Perepelitsa-Belancio and Deininger, 2003).

Cytoplasmic binding proteins

Studies have demonstrated that cytoplasmic proteins associate with the L1 RNP (Dai et al., 2012; Goodier et al., 2013; Goodier et al., 2015; Goodier et al., 2007; Moldovan and Moran, 2015; Taylor et al., 2013) (Figure 1.6). The functions of these proteins are varied and have different effects on L1 RNP biogenesis and L1

retrotransposition. Processing bodies (PBs) and stress granules (SGs) are cytoplasmic aggregations of proteins and mRNAs that affect mRNA decay and protein translation (Decker and Parker, 2012). PBs are present in unstressed cells; however, their quantity is increased in response to stress (Decker and Parker, 2012; Kedersha et al., 2005). SGs are primarily induced from various types of cellular stress (Decker and Parker, 2012; Kedersha et al., 2005) and are thought to sequester mRNAs until a decision is made to translate, sequester, or shuttle mRNAs to PBs for degradation (Kedersha et al., 2005).

Components of the L1 RNP have been demonstrated to associate with SG markers using immunofluorescence, immunoprecipitation followed by tandem mass spectrometry, and western blot experiments (Dai et al., 2012; Goodier, 2016; Goodier et al., 2013; Goodier et al., 2015; Goodier et al., 2007; Moldovan and Moran, 2015; Taylor et al., 2013). The zinc-finger antiviral protein (ZAP) associates both with stress granules and the L1 RNP (Goodier, 2016; Goodier et al., 2015; Moldovan and Moran, 2015) and when overexpressed decreases L1 retrotransposition in cell culture based assays (Goodier et al., 2015; Moldovan and Moran, 2015). The mechanism of action remains unclear, but ZAP could either inhibit L1 mRNA translation or initiate destruction of the mRNA molecule (Moldovan and Moran, 2015).

A component of PBs and SGs, 5’-3’ exonuclease XRN1, degrades mRNAs and was loosely associated with L1 ORF1p in immunofluorescence experiments (Goodier et al., 2007). Largely, L1 proteins and RNA are not directly associated with PBs; however, it has been suggested that the L1 RNP may first associate with stress granules, and then interact with PBs to facilitate degradation (Goodier et al., 2007). It has also been suggested that autophagosomes can target PBs or SGs containing L1 RNP components. Experiments using siRNA-mediated knockdown of the Autophagy Protein 5 (ATG5) in human cells resulted in an increase of endogenous L1 and Alu RNA as well as increased L1 retrotransposition in a cell culture assay (Guo et al., 2014). Additional experimentation is necessary to determine the role of SGs and PBs in disrupting the L1 retrotransposition cycle (Goodier et al., 2013; Goodier et al., 2015; Goodier et al., 2007; Moldovan and Moran, 2015).

Antiviral proteins and integration inhibition

Antiviral proteins usually targeting specific viral components to inhibit viral replication (Goff, 2004). Given that L1 is an ancient endogenous parasite, it is not surprising that some proteins that inhibit various infectious viruses also inhibit L1 retrotransposition (Burton et al., 1986; Goodier, 2016; Lander et al., 2001; Moldovan and Moran, 2015; Sawyer et al., 2004) (Figure 1.6). For example, overexpression of the Moloney Leukemia Virus 10 protein (MOV10), an RNA helicase, inhibits L1 and Alu retrotransposition (Goodier et al., 2012, 2013; Moldovan and Moran, 2015) and retrovirus replication (Wang et al., 2010) in cell culture assays. The mechanism of MOV10 is unknown but it may bind L1 RNA and inhibit translation and/or ORF2p RT activity (Goodier et al., 2012, 2013; Moldovan and Moran, 2015).

SAM domain and HD domain-containing protein 1 (SAMHD1) has been shown to repress HIV-1 infection and L1 retrotransposition through different mechanisms. In immune cells infected by HIV-1, SAMHD1 restricts retroviral replication by depleting cellular levels of dNTPs as well as by acting as an RNase to destroy the viral RNA (Ballana and Este, 2015; White et al., 2014). Contrastingly, overexpression of SAMHD1 leads to sequestration of the L1 RNP in SGs (Hu et al., 2015). Evidence suggests SAMHD1 overexpression may inhibit L1 retrotransposition by both depleting cellular levels of dNTPs, and/or by somehow disrupting ORF2p RT activity (Zhao et al., 2013). Thus, it seems clear that some proteins evolved to inhibit L1 and viral replication through different mechanisms, perhaps suggesting that the protein arose to combat one particle, and subsequently evolved an additional function to combat both particles.

A family of cytidine deaminases has been shown to restrict HIV-1 infection as well as L1 and Alu retrotransposition. Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3, A3) has undergone six duplications in the primate lineage yielding seven members (A3A, A3B, A3C, A3D/E, A3F, A3G, A3H) (Koito and Ikeda, 2013; Lindic et al., 2013; Sawyer et al., 2004). A3 proteins generally function by deaminating single-strand DNA substrates, or perhaps RNA, leading to cytidine to uridine transition mutations (Chiu and Greene, 2008) or by directly blocking RT activity (Chiu and Greene, 2008; Koito and Ikeda, 2013). The overexpression of A3A (Bogerd et

al., 2006; Muckenfuss et al., 2006; Richardson et al., 2015; Richardson et al., 2014; Stenglein and Harris, 2006), A3B (Bogerd et al., 2006; Muckenfuss et al., 2006; Stenglein and Harris, 2006), A3C (Muckenfuss et al., 2006), and A3F (Muckenfuss et al., 2006; Stenglein and Harris, 2006) proteins are able to restrict L1 retrotransposition in a cell culture model. A3G blocks replication of HIV-1 (Sheehy et al., 2002) and overexpression inhibits Alu retrotransposition, but not L1 retrotransposition (Hulme et al., 2007).

The mechanism of APOBEC inhibition of L1 and Alu retrotransposition is varied. A3B and A3C catalytic mutants are able to inhibit L1 retrotransposition; thus, they may act as a physical barrier to RT (Bogerd et al., 2006). Similarly, the Alu specific A3G protein also likely does not require deaminase activity suggesting its mechanism of repression is sequestration of Alu RNA into high molecular weight cellular complexes. (Chiu et al., 2006). In contrast, it has recently been demonstrated that A3A restricts L1 retrotransposition, in part, by deaminating transiently exposed single-strand L1 cDNA exposed during TPRT, which could subsequently lead to cDNA cleavage and L1 5' truncation (Richardson et al., 2014). Alternatively, A3A could deaminate exposed single stranded genomic DNA that would ultimately lead to the loss of the entire L1 cDNA (Richardson et al., 2014).

Locations of LINE-1 Retrotransposition

Efficient L1 propagation requires that it retrotranspose in either the germline or very early in embryogenesis. It is unclear how frequently offspring harbor a new L1 insertion, but it has been hypothesized that one out of 20 to one out of 200 human live births contain a *de novo* L1 insertion (Cordaux and Batzer, 2009).

Experimental evidence of retrotransposition in the germ line has been observed in mice. A transgenic mouse containing a tagged L1 expression construct expressing an EGFP indicator cassette was shown to retrotranspose in both male testis and female ovaries (Ostertag et al., 2002). Additionally, two offspring from these transgenic mice exhibited a *de novo* retrotransposition event and importantly did *not* contain the transgene from the parent (Ostertag et al., 2002). These results suggest that retrotransposition occurred in the male germ line prior to meiosis II, and was not the

result of retrotransposition early in offspring embryogenesis (Athanikar et al., 2002). In human testis, expression of both ORF1p and ORF2p has been observed (Ergun et al., 2004) and evidence suggests human oocytes express endogenous L1 RNA and support retrotransposition of an engineered L1 (Georgiou et al., 2009).

Retrotransposition can occur in early embryogenesis. Depending on when the retrotransposition event occurs, the new insertion can contribute *only* to somatic cells, *only* to germ cells, or to *both* somatic and germ cells; the latter two cases representing heritable insertions (Richardson et al., 2015). Cell culture experiments using both hESCs (Garcia-Perez et al., 2007b) and ECs (Garcia-Perez et al., 2010; Skowronski et al., 1988; Skowronski and Singer, 1985) demonstrated that endogenous L1 RNA and proteins are highly expressed in these cells and that these cell types support retrotransposition of an engineered L1 (Garcia-Perez et al., 2007b; Garcia-Perez et al., 2010). Additionally, *de novo* retrotransposition events can occur in both human induced pluripotent stem cells (hiPSCs) and hESCs (Klawitter et al., 2016; Wissing et al., 2011). These experiments show that in principle, L1 can retrotranspose early in human embryonic development.

A direct example of endogenous L1 retrotransposition in the early embryo was gleaned from a male patient with X-linked choriodermia. He harbored an L1 insertion in the *CMH* gene (van den Hurk et al., 2003). The mother of the patient was a somatic mosaic for the L1 insertion, suggesting that the insertion into the *CMH* gene occurred early enough during her development to give rise to both somatic and germline cells (van den Hurk et al., 2007). Of the patients' two sisters, only one had the L1 insertion into the *CMH* gene, even though they both shared the same maternal haplotype block, further supporting the notion that this L1 insertion occurred in the mother, early in her embryonic development (van den Hurk et al., 2007).

It was originally assumed that most somatic tissues were unable to support retrotransposition, as these represent 'dead ends' for L1 inheritance. This notion was radically overturned when it was found that an engineered L1 retrotransposed in cultured rat neuronal precursor cells (NPCs), as well as in various brain tissues of transgenic mice expressing an engineered L1 (Muotri et al., 2005). Additional data

suggested human neural progenitor cells (NPCs) from fetal brain, as well as NPC derived hESCs supported retrotransposition of an engineered L1 (Coufal et al., 2009). Quantitative PCR (qPCR) analysis of genomic L1 insertions supported the notion that neuronal tissues harbor more endogenous retrotransposition events than heart or liver tissues from the same individuals (Coufal et al., 2009).

Next generation sequencing (NGS) of L1, SVA, and Alu enriched libraries revealed *de novo* insertions from all three retrotransposons occurred in the brain, but were not present in the germline from the same individuals (Baillie et al., 2011). Single neuron sequencing of hippocampal neurons from four individuals suggests that, at least in this tissue, each neuron harbors ~13 somatic insertions when compared to matched liver samples (Upton et al., 2015). Single cell analysis of cortical neurons shows that, at least in this tissue, somatic retrotransposition occurs at a frequency of ~1 insertion per neuron (Evrony et al., 2012; Evrony et al., 2015). Despite potential cell-type differences in L1 retrotransposition activity, it is clear that retrotransposition occurs somatically in the human brain. Given the number of neurons in the human brain, it is intriguing to consider the possibility that there may be hundreds of millions of somatic retrotransposition events in a single brain. Future studies are needed to elucidate the frequency of retrotransposition in the human brain and the effect it may have on neuronal biology.

Conclusion

Despite Barbara McClintock's astute and pioneering observations (McClintock, 1950, 1951), it is unlikely even she could predict the scope in which mobile genetic elements affect humans. Her first suspicion that genomes are not static, but are in fact constantly changing and mutable entities, set the stage for more than 60 years of research and discovery.

Though the majorities of TEs are molecular fossils and are incapable of mobility, they still affect the genome by serving as fodder for cooption of sequences for gene regulation, mediating gene birth, and acting as arbiters of genetic variation. Yet, each individual cell harbors nearly 100 potentially active L1s, and even more Alu elements, which continue to diversify our genomes. Their existence is measured in the hundreds

of millions of years, and their reverse transcriptase domain may harken back to the transition from the RNA to the DNA world. Indeed, reverse transcriptase likely was instrumental in constructing the genetic code for life. There is a monumental amount of work yet to be done and the most exciting aspects L1 and Alu biology remain to be uncovered.

Future work should focus on TE role in disease and cancer progression. Though they may rarely be responsible for activating the cancer phenotype, it is very likely that their activity can drive cancer progression. TEs could be utilized as a model to identify antiviral proteins as a safer alternative than screening proteins using HIV. TEs continue to be an invaluable tool in understanding the genetic flow in populations and in genotyping individuals. Finally, mounting evidence suggests TE mobilization may play a role in neuronal plasticity and perhaps contribute to neuronal disease phenotypes. New technologies in gene editing and deep sequencing are going to be instrumental in investigating the budding field of neuronal TE biology.

Figure 1.1: Sequence composition of the human genome.
Completion of the human genome uncovered that Transposable Elements (TEs) (blue and dark blue) comprise nearly 50% of genomic DNA whereas protein coding genes (red) comprise only ~1.5% of genomic DNA (Lander et al., 2001). Retrotransposons alone constitute ~42% of genomic DNA (blue). Of the retrotransposons, non-LTR retrotransposons are the most abundant. Ancient, inactive LINE sequences (in mustard) represent ~6% of genomic DNA. Autonomous LINE-1 (green) sequences constitute ~17% of genomic DNA and their non-autonomous counterpart, SINE (light green) represent ~11% of genomic DNA. Autonomous and non-autonomous LTR-retrotransposons comprise ~8% (in orange). The least abundant TE, DNA transposons (dark blue) comprise twice the fraction of the human genome that protein-coding genes do.

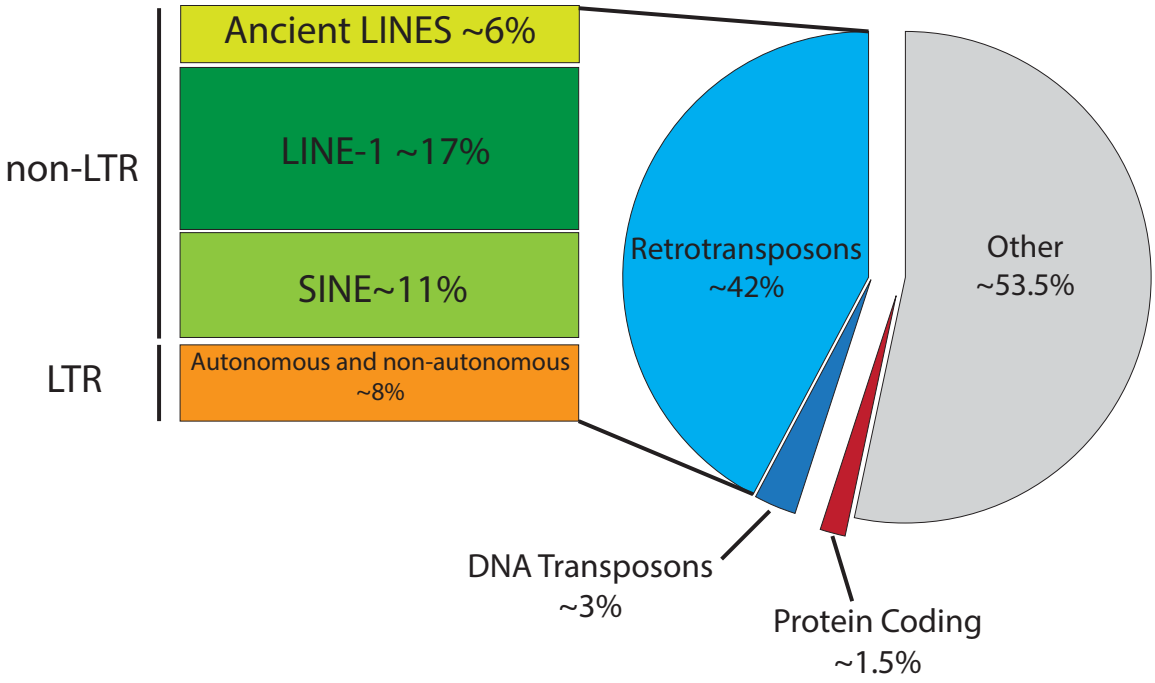


Figure 1.1: Sequence composition of the human genome.

Figure 1.2: Diagrams of transposon and transposon-derived sequences in the human genome.

Representative Transposable Elements (TEs) in the human genome. A) A full-length Long Terminal Repeat (LTR) HERV-K is depicted. LTR retrotransposons range from 6-8 kilobases (kb) in length. They encode a *gag* protein (mustard rectangle), followed by a *pol* protein (navy blue rectangle), and end with an *env* protein (orange rectangle). LTR retrotransposon coding sequences are flanked by long terminal repeats (grey boxes). The 5' LTR contains polII promoter activity (black arrow) and the 3' LTR harbors a polyadenylation sequence (P(A)). The entire element is flanked by ~4-6 base pair (bp) target site duplications (black arrows flanking element. (B-D Non-LTR Retrotransposons) B) The autonomous Non-LTR retrotransposon LINE-1 is depicted. LINE-1s are ~6 kb and begin with a 5'UTR (grey rectangle) with sense and antisense polII promoter activity (forward and backward facing arrows respectively). Following the 5'UTR is Open Reading Frame 1 (ORF1, maize rectangle), a short intergenic spacer (white box), and Open Reading Frame 2 (ORF2, blue rectangle). The sequence ends with a 3'UTR (dark grey rectangle) and a polyadenosine tail (A_N). The entire element is flanked by 7-20 bp target site duplications (black arrows flanking element) C) The non-autonomous non-LTR retrotransposon Alu is depicted. Alu elements are ~280 bp. Alus do not encode proteins and are derived from 7SL RNA. A poly-adenosine rich tract (dark green rectangle in center) separates the left and right 7SL-derived monomers (green rectangle). The left monomer contains an A and B box (vertical black lines) that contains polIII promoter activity (black arrow). Alu elements end in a polyadenosine tail (A_N). The entire element is flanked by 7-20 bp target site duplications (black arrows flanking element). D) Processed pseudogenes are not transposons, but are processed (spliced and polyadenylated) mRNAs derived from protein coding genes that have been mobilized by LINE-1 proteins. Adjacent purple rectangles are spliced exons and grey rectangles at either end are the 5' and 3' untranslated regions. As they are derived from genes their length is variable. Processed pseudogenes end in a polyadenylated tail (A_N) and are flanked by 7-20 bp target site duplications (black arrows flanking processed pseudogene. E) DNA transposons are 2-3 kb and encode a transposase gene (blue rectangle) which is flanked by terminal inverted repeats (black triangle in white box). Terminal inverted repeats can have promoter activity (black arrow). The entire element is flanked by 4-6 bp target site duplications (black arrows flanking element). (See main text for references)

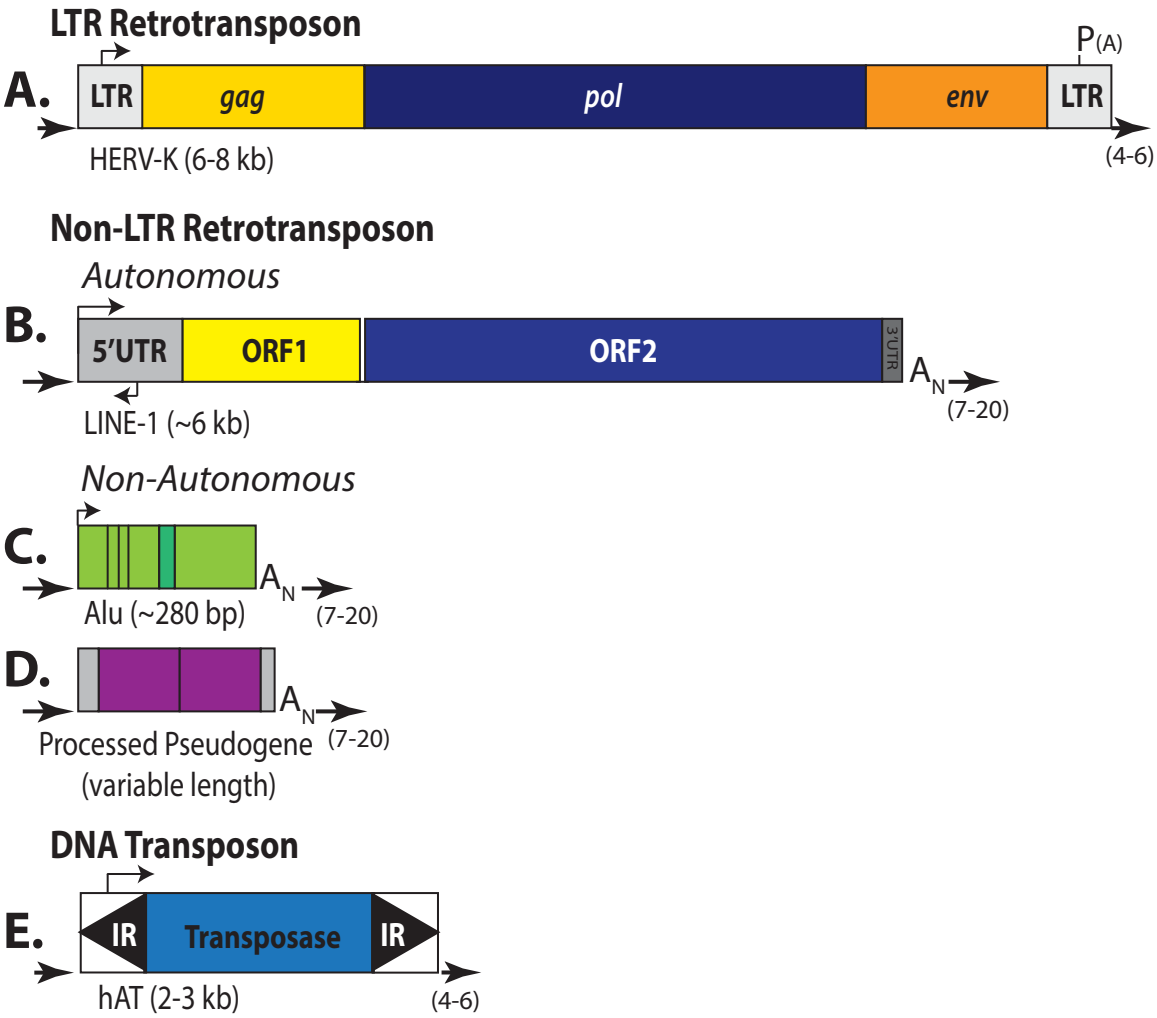


Figure 1.2: Diagrams of transposon and transposon-derived sequences in the human genome.

Figure 1.3: Timeline of LINE-1 evolution in the context of global evolutionary events.

The top of the figure lists selected evolutionary events in the mammalian lineage. The depicted timeline is measured in million years ago (MYA). Arrows and short descriptions of events are placed in relation to when they occurred. Below the timeline is a cluster of six long rectangles. The rectangles indicate the three major LINE families. LINE-3 (L3) is shown on the top and is indicated with a black to grey rectangle that ends in a slope. The slope depicts the gradual decrease of genomic amplification. Note that even though these elements ceased amplifying, fossilized copies are still present in mammalian genomes. LINE-2 (L2) is depicted by a red orange bar that fades to a yellow slope. Beneath L2 is a cluster of four rectangles indicated with a black bracket. These are all members of the LINE-1 family. L1M and L1PB subfamilies are shown as light blue rectangles that fade to white and ending with a slope. The second to the bottom rectangle (solid blue) depicts the L1PA subfamily. Note linear subfamily succession is indicated by white sloping lines and L1 subfamily label (*e.g.*, PA8) The final dark blue rectangle depicts the emergence of the human specific LINE-1 (L1Hs). It is a part of the L1PA subfamily and is separated on this diagram for clarity. Beneath the long rectangles is another timeline on the same scale as the top timeline. This timeline depicts the approximate timing of events that occurred during LINE-1 evolution. Arrows and short descriptions of events are placed in relation to when they occurred. Below the time line are two red boxes that indicate increased rates of LINE-1 amplification. Finally two brackets indicate when, during evolution, multiple LINE-1 subfamilies were amplifying and evolving simultaneously (green bracket, radiating evolution), and when linear evolution of LINE-1 occurs (orange bracket, linear evolution).

Figure 1.3: Timeline of LINE-1 evolution in the context of global evolutionary events.

Figure 1.4: Evolution of LINE-1 5'UTR sequences

The evolution of 5'UTR sequences is depicted. The arrow depicts evolutionary time with more ancient elements placed at the bottom. The green bar to the right of the arrow indicates the period of time when LINE-1s underwent radiating evolution and the orange bar indicates linear evolution (See Figure 1.3 and text for more details). Note that the ORF1 and ORF2 reading frames are kept constant in this diagram, however they too underwent changes over evolutionary history. The bottom LINE-1 depicts a hypothetical 5'UTR (dark grey rectangle with two dark brown lines). This progenitor 5'UTR gave rise to the M (light grey rectangle with two dark brown lines) family of 5'UTRs as well as a precursor(s) to the PB and PA (dark grey rectangle with single dark brown line) subfamily of elements. The PA and PB precursor 5'UTR gave rise to the PB (grey rectangle with tan line) subfamily and PA (grey rectangle with light brown line) subfamilies of 5'UTRs. The M and PA and PB subfamilies amplified concomitantly until only the PA subfamily was dominant. Members of the PA subfamily were also amplifying concomitantly until ~45 million years ago a single lineage (PA8, light grey 5'UTR) was the sole amplifying member. Since then, only a single subfamily has amplified at any given time. Note: the 5'UTR has tended to shorten over time. The 5'UTR depicted at the top of the diagram is the human specific L1Hs 5'UTR. (See text for references)

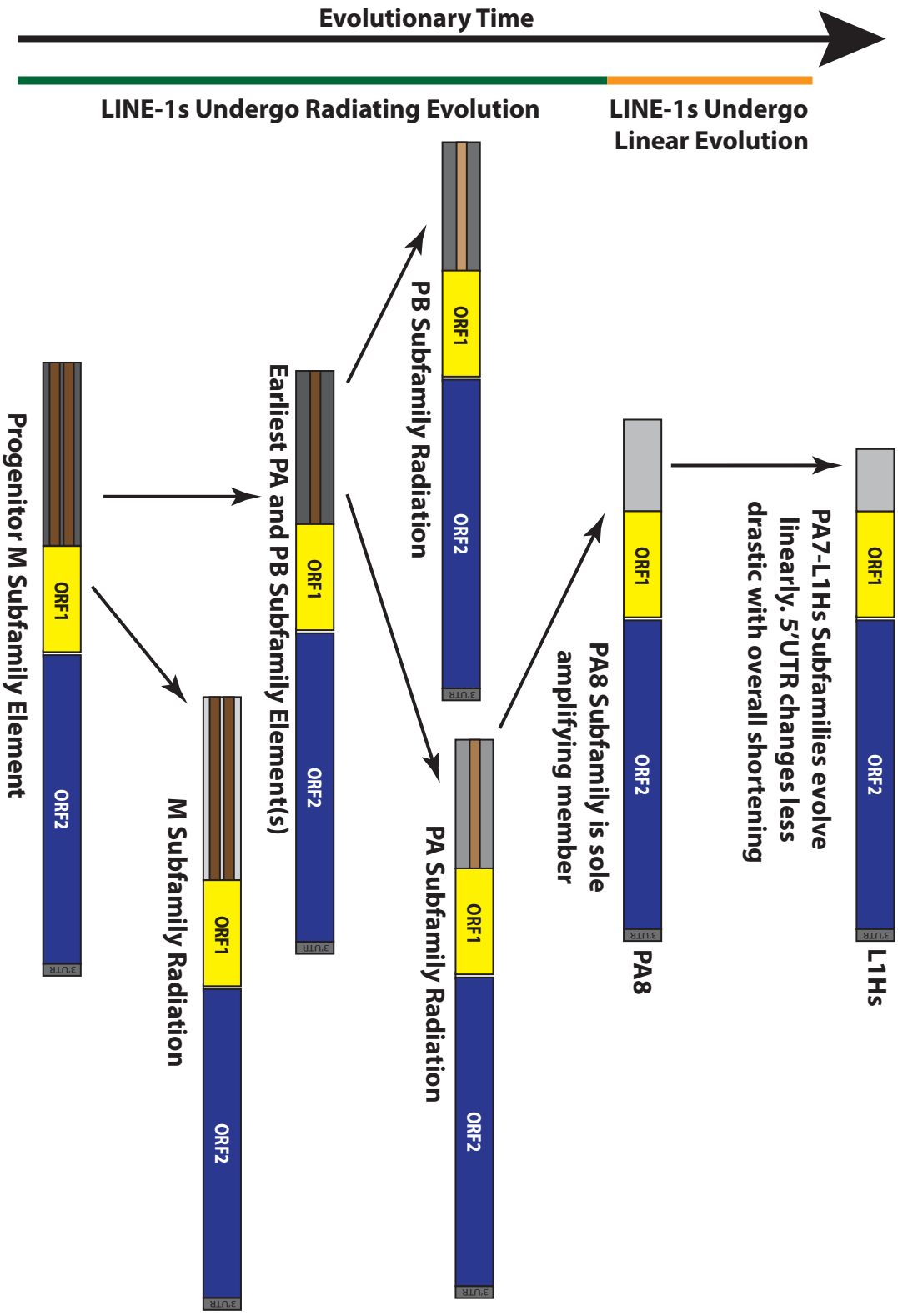


Figure 1.4: Evolution of LINE-1 5'UTR sequences

Figure 1.5: The LINE-1 retrotransposition cycle.
LINE-1 is transcribed (wavy line; black depicts 5' and 3' UTRs, maize depicts ORF1, blue depicts ORF2, backwards "G" depicts 5'methyl guanosine, and A_N depicts polyadenylation) from a genomic location (dark brown rectangle on red chromosome) and exported to the cytoplasm. Both ORFs are translated and the proteins (maize circles depict ORF1p, blue oval depicts ORF2p) bind back to the L1 mRNA from which they were translated (termed *cis*-preference). A cellular mRNA or SINE element (green wavy line) can 'hijack' an ORF2p molecule and use it for its own retrotransposition. Components of the resulting ribonucleoprotein particle (RNP) are imported into the nucleus where ORF2p mediates insertion into a new genomic location (dark green rectangle on green chromosome) by a process termed target-site primed reverse transcription (TPRT; shown as a blow up, black lines indicated top and bottom strand of genomic DNA). ORF2p EN domain initiates TPRT with a single-stranded endonucleolytic nick at a thymidine-rich sequence (*e.g.,* 5'-TTTT/A-3'), where the (/) denotes the siscile phosphate (Cost et al., 2002; Feng et al., 1996; Morrish et al., 2002, Luan et al., 1993) The result of the endonucleolytic nick is exposure of a reactive 3' hydroxyl group, which ORF2p RT activity uses as a primer to initiate minus (-) strand L1 cDNA synthesis (Cost and Boeke, 1998; Feng et al., 1996) (Figure 1.5). Completion of second strand cDNA requires elucidation. (See text for more details)

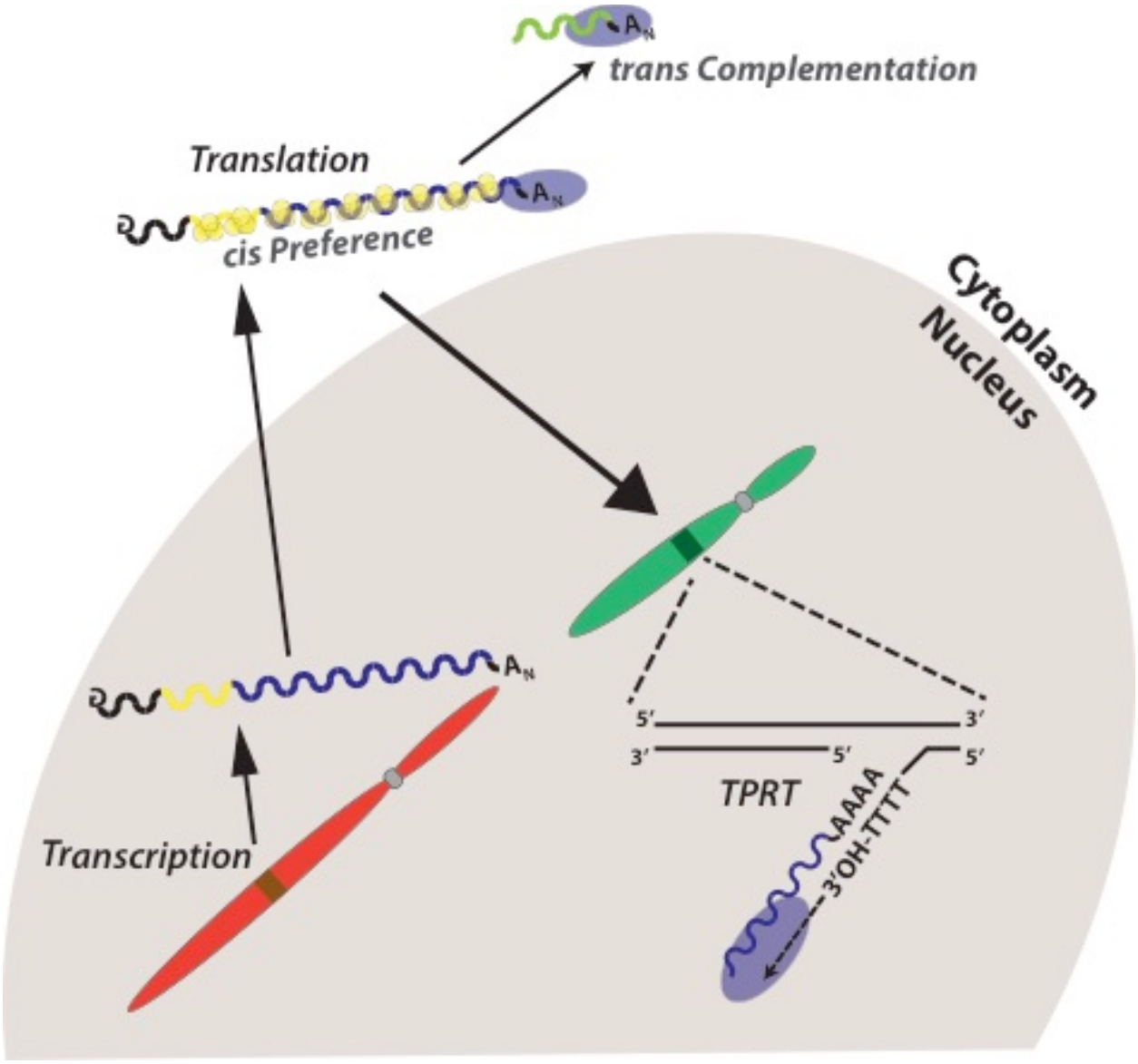


Figure 1.5: The LINE-1 retrotransposition cycle.

Figure 1.6: Mechanisms that inhibit LINE-1 retrotransposition.
The schematic shows a single round of LINE-1 retrotransposition (See Figure 1.5). Red boxes contain mechanisms or proteins that inhibit retrotransposition. The repressor bars emanating from each box indicate where in the retrotransposition cycle they act. Note that the piRNA/siRNA box has an activating red arrow pointing to the box containing DNA methylation. The activity of piRNAs can also influence genomic methylation of LINE-1s (see text).

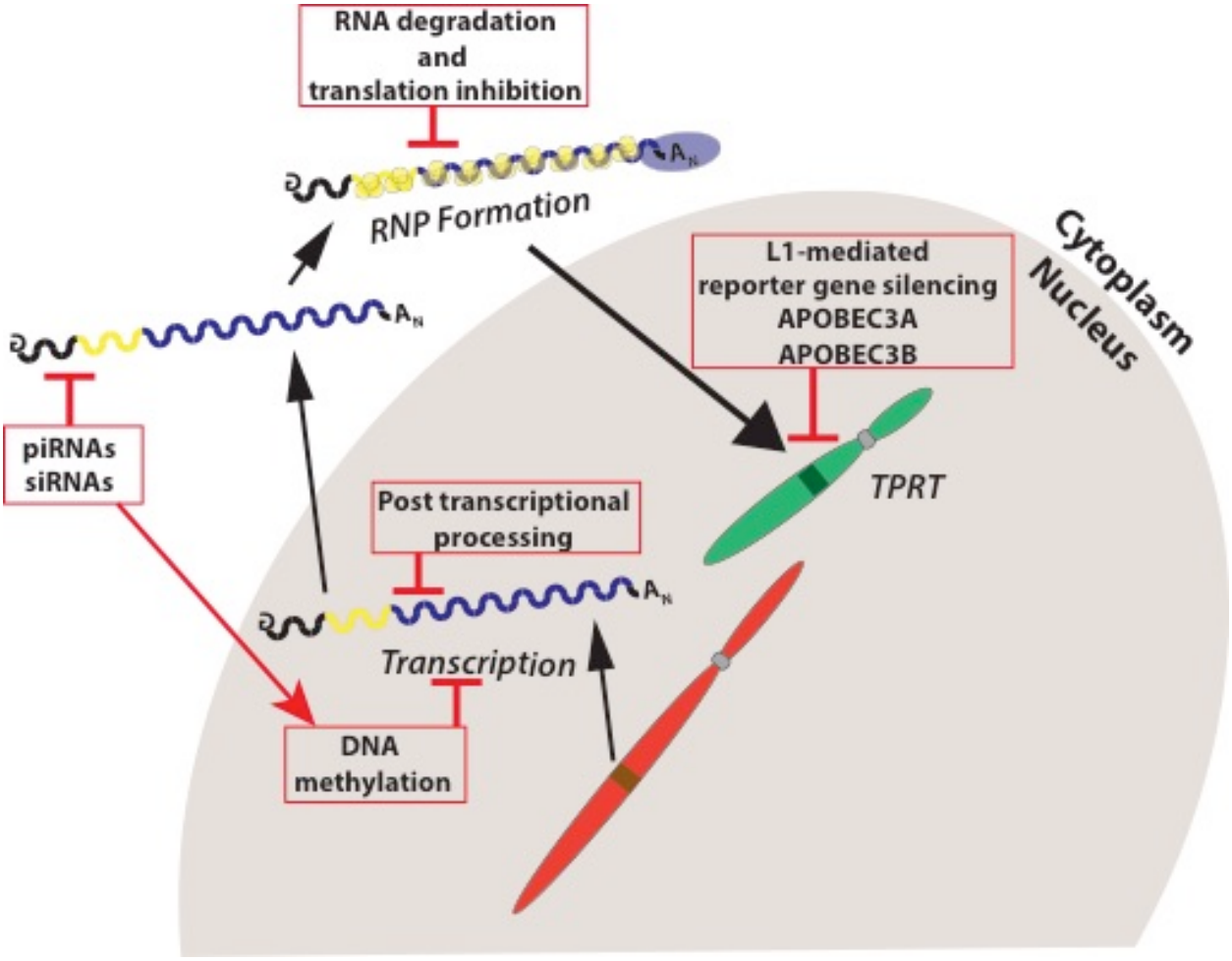


Figure 1.6: Mechanisms that inhibit LINE-1 retrotransposition

References

Ahl, V., Keller, H., Schmidt, S., and Weichenrieder, O. (2015). Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. Molecular cell *60*, 715-727.

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J.V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes & development *20*, 210-224.

Alves, G., Tatro, A., and Fanning, T. (1996). Differential methylation of human LINE-1 retrotransposons in malignant cells. Gene *176*, 39-44.

Ando, S., Yang, H., Nozaki, N., Okazaki, T., and Yoda, K. (2002). CENP-A, -B, and -C chromatin complex that contains the I-type alpha-satellite array constitutes the prekinetochore in HeLa cells. Molecular and cellular biology *22*, 2229-2241.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Molecular cell *31*, 785-799.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744-747.

Athanikar, J.N., Badge, R.M., and Moran, J.V. (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic acids research *32*, 3846-3855.

Athanikar, J.N., Morrish, T.A., and Moran, J.V. (2002). Of man in mice. Nature genetics *32*, 562-563.

Awano, H., Malueka, R.G., Yagi, M., Okizuka, Y., Takeshima, Y., and Matsuo, M. (2010). Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. J Hum Genet *55*, 785-790.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. American journal of human genetics *72*, 823-838.

Bailey, J.A., Liu, G., and Eichler, E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. American journal of human genetics *73*, 823-834.

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M.*, et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Ballana, E., and Este, J.A. (2015). SAMHD1: at the crossroads of cell proliferation, immune responses, and virus restriction. Trends in microbiology *23*, 680-692.

Bannert, N., and Kurth, R. (2006). The evolutionary dynamics of human endogenous retroviral families. Annual review of genomics and human genetics *7*, 149-173.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Herault, Y., Guillou, F., and Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science *354*, 909-912.

Basame, S., Wai-lun Li, P., Howard, G., Branciforte, D., Keller, D., and Martin, S.L. (2006). Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition. Journal of molecular biology *357*, 351-357.

Beauregard, A., Curcio, M.J., and Belfort, M. (2008). The take and give between retrotransposable elements and their hosts. Annual review of genetics *42*, 587-617.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159-1170.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 elements in structural variation and disease. Annual review of genomics and human genetics *12*, 187-215.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Human molecular genetics *2*, 1697-1702.

Belancio, V.P. (2011). Importance of RNA analysis in interpretation of reporter gene expression data. Anal Biochem *417*, 159-161.

Belancio, V.P., Hedges, D.J., and Deininger, P. (2006). LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic acids research *34*, 1512-1521.

Belancio, V.P., Roy-Engel, A.M., and Deininger, P. (2008). The impact of multiple splice sites in human L1 elements. Gene *411*, 38-45.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active Alu retrotransposons in the human genome. Genome research *18*, 1875-1883.

Birtle, Z., and Ponting, C.P. (2006). Meisetz and the birth of the KRAB motif. Bioinformatics *22*, 2841-2845.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491-500.

Bogerd, H.P., Wiegand, H.L., Hulme, A.E., Garcia-Perez, J.L., O'Shea, K.S., Moran, J.V., and Cullen, B.R. (2006). Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proceedings of the National Academy of Sciences of the United States of America *103*, 8780-8785.

Boissinot, S., Chevret, P., and Furano, A.V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Molecular biology and evolution *17*, 915-928.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., and Furano, A.V. (2004a). The insertional history of an active family of L1 retrotransposons in humans. Genome research *14*, 1221-1231.

Boissinot, S., and Furano, A.V. (2005). The recent evolution of human L1 retrotransposons. Cytogenetic and genome research *110*, 402-406.

Boissinot, S., Roos, C., and Furano, A.V. (2004b). Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. Journal of molecular evolution *58*, 122-130.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96-99.

Bratthauer, G.L., and Fanning, T.G. (1993). LINE-1 retrotransposon expression in pediatric germ cell tumors. Cancer *71*, 2383-2386.

Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science *165*, 349-357.

Brouha, B., Meischl, C., Ostertag, E., de Boer, M., Zhang, Y., Neijens, H., Roos, D., and Kazazian, H.H., Jr. (2002). Evidence consistent with human L1 retrotransposition in maternal meiosis I. American journal of human genetics *71*, 327-336.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences of the United States of America *100*, 5280-5285.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods *10*, 1213-1218.

Burton, F.H., Loeb, D.D., Voliva, C.F., Martin, S.L., Edgell, M.H., and Hutchison, C.A., 3rd (1986). Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. Journal of molecular biology *187*, 291-304.

Burwinkel, B., and Kilimann, M.W. (1998). Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. Journal of molecular biology *277*, 513-517.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. (2002). A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics *80*, 402-406.

Callahan, K.E., Hickman, A.B., Jones, C.E., Ghirlando, R., and Furano, A.V. (2012). Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. Nucleic acids research *40*, 813-827.

Carroll, L., Foreman, M., and Sterling Publishing Company. (2004). Alice's adventures in Wonderland (New York: Sterling Pub.).

Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E., and Fraser, M.J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology *172*, 156-169.

Casola, C., Hucks, D., and Feschotte, C. (2008). Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Molecular biology and evolution *25*, 29-41.

Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature reviews Genetics *14*, 100-112.

Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., Duc, J., Jang, S.M., Turelli, P., and Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. Genes & development *28*, 1397-1409.

Chen, L., Dahlstrom, J.E., Lee, S.H., and Rangasamy, D. (2012). Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. Epigenetics *7*, 758-771.

Chiu, Y.L., and Greene, W.C. (2008). The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annual review of immunology *26*, 317-353.

Chiu, Y.L., Witkowska, H.E., Hall, S.C., Santiago, M., Soros, V.B., Esnault, C., Heidmann, T., and Greene, W.C. (2006). High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. Proceedings of the National Academy of Sciences of the United States of America *103*, 15588-15593.

Christensen, S.M., and Eickbush, T.H. (2005). R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. Molecular and cellular biology *25*, 6617-6628.

Christian, C.M., Sokolowski, M., deHaro, D., Kines, K.J., and Belancio, V.P. (2017). Involvement of Conserved Amino Acids in the C-Terminal Region of LINE-1 ORF2p in Retrotransposition. Genetics.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083-1087.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nature reviews Genetics *18*, 71-86.

Clements, A.P., and Singer, M.F. (1998). The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. Nucleic acids research *26*, 3528-3535.

Conrad, B., Weissmahr, R.N., Boni, J., Arcari, R., Schupbach, J., and Mach, B. (1997). A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. Cell *90*, 303-313.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews Genetics *10*, 691-703.

Cordonnier, A., Casella, J.F., and Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. Journal of virology *69*, 5890-5897.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell reports *2*, 1025-1035.

Cost, G.J., and Boeke, J.D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry *37*, 18081-18093.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. The EMBO journal *21*, 5899-5910.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

Cox, D.N., Chao, A., and Lin, H. (2000). piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development *127*, 503-514.

Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A., Rice, P.A., and Sandmeyer, S. (2015). Mobile DNA III.

Dai, L.X., Taylor, M.S., O'Donnell, K.A., and Boeke, J.D. (2012). Poly(A) Binding Protein C1 Is Essential for Efficient L1 Retrotransposition and Affects L1 RNP Formation. Molecular and cellular biology *32*, 4323-4336.

Dawkins, R. (1976). The selfish gene (New York: Oxford University Press).

de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics *7*, e1002384.

Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harbor perspectives in biology *4*, a012286.

Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T., and Schmid, C.W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. Journal of molecular biology *151*, 17-33.

Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J.*, et al.* (2015). Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell *163*, 583-593.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nature genetics *35*, 41-48.

Dewannieux, M., and Heidmann, T. (2005). Role of poly(A) tail length in Alu retrotransposition. Genomics *86*, 378-381.

Di Giacomo, M., Comazzetto, S., Saini, H., De Fazio, S., Carrieri, C., Morgan, M., Vasiliauskaite, L., Benes, V., Enright, A.J., and O'Carroll, D. (2013). Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. Molecular cell *50*, 601-608.

Di Giacomo, M., Comazzetto, S., Sampath, S.C., Sampath, S.C., and O'Carroll, D. (2014). G9a co-suppresses LINE1 elements in spermatogonia. Epigenet Chromatin *7*.

Dmitriev, S.E., Andreev, D.E., Terenin, I.M., Olovnikov, I.A., Prassolov, V.S., Merrick, W.C., and Shatsky, I.N. (2007). Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. Molecular and cellular biology *27*, 4685-4697.

Dombroski, B.A., Feng, Q., Mathias, S.L., Sassaman, D.M., Scott, A.F., Kazazian, H.H., Jr., and Boeke, J.D. (1994). An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Molecular and cellular biology *14*, 4485-4492.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805-1808.

Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature *284*, 601-603.

Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A., Moran, J.V.*, et al.* (2010). Characterization of LINE-1 ribonucleoprotein particles. PLoS genetics *6*.

Doucet, A.J., Wilusz, J.E., Miyoshi, T., Liu, Y., and Moran, J.V. (2015). A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. Molecular cell *60*, 728-741.

Doucet-O'Hare, T.T., Rodic, N., Sharma, R., Darbari, I., Abril, G., Choi, J.A., Young Ahn, J., Cheng, Y., Anders, R.A., Burns, K.H.*, et al.* (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proceedings of the National Academy of Sciences of the United States of America *112*, E4894-4900.

Doudna, J.A., and Rath, V.L. (2002). Structure and function of the eukaryotic ribosome: the next frontier. Cell *109*, 153-156.

Dupressoir, A., Lavialle, C., and Heidmann, T. (2012). From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta *33*, 663-671.

Dupuy, A.J., Jenkins, N.A., and Copeland, N.G. (2006). Sleeping beauty: a novel cancer gene discovery tool. Human molecular genetics *15 Spec No 1*, R75-79.

Eickbush, T.H. (1997). Telomerase and retrotransposons: which came first? Science *277*, 911-912.

Eickbush, T.H., Burke, W.D., Eickbush, D.G., and Lathe, W.C., 3rd (1997). Evolution of R1 and R2 in the rDNA units of the genus Drosophila. Genetica *100*, 49-61.

Ergun, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Stratling, W.H., and Schumann, G.G. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. The Journal of biological chemistry *279*, 27753-27763.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nature genetics *24*, 363-367.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A.*, et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J.*, et al.* (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Ewing, A.D., Gacita, A., Wood, L.D., Ma, F., Xing, D., Kim, M.S., Manda, S.S., Abril, G., Pereira, G., Makohon-Moore, A.*, et al.* (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome research *25*, 1536-1545.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Human molecular genetics *23*, 5866-5878.

Fanning, T., and Singer, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic acids research *15*, 2251-2260.

Fedoroff, N.V. (2012). Presidential address. Transposable elements, epigenetics, and genome evolution. Science *338*, 758-767.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell *87*, 905-916.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annual review of genetics *41*, 331-368.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Fitch, D.H., Bailey, W.J., Tagle, D.A., Goodman, M., Sieu, L., and Slightom, J.L. (1991). Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proceedings of the National Academy of Sciences of the United States of America *88*, 7396-7400.

Galli, U.M., Sauter, M., Lecher, B., Maurer, S., Herbst, H., Roemer, K., and Mueller-Lantzsch, N. (2005). Human endogenous retrovirus rec interferes with germ cell

development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene *24*, 3223-3228.

Garcia-Perez, J.L., Doucet, A.J., Bucheton, A., Moran, J.V., and Gilbert, N. (2007a). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome research *17*, 602-611.

Garcia-Perez, J.L., Marchetto, M.C., Muotri, A.R., Coufal, N.G., Gage, F.H., O'Shea, K.S., and Moran, J.V. (2007b). LINE-1 retrotransposition in human embryonic stem cells. Human molecular genetics *16*, 1569-1577.

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P.*, et al.* (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769-773.

Georgiou, I., Noutsopoulos, D., Dimitriadou, E., Markopoulos, G., Apergi, A., Lazaros, L., Vaxevanoglou, T., Pantos, K., Syrrou, M., and Tzavaras, T. (2009). Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes. Human molecular genetics *18*, 1221-1228.

Gerdes, P., Richardson, S.R., Mager, D.L., and Faulkner, G.J. (2016). Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome biology *17*, 100.

Gilbert, N., Bomar, J.M., Burmeister, M., and Moran, J.V. (2004). Characterization of a mutagenic B1 retrotransposon insertion in the jittery mouse. Human mutation *24*, 9-13.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Molecular and cellular biology *25*, 7780-7795.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315-325.

Goff, S.P. (2004). Retrovirus restriction factors. Molecular cell *16*, 849-859.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annual review of biochemistry *74*, 481-514.

Goodier, J.L. (2016). Restricting retrotransposons: a review. Mobile DNA *7*, 16.

Goodier, J.L., Cheung, L.E., and Kazazian, H.H., Jr. (2012). MOV10 RNA helicase is a potent inhibitor of retrotransposition in cells. PLoS genetics *8*, e1002941.

Goodier, J.L., Cheung, L.E., and Kazazian, H.H., Jr. (2013). Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. Nucleic acids research *41*, 7401-7419.

Goodier, J.L., Mandal, P.K., Zhang, L., and Kazazian, H.H., Jr. (2010). Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. Human molecular genetics *19*, 1712-1725.

Goodier, J.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2000). Transduction of 3'-flanking sequences is common in L1 retrotransposition. Human molecular genetics *9*, 653-657.

Goodier, J.L., Pereira, G.C., Cheung, L.E., Rose, R.J., and Kazazian, H.H., Jr. (2015). The Broad-Spectrum Antiviral Protein ZAP Restricts Human Retrotransposition. PLoS genetics *11*, e1005252.

Goodier, J.L., Zhang, L., Vetter, M.R., and Kazazian, H.H., Jr. (2007). LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. Molecular and cellular biology *27*, 6469-6483.

Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. (1998). Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Molecular phylogenetics and evolution *9*, 585-598.

Grimaldi, G., and Singer, M.F. (1983). Members of the KpnI family of long interspersed repeated sequences join and interrupt alpha-satellite in the monkey genome. Nucleic acids research *11*, 321-338.

Grimaldi, G., Skowronski, J., and Singer, M.F. (1984). Defining the beginning and end of KpnI family segments. The EMBO journal *3*, 1753-1759.

Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y.*, et al.* (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221-225.

Guo, H., Chitiprolu, M., Gagnon, D., Meng, L., Perez-Iratxeta, C., Lagace, D., and Gibbings, D. (2014). Autophagy supports genomic stability by degrading retrotransposon RNA. Nat Commun *5*, 5276.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature *429*, 268-274.

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., and Batzer, M.A. (2008). L1 recombination-associated deletions generate human genomic variation. Proceedings of the National Academy of Sciences of the United States of America *105*, 19366-19371.

Han, K., Sen, S.K., Wang, J., Callinan, P.A., Lee, J., Cordaux, R., Liang, P., and Batzer, M.A. (2005). Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic acids research *33*, 4040-4052.

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., and Kazazian, H.H., Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Human molecular genetics *20*, 3386-3400.

Hancks, D.C., and Kazazian, H.H., Jr. (2016). Roles for retrotransposon insertions in human disease. Mobile DNA *7*, 9.

Hata, K., and Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene *189*, 227-234.

Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature *321*, 625-628.

Havecker, E.R., Gao, X., and Voytas, D.F. (2004). The diversity of LTR retrotransposons. Genome biology *5*, 225.

Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome research *24*, 1053-1063.

Heras, S.R., Macias, S., Plass, M., Fernandez, N., Cano, D., Eyras, E., Garcia-Perez, J.L., and Caceres, J.F. (2013). The Microprocessor controls the activity of mammalian retrotransposons. Nature structural & molecular biology *20*, 1173-1181.

Hohjoh, H., and Singer, M.F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. The EMBO journal *15*, 630-639.

Hohjoh, H., and Singer, M.F. (1997a). Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. Journal of molecular biology *271*, 7-12.

Hohjoh, H., and Singer, M.F. (1997b). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. The EMBO journal *16*, 6034-6043.

Hohn, O., Hanke, K., and Bannert, N. (2013). HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Frontiers in oncology *3*, 246.

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H., Jr. (1994). A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nature genetics *7*, 143-148.

Holmes, S.E., Singer, M.F., and Swergold, G.D. (1992). Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. The Journal of biological chemistry *267*, 19765-19768.

Hormozdiari, F., Konkel, M.K., Prado-Martinez, J., Chiatante, G., Herraez, I.H., Walker, J.A., Nelson, B., Alkan, C., Sudmant, P.H., Huddleston, J.*, et al.* (2013). Rates and patterns of great ape retrotransposition. Proceedings of the National Academy of Sciences of the United States of America *110*, 13457-13462.

Howell, R., and Usdin, K. (1997). The ability to form intrastrand tetraplexes is an evolutionarily conserved feature of the 3' end of L1 retrotransposons. Molecular biology and evolution *14*, 144-155.

Hu, S., Li, J., Xu, F., Mei, S., Le Duff, Y., Yin, L., Pang, X., Cen, S., Jin, Q., Liang, C.*, et al.* (2015). SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation. PLoS genetics *11*, e1005367.

Huang, S., Tao, X., Yuan, S., Zhang, Y., Li, P., Beilinson, H.A., Zhang, Y., Yu, W., Pontarotti, P., Escriva, H.*, et al.* (2016). Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell *166*, 102-114.

Hulme, A.E., Bogerd, H.P., Cullen, B.R., and Moran, J.V. (2007). Selective inhibition of Alu retrotransposition by APOBEC3G. Gene *390*, 199-205.

Ishizu, H., Nagao, A., and Siomi, H. (2011). Gatekeepers for Piwi-piRNA complexes to enter the nucleus. Current opinion in genetics & development *21*, 484-490.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, 1253-1261.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvak, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell *91*, 501-510.

Jacobs, F.M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature.

Janousek, V., Karn, R.C., and Laukaitis, C.M. (2013). The role of retrotransposons in gene family expansions: insights from the mouse Abp gene family. BMC evolutionary biology *13*, 107.

Janousek, V., Laukaitis, C.M., Yanchukov, A., and Karn, R.C. (2016). The role of retrotransposons in gene family expansions in the human and mouse genomes. Genome biology and evolution.

Januszyk, K., Li, P.W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J.A., Martin, S.L., and Clubb, R.T. (2007). Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. The Journal of biological chemistry *282*, 24893-24904.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics *13*, 484-492.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. Current opinion in structural biology *8*, 333-337.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends in genetics : TIG *16*, 418-420.

Kagawa, T., Oka, A., Kobayashi, Y., Hiasa, Y., Kitamura, T., Sakugawa, H., Adachi, Y., Anzai, K., Tsuruya, K., Arase, Y*., et al.* (2015). Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. Human mutation *36*, 327-332.

Kapitonov, V.V., and Jurka, J. (2003). The esterase and PHD domains in CR1-like non-LTR retrotransposons. Molecular biology and evolution *20*, 38-46.

Kapitonov, V.V., and Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS biology *3*, e181.

Kawakami, K. (2005). Transposon tools and methods in zebrafish. Developmental dynamics : an official publication of the American Association of Anatomists *234*, 244-254.

Kawakami, K. (2007). Tol2: a versatile gene transfer vector in vertebrates. Genome biology *8 Suppl 1*, S7.

Kazazian, H.H., Jr. (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626-1632.

Kazazian, H.H., Jr., and Moran, J.V. (1998). The impact of L1 retrotransposons on the human genome. Nature genetics *19*, 19-24.

Kazazian, H.H., Jr., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164-166.

Kedersha, N., Stoecklin, G., Ayodele, M., Yacono, P., Lykke-Andersen, J., Fritzler, M.J., Scheuner, D., Kaufman, R.J., Golan, D.E., and Anderson, P. (2005). Stress granules and processing bodies are dynamically linked sites of mRNP remodeling. The Journal of cell biology *169*, 871-884.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome research *16*, 78-87.

Khazina, E., Truffault, V., Buttner, R., Schmidt, S., Coles, M., and Weichenrieder, O. (2011). Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. Nature structural & molecular biology *18*, 1006-1014.

Khazina, E., and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proceedings of the National Academy of Sciences of the United States of America *106*, 731-736.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837-847.

Klawitter, S., Fuchs, N.V., Upton, K.R., Munoz-Lopez, M., Shukla, R., Wang, J., Garcia-Canadas, M., Lopez-Ruiz, C., Gerhardt, D.J., Sebe, A.*, et al.* (2016). Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat Commun *7*, 10286.

Kleckner, N. (1981). Transposable elements in prokaryotes. Annual review of genetics *15*, 341-404.

Klinakis, A.G., Zagoraiou, L., Vassilatis, D.K., and Savakis, C. (2000). Genome-wide insertional mutagenesis in human cells by the Drosophila mobile element Minos. EMBO reports *1*, 416-421.

Knudson, A.G. (1971). Mutation and Cancer - Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America *68*, 820-&.

Koito, A., and Ikeda, T. (2013). Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. Frontiers in microbiology *4*, 28.

Kondo-Iida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y., and Toda, T. (1999). Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Human molecular genetics *8*, 2303-2309.

Kriegs, J.O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends in genetics : TIG *23*, 158-161.

Krupovic, M., and Koonin, E.V. (2015). Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nature reviews Microbiology *13*, 105-115.

Kubo, S., Seleme, M.C., Soifer, H.S., Perez, J.L., Moran, J.V., Kazazian, H.H., Jr., and Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. Proceedings of the National Academy of Sciences of the United States of America *103*, 8036-8041.

Kulpa, D.A., and Moran, J.V. (2005). Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Human molecular genetics *14*, 3237-3248.

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nature structural & molecular biology *13*, 655-660.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K.*, et al.* (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967-971.

Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985). Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science *227*, 140-146.

Leibold, D.M., Swergold, G.D., Singer, M.F., Thayer, R.E., Dombroski, B.A., and Fanning, T.G. (1990). Translation of LINE-1 DNA elements in vitro and in human cells. Proceedings of the National Academy of Sciences of the United States of America *87*, 6990-6994.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nature reviews Genetics *12*, 615-627.

Li, P.W., Li, J., Timmerman, S.L., Krushel, L.A., and Martin, S.L. (2006). The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic acids research *34*, 853-864.

Liao, C.H., Kuang, Y.Q., Liu, H.L., Zheng, Y.T., and Su, B. (2007). A novel fusion gene, TRIM5-Cyclophilin A in the pig-tailed macaque determines its susceptibility to HIV-1 infection. Aids *21 Suppl 8*, S19-26.

Lim, A.K., and Knowles, B.B. (2015). Controlling Endogenous Retroviruses and Their Chimeric Transcripts During Natural Reprogramming in the Oocyte. J Infect Dis *212*, S47-S51.

Lindic, N., Budic, M., Petan, T., Knisbacher, B.A., Levanon, E.Y., and Lovsin, N. (2013). Differential inhibition of LINE1 and LINE2 retrotransposition by vertebrate AID/APOBEC proteins. Retrovirology *10*, 156.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595-605.

Macfarlane, C.M., and Badge, R.M. (2015). Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and chimpanzees that are absent from genome assemblies. Retrovirology *12*.

Macfarlane, C.M., Collier, P., Rahbari, R., Beck, C.R., Wagstaff, J.F., Igoe, S., Moran, J.V., and Badge, R.M. (2013). Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. Human mutation *34*, 974-985.

Macia, A., Munoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. Molecular and cellular biology *31*, 300-316.

Malfavon-Borja, R., Wu, L.I., Emerman, M., and Malik, H.S. (2013). Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. Proceedings of the National Academy of Sciences of the United States of America *110*, E583-592.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. Molecular biology and evolution *16*, 793-805.

Martin, F., Maranon, C., Olivares, M., Alonso, C., and Lopez, M.C. (1995). Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. Journal of molecular biology *247*, 49-59.

Martin, S.L. (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Molecular and cellular biology *11*, 4804-4807.

Martin, S.L., Branciforte, D., Keller, D., and Bain, D.L. (2003). Trimeric structure for an essential protein in L1 retrotransposition. Proceedings of the National Academy of Sciences of the United States of America *100*, 13815-13820.

Martin, S.L., and Bushman, F.D. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Molecular and cellular biology *21*, 467-475.

Martin, S.L., Li, W.L., Furano, A.V., and Boissinot, S. (2005). The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenetic and genome research *110*, 223-228.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D., and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science *254*, 1808-1810.

Matlik, K., Redik, K., and Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. Journal of biomedicine & biotechnology *2006*, 71753.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences of the United States of America *36*, 344-355.

McClintock, B. (1951). Chromosome organization and genic expression. Cold Spring Harbor symposia on quantitative biology *16*, 13-47.

McKinley, K.L., and Cheeseman, I.M. (2016). The molecular basis for centromere identity and function. Nature reviews Molecular cell biology *17*, 16-29.

McMillan, J.P., and Singer, M.F. (1993). Translation of the human LINE-1 element, L1Hs. Proceedings of the National Academy of Sciences of the United States of America *90*, 11533-11537.

Medstrand, P., and Mager, D.L. (1998). Human-specific integrations of the HERV-K endogenous retrovirus family. Journal of virology *72*, 9782-9787.

Meyers, R.A. (2007). Genomics and genetics : from molecular details to analysis and techniques (Weinheim: Wiley-VCH).

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer research *52*, 643-645.

Mills, R.E., Bennett, E.A., Iskow, R.C., Luttig, C.T., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. American journal of human genetics *78*, 671-679.

Mine, M., Chen, J.M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Ferec, C., Abitbol, M., Ricquier, D.*, et al.* (2007). A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Human mutation *28*, 137-142.

Mol, C.D., Kuo, C.F., Thayer, M.M., Cunningham, R.P., and Tainer, J.A. (1995). Structure and function of the multifunctional DNA-repair enzyme exonuclease III. Nature *374*, 381-386.

Moldovan, J.B., and Moran, J.V. (2015). The Zinc-Finger Antiviral Protein ZAP Inhibits LINE and Alu Retrotransposition. PLoS genetics *11*, e1005121.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530-1534.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917-927.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nature genetics *31*, 159-165.

Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Lower, J., Cichutek, K., Flory, E., Schumann, G.G., and Munk, C. (2006). APOBEC3 proteins inhibit human LINE-1 retrotransposition. The Journal of biological chemistry *281*, 22161-22172.

Munoz-Lopez, M., and Garcia-Perez, J.L. (2010). DNA transposons: nature and applications in genomics. Current genomics *11*, 115-128.

Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903-910.

Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443-446.

Murray, S.C., Haenni, S., Howe, F.S., Fischl, H., Chocian, K., Nair, A., and Mellor, J. (2015). Sense and antisense transcription are associated with distinct chromatin architectures across genes. Nucleic acids research *43*, 7823-7837.

Musova, Z., Hedvicakova, P., Mohrmann, M., Tesarova, M., Krepelova, A., Zeman, J., and Sedlacek, Z. (2006). A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: further evidence for possible bias in retroposon integration. Biochemical and biophysical research communications *347*, 145-149.

Narita, N., Nishio, H., Kitoh, Y., Ishikawa, Y., Ishikawa, Y., Minami, R., Nakamura, H., and Matsuo, M. (1993). Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. The Journal of clinical investigation *91*, 1862-1867.

Nelson, P.N., Carnegie, P.R., Martin, J., Davari Ejtehadi, H., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J., and Murray, P.G. (2003). Demystified. Human endogenous retroviruses. Molecular pathology : MP *56*, 11-18.

Nigumann, P., Redik, K., Matlik, K., and Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics *79*, 628-634.

Nisole, S., Lynch, C., Stoye, J.P., and Yap, M.W. (2004). A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. Proceedings of the National Academy of Sciences of the United States of America *101*, 13324-13328.

Notarangelo, L.D., Kim, M.S., Walter, J.E., and Lee, Y.N. (2016). Human RAG mutations: biochemistry and clinical implications. Nature reviews Immunology *16*, 234-246.

Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven symposia in biology *23*, 366-370.

Okada, N., Hamada, M., Ogiwara, I., and Ohshima, K. (1997). SINEs and LINEs share common 3' sequences: a review. Gene *205*, 229-243.

Orgel, L.E., and Crick, F.H. (1980). Selfish DNA: the ultimate parasite. Nature *284*, 604-607.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian, H.H., Jr. (2002). A mouse model of human L1 retrotransposition. Nature genetics *32*, 655-660.

Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H., Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. American journal of human genetics *73*, 1444-1451.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome research *11*, 2059-2065.

Perepelitsa-Belancio, V., and Deininger, P. (2003). RNA truncation by premature polyadenylation attenuates human mobile element activity. Nature genetics *35*, 363-366.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome research *10*, 411-415.

Piskareva, O., Denmukhametova, S., and Schmatchenko, V. (2003). Functional reverse transcriptase encoded by the human LINE-1 from baculovirus-infected insect cells. Protein expression and purification *28*, 125-130.

Piskareva, O., and Schmatchenko, V. (2006). DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. Febs Lett *580*, 661-668.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Lower, J., Stratling, W.H., Lower, R., and Schumann, G.G. (2012a). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic acids research *40*, 1666-1683.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Lower, J., Stratling, W.H., Lower, R., and Schumann, G.G. (2012b). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic acids research *40*, 1666-1683.

Ray, D.A., Feschotte, C., Pagan, H.J., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome research *18*, 717-728.

Ray, D.A., Pagan, H.J., Thompson, M.L., and Stevens, R.D. (2007). Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. Molecular biology and evolution *24*, 632-639.

Rebollo, R., Farivar, S., and Mager, D.L. (2012). C-GATE - catalogue of genes affected by transposable elements. Mobile DNA *3*, 9.

Richardson, S.R., Doucet, A.J., Kopera, H.C., Moldovan, J.B., Garcia-Perez, J.L., and Moran, J.V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiology spectrum *3*, MDNA3-0061-2014.

Richardson, S.R., Narvaiza, I., Planegger, R.A., Weitzman, M.D., and Moran, J.V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. Elife *3*, e02008.

Rodriguez-Martin, C., Cidre, F., Fernandez-Teijeiro, A., Gomez-Mariano, G., de la Vega, L., Ramos, P., Zaballos, A., Monzon, S., and Alonso, J. (2016). Familial

retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. J Hum Genet *61*, 463-466.

Sahakyan, A.B., Murat, P., Mayer, C., and Balasubramanian, S. (2017). G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. Nature structural & molecular biology.

Samuelson, L.C., Phillips, R.S., and Swanberg, L.J. (1996). Amylase gene structures in primates: retroposon insertions and promoter evolution. Molecular biology and evolution *13*, 767-779.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H., Jr. (1997). Many human L1 elements are capable of retrotransposition. Nature genetics *16*, 37-43.

Sawyer, S.L., Emerman, M., and Malik, H.S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS biology *2*, E275.

Sayah, D.M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature *430*, 569-573.

Sciamanna, I., Gualtieri, A., Cossetti, C., Osimo, E.F., Ferracin, M., Macchia, G., Arico, E., Prosseda, G., Vitullo, P., Misteli, T.*, et al.* (2013). A tumor-promoting mechanism mediated by retrotransposon-encoded reverse transcriptase is active in human transformed cell lines. Oncotarget *4*, 2271-2287.

Sciamanna, I., Gualtieri, A., Piazza, P.V., and Spadafora, C. (2014). Regulatory roles of LINE-1-encoded reverse transcriptase in cancer onset and progression. Oncotarget *5*, 8039-8051.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics *1*, 113-125.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome research *26*, 745-755.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature *418*, 646-650.

Shen, S.H., Slightom, J.L., and Smithies, O. (1981). A history of the human fetal globin gene duplication. Cell *26*, 191-203.

Shin, W., Lee, J., Son, S.Y., Ahn, K., Kim, H.S., and Han, K. (2013). Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome. PloS one *8*.

Shukla, R., Upton, K.R., Munoz-Lopez, M., Gerhardt, D.J., Fisher, M.E., Nguyen, T., Brennan, P.M., Baillie, J.K., Collino, A., Ghisletti, S.*, et al.* (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell *153*, 101-111.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology *12*, 246-258.

Skowronski, J., Fanning, T.G., and Singer, M.F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. Molecular and cellular biology *8*, 1385-1397.

Skowronski, J., and Singer, M.F. (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proceedings of the National Academy of Sciences of the United States of America *82*, 6050-6054.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature reviews Genetics *8*, 272-285.

Smit, A.F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic acids research *21*, 1863-1872.

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. Current opinion in genetics & development *6*, 743-748.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current opinion in genetics & development *9*, 657-663.

Smit, A.F., Toth, G., Riggs, A.D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. Journal of molecular biology *246*, 401-417.

Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon, D.F., Casella, A., Erlanger, B.*, et al.* (2012). Extensive somatic L1 retrotransposition in colorectal tumors. Genome research *22*, 2328-2338.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Molecular and cellular biology *21*, 1973-1985.

Stacey, S.N., Kehr, B., Gudmundsson, J., Zink, F., Jonasdottir, A., Gudjonsson, S.A., Sigurdsson, A., Halldorsson, B.V., Agnarsson, B.A., Benediktsdottir, K.R.*, et al.* (2016). Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. Human molecular genetics *25*, 1008-1018.

Startek, M., Szafranski, P., Gambin, T., Campbell, I.M., Hixson, P., Shaw, C.A., Stankiewicz, P., and Gambin, A. (2015). Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. Nucleic acids research *43*, 2188-2198.

Stenglein, M.D., and Harris, R.S. (2006). APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. The Journal of biological chemistry *281*, 16837-16841.

Suminaga, R., Takeshima, Y., Yasuda, K., Shiga, N., Nakamura, H., and Matsuo, M. (2000). Non-homologous recombination between Alu and LINE-1 repeats caused a 430-kb deletion in the dystrophin gene: a novel source of genomic instability. J Hum Genet *45*, 331-336.

Suter, C.M., Martin, D.I., and Ward, R.L. (2004). Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue. International journal of colorectal disease *19*, 95-101.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Molecular and cellular biology *10*, 6718-6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Tavare, S., Marshall, C.R., Will, O., Soligo, C., and Martin, R.D. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature *416*, 726-729.

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.*, et al.* (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell *155*, 1034-1048.

Tchenio, T., Casella, J.F., and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic acids research *28*, 411-415.

Temtamy, S.A., Aglan, M.S., Valencia, M., Cocchi, G., Pacheco, M., Ashour, A.M., Amr, K.S., Helmy, S.M., El-Gammal, M.A., Wright, M.*, et al.* (2008). Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence. Human mutation *29*, 931-938.

Thayer, R.E., Singer, M.F., and Fanning, T.G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. Gene *133*, 273-277.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome research *21*, 1800-1812.

Tica, J., Lee, E., Untergasser, A., Meiers, S., Garfield, D.A., Gokcumen, O., Furlong, E.E., Park, P.J., Stutz, A.M., and Korbel, J.O. (2016). Next-generation sequencing-based detection of germline L1-mediated transductions. BMC genomics *17*, 342.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome research *14*, 62-66.

Tubio, J.M., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K.*, et al.* (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Ullu, E., and Weiner, A.M. (1985). Upstream sequences modulate the internal promoter of the human 7SL RNA gene. Nature *318*, 371-374.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M.*, et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

Usdin, K., and Furano, A.V. (1989). The structure of the guanine-rich polypurine:polypyrimidine sequence at the right end of the rat L1 (LINE) element. The Journal of biological chemistry *264*, 15681-15687.

van den Hurk, J.A., van de Pol, D.J., Wissinger, B., van Driel, M.A., Hoefsloot, L.H., de Wijs, I.J., van den Born, L.I., Heckenlively, J.R., Brunner, H.G., Zrenner, E.*, et al.* (2003). Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. Human genetics *113*, 268-275.

van den Hurk, J.A.J.M., Meij, I.C., Seleme, M.D.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S.*, et al.* (2007). L1 retrotransposition can occur early in human embryonic development. Human molecular genetics *16*, 1587-1592.

Vanvalen, L. (1973). Theory in Abstracts. Bioscience *23*, 626-626.

Varshney, D., Vavrova-Anderson, J., Oler, A.J., Cowling, V.H., Cairns, B.R., and White, R.J. (2015). SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. Nat Commun *6*.

Virgen, C.A., Kratovac, Z., Bieniasz, P.D., and Hatziioannou, T. (2008). Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. Proceedings of the National Academy of Sciences of the United States of America *105*, 3563-3568.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA elements: a hominid-specific retroposon family. Journal of molecular biology *354*, 994-1007.

Wang, X., Han, Y., Dang, Y., Fu, W., Zhou, T., Ptak, R.G., and Zheng, Y.H. (2010). Moloney leukemia virus 10 (MOV10) protein inhibits retrovirus replication. The Journal of biological chemistry *285*, 14346-14355.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T.*, et al.* (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539-543.

Weber, M.J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. PLoS genetics *2*, e205.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., and Moran, J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. Molecular and cellular biology *21*, 1429-1439.

Weichenrieder, O., Wild, K., Strub, K., and Cusack, S. (2000). Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature *408*, 167-173.

Weiss, R.A. (2016). Human endogenous retroviruses: friend or foe? APMIS : acta pathologica, microbiologica, et immunologica Scandinavica *124*, 4-10.

White, T.E., Brandariz-Nunez, A., Valle-Casuso, J.C., Knowlton, C., Kim, B., Sawyer, S.L., and Diaz-Griffero, F. (2014). Effects of human SAMHD1 polymorphisms on HIV-1 susceptibility. Virology *460-461*, 34-44.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proceedings of the National Academy of Sciences of the United States of America *113*, E2326-2334.

Wilson, R.C., and Doudna, J.A. (2013). Molecular mechanisms of RNA interference. Annu Rev Biophys *42*, 217-239.

Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS genetics *7*, e1002371.

Wissing, S., Montano, M., Garcia-Perez, J.L., Moran, J.V., and Greene, W.C. (2011). Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. The Journal of biological chemistry *286*, 36427-36437.

Wissing, S., Munoz-Lopez, M., Macia, A., Yang, Z., Montano, M., Collins, W., Garcia-Perez, J.L., Moran, J.V., and Greene, W.C. (2012). Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. Human molecular genetics *21*, 208-218.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., and Batzer, M.A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. Proceedings of the National Academy of Sciences of the United States of America *103*, 17608-17613.

Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO journal *9*, 3353-3362.

Yang, N., and Kazazian, H.H., Jr. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature structural & molecular biology *13*, 763-771.

Yang, N., Zhang, L., Zhang, Y., and Kazazian, H.H., Jr. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic acids research *31*, 4929-4940.

Yang, Z., Boffelli, D., Boonmark, N., Schwartz, K., and Lawn, R. (1998). Apolipoprotein(a) gene enhancer resides within a LINE element. The Journal of biological chemistry *273*, 891-897.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends in genetics : TIG *13*, 335-340.

Yoshida, K., Nakamura, A., Yazaki, M., Ikeda, S., and Takeda, S. (1998). Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. Human molecular genetics *7*, 1129-1132.

Yu, F., Zingler, N., Schumann, G., and Stratling, W.H. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. Nucleic acids research *29*, 4493-4501.

Yu, Y., Gu, J., Jin, Y., Luo, Y., Preall, J.B., Ma, J., Czech, B., and Hannon, G.J. (2015). Panoramix enforces piRNA-dependent cotranscriptional silencing. Science *350*, 339-342.

Zhao, K., Du, J., Han, X., Goodier, J.L., Li, P., Zhou, X., Wei, W., Evans, S.L., Li, L., Zhang, W.*, et al.* (2013). Modulation of LINE-1 and Alu/SVA retrotransposition by Aicardi-Goutieres syndrome-related SAMHD1. Cell reports *4*, 1108-1115.

Chapter 2

Spliced Integrated Retrotransposed Element (SpIRE) Formation in the Human Genome

Peter A. Larson, Christine R. Beck, and John V. Moran

This chapter is a working draft of a paper that is near submission. Dr. Christine Beck identified $SpIRE_{97\text{-}622}$ and performed initial searches of the human genome reference for $SpIRE_{97\text{-}622}$ sequences. Mr. Peter Larson performed all experiments contained within this chapter.

Abstract

Long INterspersed Element-1 retrotransposons (LINE-1s or L1s) contain an internal RNA polymerase II promoter within their 5' untranslated region (UTR) and encode two proteins (ORF1p and ORF2p) required for their mobilization (*i.e.*, retrotransposition). The evolutionary success of L1 relies on the reiterative retrotransposition of full-length L1 sequences. Previous studies identified functional splice donor, splice acceptor, and polyadenylation sequences in L1 RNA and provided evidence that spliced L1 RNAs can retrotranspose. Here, we report that the retrotransposition of intra-5'UTR or 5'UTR/ORF1 spliced L1 transcripts leads to the generation of Spliced Integrated Retrotransposed Elements (SpIREs) and describe a new Intra-5'UTR SpIRE that is 10 times more abundant than previously identified SpIREs. Intra-5'UTR SpIREs lack *cis*-acting transcription factor binding sites and exhibit reduced promoter activity. In addition to lacking *cis*-acting sequences 5'UTR/ORF1 SpIREs produce non-functional ORF1p variants. We show that these deletions render SpIREs retrotransposition-defective and conclude that splicing negatively affects L1 retrotransposition. Surprisingly, two percent of annotated full-length L1s from the L1PA1-L1PA6 subfamilies are actually SpIREs,

demonstrating that L1 splicing has occurred for at least ~27 million years. Finally, we describe how sequence changes in the 5'UTR can result in new, deleterious splicing events, representing another layer of negative selective pressure on L1 evolution.

Introduction

Long INterspersed Element-1 (LINE-1 or L1) retrotransposon sequences comprise approximately 17% of human genomic DNA (Lander et al. 2001). Over 99.9% of human L1s cannot retrotranspose due to 5' truncations, internal DNA rearrangements, or point mutations that inactivate the L1-encoded proteins (Grimaldi et al. 1984; Kazazian and Moran 1998; Lander et al. 2001). However, the average diploid genome harbors approximately 80-100 full-length retrotransposition-competent (RC-L1s) L1s (Brouha et al. 2003), and a small number of these, called highly active or 'hot' L1s, can retrotranspose efficiently in cultured cells (Sassaman et al. 1997; Brouha et al. 2003; Beck et al. 2010; Macfarlane et al. 2013). Active L1 retrotransposition affects both intra- and inter-individual human genetic variation (Kazazian et al. 1988; Holmes et al. 1994; Beck et al. 2010; Beck et al. 2011; Wimmer et al. 2011; Hancks and Kazazian 2016). Indeed, L1-mediated retrotransposition events are responsible for approximately 1/250 disease-producing mutations in man (Kazazian et al. 1988; Wimmer et al. 2011; Hancks and Kazazian 2016)

Human RC-L1s are approximately six-kilobases (kb) in length (Scott et al. 1987; Dombroski et al. 1991). They contain a 5'UTR that harbors both sense (Swergold 1990) and anti-sense (Speek 2001) RNA polymerase II promoters, as well as an anti-sense open reading frame (ORF-0) (Denli et al. 2015) of unknown function. The 5'UTR is followed by two open reading frames (ORF1 and ORF2), which are separated by a 63 base pair (bp) inter-ORF spacer (Dombroski et al. 1991; Alisch et al. 2006). L1s end with a 3'UTR containing a conserved polypurine motif, an RNA polymerase II polyadenylation signal, followed by a variable length polyadenosine (poly(A)) tract (Scott et al. 1987; Usdin and Furano 1989; Moran et al. 1999; Perepelitsa-Belancio and Deininger 2003).

ORF1 encodes an ~40 kiloDalton (kDa) protein (ORF1p) that contains an amino-terminal coiled-coil domain required for ORF1p trimerization (Martin et al. 2003; Khazina et al. 2011; Naufer et al. 2016), a centrally located, non-canonical RNA recognition motif (RRM) domain (Khazina and Weichenrieder 2009; Khazina et al. 2011), and a carboxyl-terminal domain that harbors conserved basic amino acid residues (Moran et al. 1996;

Januszyk et al. 2007). The RRM and carboxyl-terminal domains are critical for ORF1p nucleic acid binding (Martin 1991; Holmes et al. 1992; Hohjoh and Singer 1996; Hohjoh and Singer 1997; Januszyk et al. 2007; Khazina and Weichenrieder 2009). ORF1p also exhibits nucleic acid chaperone activity, which is postulated to play a role in L1 integration (Martin and Bushman 2001; Khazina and Weichenrieder 2009).

ORF2 encodes an ~150 kDa protein (ORF2p) (Ergun et al. 2004; Doucet et al. 2010; Goodier et al. 2010) that contains conserved apurinic/apyrimidinic endonuclease-like (EN) (Martin et al. 1995; Feng et al. 1996) and reverse transcriptase (RT) domains (Hattori et al. 1986; Mathias et al. 1991; Moran et al. 1996), as well as a conserved cysteine-rich (C) domain (Fanning and Singer 1987; Moran et al. 1996). Biochemical activities contained within both ORF1p and ORF2p are required for L1 retrotransposition in cultured human cells (Moran et al. 1996).

A round of human RC-L1 retrotransposition begins with the internal sense-strand promoter initiating transcription at or near the first nucleotide of the 5'UTR (Athanikar et al. 2004; Dmitriev et al. 2007; Richardson et al. 2015). The resultant mRNA is exported to the cytoplasm where it undergoes translation (Leibold et al. 1990; McMillan and Singer 1993; Alisch et al. 2006; Dmitriev et al. 2007). Following translation, ORF1p and ORF2p preferentially bind to their encoding mRNA in *cis* to form a ribonucleoprotein (RNP) particle (Martin 1991; Hohjoh and Singer 1996; Wei et al. 2001; Kulpa and Moran 2005; Kulpa and Moran 2006; Doucet et al. 2015). Nascent ORF2p recruitment in *cis* is likely mediated by the L1 poly(A) tail (Ahl et al. 2015; Doucet et al. 2015). Components of the L1 RNP then gain access to the nucleus by a mechanism that does not require nuclear envelope breakdown (Kubo et al. 2006). L1 integration occurs by target-site primed reverse transcription (TPRT) (Luan et al. 1993; Feng et al. 1996; Cost et al. 2002; Kulpa and Moran 2006). The L1 EN makes a single-strand endonucleolytic nick at a thymidine-rich sequence (*e.g.*, 5'-TTTT/A-3', 5'-TTTC/A-3', *etc*.) to liberate a 3' hydroxyl group that is used as a primer by the ORF2p RT to initiate the reverse transcription of L1 mRNA (Feng et al. 1996; Cost et al. 2002; Morrish et al. 2002). How TPRT is completed requires elucidation. However, as demonstrated for the related R2 retrotransposon from *Bombyx mori*, (Christensen et al. 2005), it is possible that L1

ORF2p participates in both second-strand genomic DNA cleavage and second-strand L1 cDNA synthesis.

Although the L1-encoded proteins preferentially retrotranspose their encoding mRNA (Esnault et al. 2000; Wei et al. 2000; Kulpa and Moran 2006), L1 ORF1p and/or ORF2p also can act in *trans* (*trans*-complementation) to mobilize RNAs encoded by non-autonomous Short INterspersed Elements (SINEs; *e.g.*, Alu RNA (Deininger et al. 1981; Dewannieux et al. 2003) and SINE-R/VNTR/Alu (SVA) RNA (Ostertag et al. 2003; Hancks et al. 2011; Raiz et al. 2012) and, more rarely, cellular mRNAs, which leads to processed pseudogene formation (Esnault et al. 2000; Wei et al. 2001; Buzdin et al. 2002; Gilbert et al. 2005; Garcia-Perez et al. 2007).

The evolutionary success of L1 requires the faithful retrotransposition of full-length L1 mRNAs and their amplification in subsequent generations. Previous studies have revealed the presence of functional splice donor (SD), splice acceptor (SA), and premature polyadenylation signals in primary full-length RC-L1 transcripts (Perepelitsa-Belancio and Deininger 2003; Belancio et al. 2006; Belancio et al. 2008; Belancio et al. 2010). Paradoxically, the use of these sites during post-transcriptional processing leads to the production of truncated and/or internally deleted L1 mRNAs (Perepelitsa-Belancio and Deininger 2003; Belancio et al. 2006; Belancio et al. 2008; Belancio et al. 2010), which could adversely affect L1 retrotransposition. Thus, it is somewhat surprising that *cis*-acting sequences that could negatively affect L1 retrotransposition have been maintained and not subject to negative selection

Here, we address how the retrotransposition of spliced L1 RNAs leads to the generation of Spliced Integrated Retrotransposed Elements (SpIREs). We describe two classes of SpIREs: those that splice within the 5'UTR (intra-5'UTR SpIREs) and those that splice from within the 5'UTR and into the ORF1 sequence (5'UTR/ORF1 SpIREs). By combining genetic, genomic, molecular biological, biochemical, and computational approaches, our data demonstrate that SpIREs are evolutionary "dead-ends" that likely are unable to undergo efficient retrotransposition in subsequent generations. Additionally, our data suggest a mechanism for how some apparently deleterious *cis*-acting splicing sequences are retained in the currently amplifying L1 lineage.

Results

A previously identified polymorphic L1 likely resulted from the retrotransposition of a spliced L1 transcript

We previously identified a polymorphic L1 in the human population (accession #AC225317) that contains a 524 nucleotide deletion within its 5'UTR (Beck et al. 2010). Upon closer inspection, we determined that this deletion may have resulted from the retrotransposition of a spliced L1 transcript that used a previously identified splice donor (SD: $G_{98}U_{99}$) (Belancio et al. 2006) and an unreported splice acceptor (SA: $A_{620}G_{621}$) within the 5'UTR of a full-length L1 (numbering based on L1.3, accession # L19088) (Figures 2.1A and 2.1B). The structure of this element resembled previous L1s characterized by Belancio and colleagues, lending support to the hypothesis that spliced L1 transcripts occasionally could undergo retrotransposition in the human genome (Belancio et al. 2006; Belancio et al. 2008). We named these L1s SpIREs (Spliced Integrated Retrotransposed Elements). The three SpIREs investigated here (Figures 2.1B, 2.1C, and 2.1D) all use the same splice donor (SD: $G_{98}U_{99}$), but pair with different splice acceptor sequences that reside within either the L1 5'UTR (SA: $A_{620}G_{621}$ or SA: $A_{788}G_{789}$) or L1 ORF1 (SA: $A_{974}G_{975}$).

SpIREs are present in the human genome

To determine the copy number of SpIRE $G_{98}U_{99}/A_{620}G_{621}$ sequences (now named $SpIRE_{97/622}$) present in the human genome reference sequence (HGR), we conducted BLAT (https://genome.ucsc.edu/index.html) (Kent 2002) searches using a 100 nucleotide *in silico* probe that spanned the intra-5'UTR splice junction present in $SpIRE_{97/622}$ (nucleotides 47-97 and 622-672 of L1.3). We screened build 37 of the HGR (GRCh37/hg19) for L1 sequences using Repeat Masker (Jurka et al. 2005).

The HGR contains an annotated record of L1s that have accumulated over evolutionary time. Thus, querying the genome allows a determination of how SpIREs contribute to the L1 genomic repertoire. We identified 116 $SpIRE_{97/622}$ sequences that spanned the youngest L1PA1 (also known as L1Hs, currently amplifying in the human population) to L1PA6 subfamilies (which amplified approximately 27 million years ago

(MYA)), but none in older L1 subfamilies (Figure 2.6A) (Khan et al. 2006; Song and Boissinot 2007). Thus, 116/6609 or ~1.8% of previously determined full-length L1s in the L1PA1-L1PA6 subfamilies are $SpIRE_{97/622}$ sequences (Figure 2.6).

Almost half of the $SpIRE_{97/622}$ sequences we identified belong to the L1PA3 subfamily (53 sequences, comprising ~3.4% of annotated full-length L1s in that subfamily) (Figure 2.6A; Data Set 1, Table 2.1). The L1PA1 subfamily harbors 6 $SpIRE_{97/622}$ (comprising 2.0% of annotated full length L1s in that subfamily) and the L1PA6 subfamily contains only one $SpIRE_{97/622}$ (comprising 0.1% of annotated full-length L1s in that subfamily) (Figure 2.6; Data Set 1; Table 2.1). Seven $SpIRE_{97/622}$ sequences could not be unambiguously assigned to a specific L1 subfamily and are classified as either L1PA2-L1PA3 or L1PA4-L1PA6 sequences (Figure 2.6A; Data Set 1; Table 2.1) (Smit et al. 1995).

Given the above data, we used BLAT to search the HGR for additional L1s containing $G_{98}U_{99}/A_{788}G_{789}$ and $G_{98}U_{99}/A_{974}G_{975}$ splicing events (now named $SpIRE_{97/790}$ and $SpIRE_{97/976}$ respectively) (Belancio et al. 2006; Belancio et al. 2008). These searches confirmed the presence of four previously identified $SpIRE_{97/790}$ sequences in the L1PA1-L1PA2 subfamilies (Figure 2.6A, Data Set 1, Table 2.1). We added an additional $SpIRE_{97/976}$ sequence to the ten previously identified (Figure 2.6C; Data 1; Table 2.1) (Belancio et al. 2006; Belancio et al. 2008). In total, these three classes of SpIREs comprise a small, but significant (131/6609 or ~2%), percentage of annotated full-length L1s from the L1PA1-L1PA6 subfamilies. The newly identified $SpIRE_{97/622}$ represents the majority (116/131 or ~89%) of those sequences.

SpIREs contain L1 structural hallmarks

Sequence characterization of the 131 $SpIRE_{97/622}$, $SpIRE_{97/790}$, and $SpIRE_{97/976}$ sequences revealed SpIREs are general flanked by target site duplications that ranged from 6-32 bp, end in a 3' poly(A) tract, and integrated into an L1 EN consensus cleavage site (5'-TTTT/A-3' and variants of that sequence) (Data Set 1; Table 2.1). The majority (119/131) of SpIRES generally did not contain additional structural alterations (*e.g.*, internal deletions). However, three SpIREs contained additional internal deletions, one contained an inversion/deletion, and eight appeared to be truncated at there 3'

ends. Four of the eight 3' truncated SpIREs retrotransposed into an older L1 generating chimeric L1s (Kriegs et al. 2007) that were easily identifiable in the genome browser (Data Set 1; Table 2.1). For the remaining four 3' truncated elements we were unable to determine the reason for the apparent truncation (see Methods).

Consistent with previous studies, approximately one-third of the SpIREs (46/131) are present within the introns of RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) (Pruitt et al. 2007) annotated genes and the majority (29/46 or ~63%) are present in the opposite transcriptional orientation of the annotated gene (Table 2.1) (Smit 1999; Gilbert et al. 2002; Symer et al. 2002; Gilbert et al. 2005; Boissinot et al. 2006; Chen et al. 2006b). Two SpIREs (~1.5%) have a 5' transductions of 130 and 560 bp (Lander et al. 2001; Chen et al. 2006a; Evrony et al. 2012), whereas 11/131 (~8.4%) contained 3' transductions that range in length from 61 to 1036 bp (Moran et al. 1999; Goodier et al. 2000; Pickeral et al. 2000; Macfarlane et al. 2013). Despite identifying SpIREs containing transduction events, we were unable to uncover evidence that any SpIRE acted as a progenitor element that gave rise to a new SpIRE (Table 2.1).

Intra-5'UTR splicing reduces L1 promoter activity

Since the formation of $SpIRE_{97/622}$ resulted in the deletion of five of six known transcription factor binding sites within the L1 5'UTR (Hohjoh et al. 1990; Minakami et al. 1992; Becker et al. 1993; Yang et al. 1998; Tchenio et al. 2000; Yang et al. 2003; Athanikar et al. 2004) (Figure 2.1A), we hypothesized the $SpIRE_{97/622}$ 5'UTR would have reduced promoter activity. To test this, we created L1/firefly luciferase expression vectors by subcloning the wild type (WT) and $SpIRE_{97/622}$ 5'UTR sequences upstream of a promoterless firefly (*Photinus pyralis)* luciferase expression vector (pGL4.11), creating $pPL_{WT}LUC$ and $pPL_{97/622}LUC$ respectively (Figure 2.2A). We then characterized the expression from these 5'UTRs in functional assays.

We first conducted northern blot analyses using polyadenylated mRNAs isolated from untransfected (UTF) HeLa-JVM cells and HeLa-JVM cells transfected with the different luciferase expression vectors (Figure 2.2). An RNA probe complementary to the first 100 ribonucleotides of the L1 5'UTR (Figure 2.2A; purple line) detected a strong signal at the expected size of roughly 2.7 kb in mRNAs derived from HeLa-JVM cells

transfected with $pPL_{WT}LUC$. We did not detect signals in mRNAs derived from untransfected HeLa-JVM cells or HeLa-JVM cells transfected with $pPL_{97/622}LUC$, or pGL4.11 (Figure 2.2B, first panel). Similar results were obtained using RNA probes complementary to either ribonucleotides 103-330 of the L1 5'UTR (Figure 2A, red line; Figure 2.2B, second panel) or the 3' end of luciferase mRNA (Figure 2.2A, blue line; Figure 2.2B, third panel). Control experiments verified the integrity and quality of the mRNAs (Figure 2.2B, actin probe). These data are consistent with previously published findings, which demonstrated that L1 transcription faithfully begins at or near the first nucleotide of the L1 5'UTR (Athanikar et al. 2004).

We were able to detect the predicted 2.2 kb transcript emanating from $pPL_{97/622}LUC$ upon the prolonged exposure of the Northern blots using probes complementary to either the first 100 ribonucleotides of the L1 5'UTR or the 3' end of the luciferase gene, but not using a probe complementary to ribonucleotides 103-330 of the L1 5'UTR (Figure 2.7), suggesting that $SpIRE_{97/622}$ 5'UTR retained weak promoter activity. Since the splicing events that gave rise to $SpIRE_{97/790}$ and $SpIRE_{97/976}$ led to larger deletions of the 5'UTR when compared to $SpIRE_{97/622}$, we reasoned that they would lead to a similar, if not greater, reduction in transcriptional activity; thus, they were not tested in this assay.

To corroborate the northern blot analyses, we conducted dual luciferase expression assays on whole cell lysates derived from HeLa-JVM cells co-transfected with firefly luciferase-based vectors ($pPL_{WT}LUC$, $pPL_{97/622}LUC$, or pGL4.11) and a constitutively expressed Renilla luciferase (*Renilla reniformis)* plasmid (pRL-TK; see Methods). Consistent with the northern blot data, HeLa-JVM cells transfected with $pPL_{WT}LUC$ exhibited ~267-fold increase of normalized firefly luciferase activity over those transfected with the promoterless pGL4.11 vector (Figure 2.2C). By comparison, HeLa-JVM cells transfected with $pPL_{97/622}LUC$ exhibited only ~7-fold increase of normalized firefly luciferase activity over those transfected with the promoterless pGL4.11 vector (Figure 2.2C). Together, the above data suggest that the splicing event leading to the generation of $SpIRE_{97/622}$ severely compromises its promoter activity.

Mutating the 5' splice donor site results in decreased L1 promoter activity

Given that splicing reduces transcriptional activity we wanted to determine why the $G_{98}U_{99}$ splice donor might be conserved. Previous studies revealed that a RUNX3 binding site within the 5'UTR is important for maximal L1 promoter activity (Yang, Zhang et al. 2003). Interestingly, the splice donor site used to generate the three classes of SpIREs reported here is contained within a core RUNX3 binding site in L1 DNA that is conserved from the L1PA1-PA10 subfamilies (Figure 2.1A; SD: $G_{98}U_{99}$; Figure 2.6) (Khan et al. 2006). Thus, we hypothesized that this SD is retained to maintain an active RUNX3 site. To test this hypothesis we mutated the splice donor sequence in the WT 5'UTR ($U_{99}C$, creating $pPL_{SDm}LUC$) (Krawczak et al. 1992) and asked if this mutation affects 5'UTR promoter activity. Northern blot analyses using the previously described riboprobes detected a signal at ~2.7kb in mRNAs derived from HeLa-JVM cells transfected with $pPL_{SDm}LUC$. Notably, there is markedly less of this transcript when compared to cells transfected with $pPL_{WT}LUC$ (Figure 2.2B; ~18% of $pPL_{WT}LUC$). By comparison, mutating the splice acceptor site within the 5'UTR did not drastically affect promoter activity (Figure 2.2B; $pPL_{SAm}LUC$; $A_{620}C$). Thus, our data suggest retention of the complete RUNX3 site is critical for L1 promoter activity and are consistent with previous findings (Yang et al. 2003).

Reverse Transcription PCR (RT-PCR) shows that intra-5'UTR splicing is a rare event

We next wanted to identify spliced L1 mRNAs that might give rise to SpIREs. To this end we conducted qualitative RT-PCR experiments using poly(A) mRNAs isolated from HeLa-JVM cells transfected with a series of L1/firefly luciferase expression vectors (Figure 2.2D; see Methods). The REV-LUC oligonucleotide (Figure 2D, purple line) was used to initiate L1/firefly luciferase first strand cDNA synthesis; the cDNA products then were PCR amplified using FWD-5'UTR (Figure 2.2D, red line) and REV-LUC oligonucleotide primers. The resultant cDNAs were cloned and characterized by Sanger DNA sequencing. Control experiments conducted in the absence of reverse transcriptase (Figure 2.2D, bottom gel) revealed that the characterized products were derived from the PCR amplification of cDNAs (Figure 2.2D, top gel).

We detected the predicted L1/firefly luciferase cDNA products from HeLa-JVM cells transfected with $pPL_{WT}LUC$, $pPL_{SDm}LUC$, and $pPL_{SAm}LUC$ (Figure 2.2D, top gel, yellow "*" in lanes 1, 3, and 4), as well as the shorter predicted L1/firefly luciferase cDNA product from HeLa-JVM cells transfected with $pPL_{97/622}LUC$ (Figure 2.2D, yellow "#" in lane 2). However, consistent with the northern blot experiments (Figure 2.2B), we could not detect the $SpIRE_{97/622}$ L1/firefly luciferase cDNA product from mRNAs derived from HeLa-JVM cells transfected with $pPL_{WT}LUC$ (Figure 2.2D). Instead, we detected a minor L1/firefly luciferase cDNA product that corresponds to the $SpIRE_{97/790}$ splicing event from cells transfected with $pPL_{WT}LUC$ and $pPL_{SAm}LUC$ (Figure 2.2D, top gel, yellow "+", lanes 1 and 4; Figure 2.1C) (Belancio et al. 2006). Importantly, this product was not detected in untransfected HeLa-JVM cells or in HeLa-JVM cells transfected with either pGL4.11 or $pPL_{SDm}LUC$. Thus, in agreement with our northern blot data, it appears that full-length L1/firefly luciferase mRNAs represent the major RNA species in these assays and that intra-5'UTR splicing is a relatively rare event.

Intra-5'UTR splicing does not dramatically affect L1 mRNA translation

Given that spliced L1 RNAs can integrate suggests that L1 translation is intact. We next tested whether intra-5'UTR splicing affects L1 mRNA translation. To accomplish this goal, we transfected HeLa-JVM cells with a WT L1 (pJM101/L1.3), an L1 that lacks the 5'UTR (pJM102/L1.3), or an L1 that contains an intra-5'UTR splicing event (pPL_{97-622}/L1.3) (Sassaman et al. 1997; Morrish et al. 2002). A cytomegalovirus early (CMV) promoter augments the expression of each L1 and a hygromycin resistance gene on the pCEP4 plasmid ensures selection of transfected cells.

Western blot analyses were conducted using whole cell lysates (WCLs) derived from hygromycin-resistant HeLa-JVM cells transfected with the above constructs nine days post-transfection. An ORF1p polyclonal antibody (α-N-ORF1p; directed against amino acids +31 to +49 in L1.3(Moldovan and Moran 2015) (UniProtKB accession #Q9UN81)) detected an ~40 kDa product in cells transfected with pJM101/L1.3, pJM102/L1.3, and pPL_{97-622}/L1.3, but not in untransfected cells (Figure 2.3B). HeLa-JVM cells transfected with pPL_{97-622}/L1.3 exhibited a slight reduction in the steady state level of ORF1p when compared to HeLa-JVM cells transfected with pJM101/L1.3 or pJM102/L1.3 (Figure

2.3B). Since a CMV promoter augmented L1 transcription, it is unlikely that this reduction is due to reduced L1 expression, although it remains possible that the slight reduction in ORF1p is due to an alteration of the sequence or secondary structure of the 5'UTR in $pPL_{97\text{-}622}$/L1.3 mRNA.

5'UTR/ORF1 splicing leads to N-terminal truncated ORF1p

The 5'UTR/ORF1 splicing event generates a deletion that lacks the first 66 nucleotides of ORF1, which includes the canonical ORF1p methionine start codon (Figure 2.3C, black AUG, 37 kDa). Thus, we hypothesized that ORF1p synthesis might initiate from one of two in-frame methionine codons (AUG) that are located in weak Kozak consensus sequences either 102 or 270 ribonucleotides downstream from the canonical AUG start codon (Figure 2.3C) (Kozak 1984). If the downstream methionine codons are used for translation initiation we expected truncated ORF1 proteins of ~33 kDa and ~27 kDa respectively.

Western blot analyses were conducted as above using WCLs derived from hygromycin-resistant HeLa-JVM cells transfected with pJM101/L1.3, $pPL_{97\text{-}976}$/L1.3, or pCEP/GFP (Alisch et al. 2006). As predicted, the α-N-ORF1p and α-C-ORF1p antibodies detected an ~40 kDa protein in WCLs derived from HeLa-JVM cells transfected with pJM101/L1.3, but did not detect a protein in WCLs derived from HeLa-JVM cells transfected with the pCEP/GFP control (Figure 2.3C). The α-N-ORF1p antibody also detected an ~33kDa protein in WCLs derived from HeLa-JVM cells transfected with $pPL_{97\text{-}976}$/L1.3, whereas the α-C-ORF1p antibody detected ~33kDa and ~27kDa proteins in the same extracts (Figure 2.33C). Similar results were obtained when RNP extracts were used in western blot experiments (Figure 2.8A). To confirm that these products were ORF1p derived we added a T7-*gene10* epitope tag to ORF1 in $pPL_{97\text{-}976}$/L1.3. Western blots using a α–T7 antibody recapitulated our ORF1p antibody results (Figure 2.8B). Thus, the 5'UTR/ORF1 splicing event leads to the generation of an mRNA that if translated results in amino-terminal truncated derivatives of ORF1p.

Intra-5'UTR splicing drastically decreases L1 retrotransposition efficiency

Our data indicate that the intra-5'UTR splicing event that leads to the generation of $SpIRE_{97/622}$ contains a defective promoter and, if transcribed, produces slightly less ORF1p than a WT L1. Thus, we hypothesized that an intra-5'UTR spliced L1 mRNA would be capable of undergoing an initial round of L1 retrotransposition. However, the resultant full-length retrotransposition events would contain a defective promoter, compromising subsequent retrotransposition.

To test the above hypothesis, we examined whether RNAs derived from the L1 expression constructs with or without an exogenous CMV, could retrotranspose using a cultured cell retrotransposition assay (Moran et al. 1996). The 3'UTR of each of these constructs contains a retrotransposition indicator cassette (*mneoI*). The *mneoI* cassette consists of an antisense copy of a neomycin transferase gene whose coding sequence is interrupted by an intron that resides in the same transcriptional orientation as the L1 (Freeman et al. 1994; Moran et al. 1996). This arrangement ensures that a functional neomycin transferase gene only will be activated upon L1 retrotransposition (Freeman et al. 1994; Moran et al. 1996). Retrotransposition efficiency then can be quantified by counting the resultant numbers of G418-resistant foci (Moran et al. 1996; Wei et al. 2000).

Consistent with previous reports, RNAs derived from RC-L1s that either lack (pJM101/L1.3ΔCMV, grey bar) or contain an exogenous CMV promoter (pJM101/L1.3 and pJM102/L1.3, black bar) could undergo efficient retrotransposition (Figure 4A), whereas RNAs derived from a retrotransposition-deficient L1 containing a missense mutation (D702A) that disrupts ORF2p RT activity (pJM105/L1.3) could not undergo retrotransposition (Figure 2.4A) (Wei et al. 2001). By comparison, the pPL_{97-622}/L1.3 expression construct produced RNAs that could undergo efficient retrotransposition only when a CMV promoter augmented L1 expression (Figure 2.4A, black bar, ~70% the activity of pJM101/L1.3), but not when L1 expression was driven from the 5'UTR harboring the intra-5'UTR splicing event (Figure 2.4A, grey bar, ~7.1% the activity of pJM101/L1.3). Consistent with this conclusion, we also showed that an L1 lacking promoter sequences (JM102/L1.3ΔCMV) could not retrotranspose in cultured cells

(Figure 2.4A) (Wei et al. 2001). These data suggest that the intra-5'UTR splicing event present in $SpIRE_{97/622}$ severely compromises L1 5'UTR promoter activity and, by proxy, L1 retrotransposition.

5'UTR/ORF1 SpIREs rely on ORF1p supplied in trans for mobilization

The retrotransposition of an mRNA derived from a 5'UTR/ORF1 splicing event would generate a SpIRE (*e.g.*, $SpIRE_{97/976}$) that contains a defective promoter and, if transcribed and translated, would produce amino-terminal truncated versions of ORF1p. If the truncated version(s) of ORF1p were non-functional, we reasoned that the 5'UTR/ORF1 splicing event would lead to an L1 mRNA that is compromised for an initial round of retrotransposition in *cis*. Indeed, RNAs derived from pPL_{97-976}/L1.3 could not retrotranspose despite expression being driven by the CMV promoter (Figure 2.4B).

We next hypothesized that a source of wild-type ORF1p would be required to act in *trans* to promote the retrotransposition of L1 mRNAs containing a 5'UTR/ORF1 splicing event. To test this hypothesis, we co-transfected pPL_{97-976}/L1.3 with a series of retrotransposition "driver" plasmids that lack the *mneoI* retrotransposition indicator cassette and either express or do not express WT ORF1p (Wei et al. 2001; Alisch et al. 2006). The co-transfection of pPL_{97-976}/L1.3 with "driver" plasmids that express WT ORF1p (pJBM561 (a monocistronic ORF1p expression vector)), pJM101/L1.3NN, or pJM105/L1.3NN), promoted a low level of pPL_{97-976}/L1.3 mRNA retrotransposition in *trans* (Figure 2.4C; columns 1, 2, and 3, respectively). By comparison the co-transfection of pPL_{97-976}/L1.3 with "driver" plasmids that do not express ORF1p (pO2NN (a monocistronic ORF2p expression vector) or pCEP4) did not promote pPL_{97-976}/L1.3 mRNA retrotransposition (Figure 2.4C; columns 4 and 5, respectively) (Alisch et al. 2006). Thus, the expression of ORF1p, but not ORF2p, can promote low levels of retrotransposition of pPL_{97-976}/L1.3 mRNA in *trans*.

Discussion

The continued evolutionary success of L1 requires the reiterative retrotransposition of full-length L1 RNAs. Previous studies have provided compelling evidence that L1 ORF1p and L1 ORF2p exhibit *cis*-preference and preferentially bind to their encoding mRNA to promote its retrotransposition (Figure 2.5A) (Martin 1991; Hohjoh and Singer

1996; Wei et al. 2001; Kulpa and Moran 2005; Kulpa and Moran 2006; Doucet et al. 2015). Thus, it was surprising when Belancio and colleagues identified a small number of L1 retrotransposition events in the HGR that apparently were derived from spliced L1 RNAs (Belancio et al. 2006; Belancio et al. 2008). Here, we confirmed and extended those findings and report a novel class of retrotransposed L1s that are derived from an L1 RNA containing an intra-5'UTR splicing event ($SpIRE_{97/622}$; Figure 1). $SpIRE_{97/622}$ is 10 times more prevalent than SpIREs previously identified, present at 116 copies in the HGR, and comprises ~1.8% of the full-length L1 retrotransposition events in the L1PA1-PA6 subfamilies accumulated during the past ~27 million years (MY)(Figure 2.6).

The retrotransposition of spliced L1 RNAs leads to the generation of SpIREs that are compromised for subsequent rounds of retrotransposition. L1 RNAs that contain intra-5'UTR splicing events can produce ORF1p and ORF2p and undergo an initial round of retrotransposition in *cis* (Figure 2.4A). However, the resultant SpIREs lack *cis*-acting sequences required for efficient L1 transcription (Figure 2) (Tchenio et al. 2000; Yang et al. 2003; Athanikar et al. 2004); thus, they are compromised for subsequent rounds of retrotransposition (Figure 2.4A; Figure 2.5B). By comparison, L1 RNAs containing 5'UTR/ORF1 splicing produce non-functional, amino-terminal truncated versions of ORF1p (Figure 2.3C; Figure 2.8A, B). As a result, these RNAs are retrotransposition-defective in *cis* and must rely on exogenous sources of ORF1p to promote their retrotransposition in *trans* (Figure 2.4B, C; Figure 2.5C). In the rare cases where *trans*-complementation occurs, the resultant 5'UTR/ORF1 SpIREs will lack *cis*-acting sequences required for efficient L1 transcription and, if transcribed, produce non-functional versions of ORF1p, making it highly unlikely that they will undergo subsequent rounds of retrotransposition (Figure 2.5C).

The above data strongly indicate that SpIREs represent evolutionary 'dead ends' in the L1 amplification process. It is possible that a small number of SpIREs could give rise to new L1 retrotransposition events. For example, the insertion of $SpIRE_{97/622}$ downstream of a cellular promoter could, in principle, enhance its expression and subsequent retrotransposition. However, any resultant retrotransposition events would contain a defective promoter and would still be compromised for subsequent rounds of

retrotransposition. Indeed, the examination of 3' transduction sequences associated with SpIREs did not uncover any examples where one SpIRE served as a progenitor for a subsequent retrotransposition event. Thus, we conclude that splicing negatively affects L1 retrotransposition.

The three classes of SpIREs examined in our study each use a common splice donor site (SD: $G_{98}U_{99}$), but different splice acceptor sites (SA: $A_{620}G_{621}$, SA: $A_{788}G_{789}$, or SA: $A_{974}G_{975}$)(Belancio et al. 2006; Belancio et al. 2008). These findings raise the following question: if splicing adversely affects L1 retrotransposition, why is the $G_{98}U_{99}$ splice donor site retained in L1 RNA? The $G_{98}U_{99}$ splice donor site has been conserved in the L1PA1-L1PA10 subfamilies for at least 46 MY (Figure 2.6B) (Khan et al. 2006) and resides within a core binding site for the RUNX3 transcription factor (Yang et al. 2003). Indeed, previous studies indicated that RUNX3 is required for efficient L1 transcription (Yang et al. 2003) and we found that mutating the splice donor sequence leads to an ~5-fold reduction in L1 promoter activity (Figure 2B). Together, these finding strongly suggest that the importance of the RUNX3 binding site in L1 expression far outweighs the cost of harboring the splice donor site (SD: $G_{98}U_{99}$) in the L1 5'UTR.

Despite the evolutionary conservation of the $G_{98}U_{99}$ splice donor, northern blotting and RT-PCR experiments revealed that the vast majority of L1 transcripts do not undergo splicing (Figures 2.2B and 2.2D). SpIREs only are formed when L1 RNAs containing rare splicing events undergo retrotransposition. The reason(s) why $G_{98}U_{99}$ is not efficiently utilized as a functional splice donor site require elucidation. However, it is possible that the $G_{98}U_{99}$ sequence is sequestered into a secondary structure in L1 RNA that restricts its access to U1 snRNA (reviewed in (Hastings and Krainer 2001; Buratti and Baralle 2004)). Indeed, such a scenario provides a plausible mechanism for how L1 can tolerate a functional splice donor sequence in its mRNA and could, in part, explain why SpIREs only represent ~2% of full-length L1 retrotransposition events that occurred during the past ~27 MY.

The notion that sequences within the L1 5'UTR are subject to selective pressure has precedent. A recent study demonstrated that a Krüppel-associated Box-containing Zinc-Finger Protein (ZNF93) could bind sequences within the 5'UTR of full-length L1s from

the L1PA3 subfamily to repress their expression (Jacobs et al. 2014). By comparison the 5'UTRs of full-length L1s from the L1PA2 and currently amplifying L1PA1 subfamilies harbor a 129-bp deletion that eliminates the ZNF93 binding site, allowing them to escape ZNF93-mediated repression (Jacobs et al. 2014).

Our RT-PCR experiments using transfected L1/firefly luciferase expression vectors only uncovered evidence of rare $SpIRE_{97/790}$ splicing events. Intriguingly, we only identified genomic $SpIRE_{97/790}$ retrotransposition events in the L1PA2 and L1PA1 subfamilies despite the fact that the SA: $A_{788}G_{789}$ sequence has been conserved from the L1Hs to L1PA15B subfamilies for ~70 MY (Figure 2.6B) (Khan et al. 2006). We propose that the 129bp deletion may have allowed the $SpIRE_{97/790}$ splice acceptor ($A_{916}G_{917}$ in L1PA3) to come into closer proximity ($A_{786}G_{787}$ in L1PA2) with a putative splicing branch point sequence in L1PA1-PA2 RNAs ($ACCTCAC_{761\text{-}767}$ in L1PA2) (Figure 2.9). If so, the 129-bp deletion present not only allowed L1PA2 and L1Hs 5'UTRs to escape ZNF93 mediated repression, but also may have altered the intra-5'UTR splicing dynamics, leading to new, low-level splicing events that can lead to the generation of new SpIREs (Figure 2.9).

In sum, our data strongly indicate that L1 mRNA splicing is detrimental to L1 retrotransposition and further strengthen the hypothesis that ORF1p and ORF2p predominantly retrotranspose their encoding full-length L1 RNAs to new genomic locations in *cis*. In addition, we demonstrated that, despite harboring evolutionarily conserved functional splice donor and splice acceptor sites within their 5'UTR, the vast majority of L1 transcripts evade splicing. These data provide insights into the evolutionary dynamics of the L1 5'UTR and raise the intriguing possibility that host factors that promote L1 splicing or alter L1 splicing profiles may represent a mechanism by which the cell can restrict the expression of full-length L1 RNA and unabated L1 retrotransposition.

Materials and Methods

E. coli and the Propagation of Plasmids

All plasmids were propagated in DH5α *E. coli* (genotype: F- φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rk-, mk+) *pho*A *sup*E44 λ- *thi*-1 *gyr*A96 *rel*A1) (Invitrogen, Carlsbad, CA). Competent cells were generated as described previously (Inoue et al. 1990). Plasmids were prepared using the Plasmid Midi Kit (Qiagen, Germany) according to the protocol provided by the manufacturer.

Cell Lines and Cell Culture Conditions

HeLa-JVM cells were cultured in high glucose Dulbecco's Modified Eagle Medium (DMEM) lacking pyruvate (Invitrogen). DMEM was supplemented with 10% fetal bovine calf serum (FBS) and 1X penicillin/streptomycin/glutamine to create DMEM-complete medium as described previously (Moran et al. 1996). HeLa-JVM cells were grown in a humidified tissue culture incubator (Thermo Scientific, Waltham MA) at 37^{o}C in the presence of 7% CO_2.

BLAT Searches and SpIRE Sequence Curation

BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) was used to screen build 37 (GRCh37/hg19) of the UCSC genome browser (https://genome.ucsc.edu) with the repeat masker track “on” using 100 bp *in silico* probes that spanned (50 bases upstream and downstream) the splice junctions of $SpIRE_{97/622}$, $SpIRE_{97/790}$ and $SpIRE_{97/976.}$ The *in silico* probes were designed using the L1.3 sequence (Accession #L19088) as a template. Putative SpIREs shared >95% sequence identity to the *in silico* probes.

Putative SpIREs were downloaded from the UCSC genome browser and manually curated with the aid of repeat masker (http://repeatmasker.org). Each sequence was inspected to ensure it contained a splicing event and represented a *bona fide* SpIRE. For four events that were prematurely 3’ truncated, we analyzed 4kb of genomic DNA flanking the 3’ end of the SpIRE to determine if it shared >95% sequence identity with L1.3 using the Serial Cloner alignment tool (http://serialbasics.free.fr/Home/Home.html).

We were unable to identify any L1 sequence in the flanking DNA; thus we cannot determine the reason for the apparent 3' truncation in these four SpIREs. Structural hallmarks of L1 integration events that occur by canonical TPRT (*i.e.*, the presence of target site duplications, a 3' poly(A) tract, L1-mediated sequence transductions, and the L1 integration site) were determined manually by analyzing sequences flanking the 5' and 3' ends of each SpIRE (Goodier et al. 2000; Szak et al. 2002; Gilbert et al. 2005). Sequences are named based on the class of SpIRE ($SpIRE_{97/622}$, $SpIRE_{97/790,}$ or $SpIRE_{97/976}$) and a corresponding number for easy referral between Tabale 2.1 and data set 1 (for example; $SpIRE_{97/622}$-1 is the first of the analyzed 116 $SpIRE_{97/622}$ sequences).

Determining the conservation of splice donor and splice acceptor sites

Khan et al. 2006 provided full-length L1 consensus sequences of L1PA1(L1Hs) through L1PA16 and assembled an alignment of the respective 5'UTRs (Khan et al. 2006). We manually inspected these alignments to determine the oldest L1 subfamily that contained the 5'UTR splice donor/splice acceptor sequences utilized in generating the reported SpIREs. We next determined the conservation of the ORF1 splice acceptor sequence (SA: $A_{974}G_{975}$) by aligning full-length L1 consensus sequences using the ClustalW alignment function (Thompson et al. 2002; Khan et al. 2006) from the MegAlign (http://www.dnastar.com/t-megalign.aspx) software. As with the 5'UTR, we manually inspected the resulting alignment to determine the oldest L1 subfamily that contained the ORF1 splice acceptor sequence (SA: $A_{974}G_{975}$).

Identification of putative branch point sequences

To identify putative branch point sequences, we downloaded full-length 5'UTR sequences as described above (Khan et al. 2006), and submitted them for analysis using the Human Splicing Finder v3.0 online prediction program (http://www.umd.be/HSF3/HSF.html) (Desmet et al. 2009). The resultant analyses identify potential SD, SA, and branch point sequences and assigns consensus value scores for each motif (Desmet et al. 2009). Motif scores greater than 80 represent "strong" splice sites; sequences with scores less than 80 represent "weaker" splice sites. The 5'UTR sequence of each L1 subfamily was uploaded and analyzed by the general 'Analyze a sequence' function. We then matched predicted branch points with

the known splice acceptor SA: $A_{786}G_{787}$ (L1PA2) using the criteria from Gao et al. 2008, where 100% of branch points are at positions −50 to −5 relative to the last nucleotide of the intron (in this case, the -1 position is G_{789}). Using this procedure, we identified the strong branch point ($A_{761}C_{762}C_{763}T_{764}C_{765}A_{766}C_{767}$) with a score of 95.75 that is 20 bp upstream of the SA: $A_{786}G_{787}$ in the L1PA2 subfamily. This putative branch point sequence is 147 bp upstream of the conserved SA: $A_{916}G_{917}$ in the L1PA3 family (See Figure 2.9).

L1 Expression Constructs

The following L1 constructs contain a derivative of a retrotransposition-competent L1 (L1.3, accession number L19088 (Dombroski et al. 1993; Sassaman et al. 1997)) cloned into the pCEP4 plasmid backbone (Life Technologies) unless indicated otherwise. Cloning strategies used to create these constructs are available upon request.

pJM101/L1.3: contains a full-length version of L1.3 in the pCEP4 backbone. The 3'UTR of L1.3 contains the *mneoI* retrotransposition indicator cassette (Dombroski et al. 1993; Moran et al. 1996; Sassaman et al. 1997).

p*JM101/L1.3ΔCMV*: is identical to *pJM101/L1.3*, but the CMV promoter was deleted from the pCEP4 plasmid (Dombroski et al. 1993; Moran et al. 1996; Sassaman et al. 1997).

pJM101/L1.3NN: is a derivative of *pJM101/L1.3* that lacks the *mneoI* retrotransposition indicator cassette (Wei et al. 2001).

pDK101/L1.3: is a derivative of *pJM101/L1.3* that expresses a version of ORF1p that contains a T7 *gene10* epitope tag on its carboxyl-terminus (Kulpa and Moran 2005).

pJM105/L1.3: is identical to *pJM101/L1.3*, but contains a D702A missense mutation in the ORF2p RT active site (Wei et al. 2001).

pJM105NN: is a derivative of *pJM105/L1.3* that lacks the *mneoI* retrotransposition indicator cassette (Wei et al. 2001).

pJM102/L1.3: is a derivative of *pJM101/L1.3* that lacks the L1 5'UTR (Morrish et al. 2002).

pJM102/L1.3ΔCMV: is identical to *pJM102/L1.3*, but the CMV promoter was deleted from the pCEP4 plasmid (Wei et al. 2001).

$pPL_{97\text{-}622}$/L1.3: is a derivative of pJM101/L1.3 that contains a 524 intra-5'UTR deletion (L1.3 nucleotides 98-621) present in $SpIRE_{97/622}$ (Beck et al. 2010).

$pPL_{97\text{-}622}$ /L1.3ΔCMV: is identical to *$pPL_{97\text{-}622}$/L1.3*, but the CMV promoter was deleted from the pCEP4 plasmid.

$pPL_{97\text{-}976}$/L1.3: is a derivative of pJM101/L1.3 that contains an 878 bp 5'UTR/ORF1 deletion (L1.3 nucleotides 98-975) present in $SpIRE_{97/976}$.

$pPL_{97\text{-}976}$/L1.3-T7: is a derivative of *$pPL_{97\text{-}976}$/L1.3* that expresses a version of ORF1p that contains a T7 *gene10* epitope tag on its carboxyl-terminus.

pORF2/L1.3NN: is a monocistronic L1 ORF2 expression plasmid that lacks the *mneoI* retrotransposition indicator cassette (Alisch et al. 2006).

pJBM561: is a monocistronic L1 ORF1 expression plasmid that lacks the *mneoI* retrotransposition indicator cassette. Plasmid created by Dr. John B. Moldovan.

pCEP/GFP: is a pCEP4-based plasmid that expresses a humanized Renilla green fluorescent protein (hrGFP) from phrGFP-C (Stratagene). A CMV promoter drives the expression of the *hrGFP* gene (Alisch et al. 2006).

Luciferase Expression Constructs

The following plasmids are based on the pGL4.11 promoterless firefly luciferase expression vector (Promega, Madison, WI). Oligonucleotides and cloning strategies used to create these constructs are available upon request.

pPL_{WT}LUC: is a derivative of pGL4.11 that contains the wild type L1.3 5'UTR upstream of the firefly luciferase reporter gene.

$pPL_{97\text{-}622}$LUC: is a derivative of pGL4.11 that contains the $pPL_{97\text{-}622}$/L1.3 5'UTR deletion derivative upstream of the firefly luciferase reporter gene.

pPL$_{SDm}$LUC: is a derivative of *pPL$_{WT}$LUC* that contains a $U_{99}C$ splice donor mutation in the L1.3 5'UTR upstream of the firefly luciferase reporter gene.

pPL$_{SAm}$LUC: is a derivative of *pPL$_{WT}$LUC* that contains an $A_{620}C$ splice acceptor mutation in the L1.3 5'UTR upstream of the firefly luciferase reporter gene.

pGL3-Control: is an expression plasmid where the SV40 promoter drives firefly luciferase transcription (Promega).

pRL-TK: is an expression plasmid where the HSV-TK promoter drives Renilla luciferase transcription (Promega).

RNA isolation

Briefly, $8X10^6$ HeLa-JVM cells were seeded into a T-175 Falcon tissue culture flask (BD Biosciences, San Jose, CA). On the following day, transfections were conducted using the FuGene HD transfection reagent (Promega, Madison, WI). The transfection reactions contained 1 mL of Opti-MEM® (Life Technologies), 120μl of the FuGene HD transfection reagent, and 20μg of plasmid DNA per flask. The tissue culture medium was changed 24 hours post-transfection. The cells were collected 48 hours post-transfection. Briefly, cells were washed in ice-cold 1X phosphate buffered saline (PBS) (Life Technologies). The cells then were scraped from the tissue culture flasks, transferred to a 15 mL conical tube (BD Biosciences), and centrifuged at 3000 x g for 5 minutes at 4^oC. Cell pellets were frozen at -20^oC overnight. The frozen pellets were thawed and total RNA was prepared using the TRIzol reagent following the protocol provided by the manufacturer (Life Technologies). Poly(A) RNAs then were isolated from the total RNAs using a Oligotex mRNA Midi Kit (Qiagen), suspended in UltraPure™ DNase/RNase-Free distilled water (Thermo Fisher Scientific, Waltham, MA), and quantified using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific).

Northern Blots

Northern blot experiments were conducted using the NorthernMax-Gly Kit (Thermo Fisher Scientific) following the protocol provided by the manufacturer. Briefly, aliquots of

poly(A) RNAs (2 μg) were incubated for 30 minutes at 50°C in Glyoxal Load Dye (containing DMSO and ethidium bromide) and then were separated on a 1.2% agarose gel. The RNAs were transferred by capillary action to a Hybond-N nylon membrane (GE Healthcare, Marlborough, MA) for four hours, and cross-linked to the membrane using the Optimum Crosslink setting of a Stratalinker (Stratagene, LaJolla, CA). Membranes were then baked at 80°C for 15 minutes. Membranes were prehybridized for approximately four hours at 68°C in NorthernMax® Prehybridization/Hybridization Buffer (Thermo Fisher Scientific) and then were incubated overnight at 68°C with a strand specific RNA probe (final concentration of probe ~ 3×10^6 cpm/ml). The following day, the membranes were washed once with low stringency wash solution (2x saline sodium citrate (SSC), 0.1% sodium dodecylsulfate (SDS)) and then twice with high stringency wash solution (0.1x SSC, 0.1% SDS). The washed membranes were placed in a film cassette (Thermo Fisher Scientific, Autoradiography Cassette FBCA 57) and exposed to Amersham Hyperfilm ECL (GE Healthcare) overnight at -80°C. Films were developed using a JP-33 X-Ray Processor (JPI America Inc., New York, NY).

Preparation of Northern Blot Probes

Strand-specific αP^{32}-UTP radiolabeled riboprobes were generated using the MAXIscript T3 system (Thermo Fisher Scientific). Briefly, oligonucleotide primers were used to PCR amplify portions of the L1.3 5'UTR (Moldovan and Moran 2015)(L1.3 nucleotides 1-100 or L1.3 nucleotides 103-330) or the 3' end of the luciferase gene (see below). The resultant PCR products were separated on a 1% agarose gel and were purified using QIAQuik gel extraction (Qiagen). Notably, a T3 RNA polymerase promoter sequence was included on the reverse primer used to generate the antisense riboprobe (underlined below). The labeling reaction was carried out at 37°C using the following reaction conditions: 500ng of gel purified DNA template, 2 μL of transcription buffer supplied by the manufacturer, 1 μL each of unlabeled 10 mM ATP, CTP, GTP, 5 μL of αP^{32}-UTP (10 mCi/mL), and 2 μL of T3 RNA polymerase. The reaction components then were mixed and brought to a total volume of 20μL using nuclease-free water in a 1.5 mL Eppendorf tube, which was incubated at 37°C for 10 minutes in a heating block. Unincorporated nucleotides were subsequently depleted using the

Ambion® NucAway™ Spin Columns (Thermo Fisher Scientific) following the protocol provided by the manufacturer. To generate a control β-actin riboprobe, the pTRI-β-actin-125-Human Antisense Control Template (Applied Biosystems) was used in T3 labeling reactions. Biological triplicates of each northern blot exhibited similar results.

Oligonucleotide sequences used to generate northern blot probes:

L1.3 5'UTR 1-100 Sense: 5'-GGAGCCAAGATGGCCGAATAGGAACAGCT-3'

L1.3 5'UTR 1-100 AS: 5'-AATTAACCCTCAAAGGGACCTCAGATGGAAATGCAG-3'

L1.3 5'UTR 103-336 Sense: 5'-GGGTTCATCTCACTAGGGAGTG-3'

L1.3 5'UTR 103-336 AS: 5'-AATTAACCCTCACTAAAGGGTATAGTCTCGTGGTGCGCCG-3'

Quantification of Northern Blots

Northern blot bands were quantified using the ImageJ software (https://imagej.nih.gov/ij/ software) (Schneider et al. 2012). The intensity of the bands in the $pPL_{WT}LUC$ and $pPL_{SDm}LUC$ lanes were determined and normalized to the actin loading control. Three independent northern blots were subject to quantification. We then computed that average intensity of the bands and calculated a standard deviation. As reported in the text, the steady state level of $pPL_{SDm}LUC$ RNA is ~18% the level of $pPL_{WT}LUC$ RNA with a standard deviation of +/-3.1%.

Dual Luciferase Assays

Luciferase assays were performed using the Dual-Luciferase® Reporter Assay System (Promega, Madison, WI) following the manufacturers protocol. Briefly, $2X10^4$ HeLa cells were plated into each well of a 6-well plate (BD Biosciences). Approximately 24 hours later each well was transfected using a transfection mixture of 100µl Opti-MEM® (Life Technologies), 3 µl of FuGENE6 transfection reagent (Promega), and 1 µg plasmid DNA (0.5 µg of a firefly luciferase test plasmid and 0.5 µg of an internal control Renilla luciferase expression). Each transfection was performed as a technical duplicate (*i.e.*, in two wells of a 6-well tissue culture plate). Approximately 24 hours post-transfection, the transfected cells were washed once with ice-cold 1X PBS and the cells

in each well were subjected to lysis for 15 minutes at room temperature using 500µl of the 1X Passive Lysis Buffer supplied by the manufacturer. Following homogenization of the lysate by manual pipetting, 60µl of the lysate from each well of the 6-well tissue culture plate was distributed equally in 3-wells of a 96-well white opaque, optically transparent top plate (BD Biosciences), allowing six luminescence readings for each transfection condition (six technical replicates – 3 readings per well of a 6-well plate). The 96-well plate then was subject to luciferase detection using a GloMax®-Multi Detection System (Promega) following the protocol provided by the manufacturer. Luminescence readings from the six technical replicates were averaged to give a single normalized luminescence reading (NLR). This assay then was performed in biological triplicate, yielding three independent NLRs. The resultant data were subsequently analyzed using a Student's one-tailed t-test. Error bars indicate the standard deviation. Luminescence readings from lysis buffer alone and from lysates derived from untransfected cells were included used as negative controls.

Reverse Transcription (RT)-PCR

Poly(A) selected mRNA from transfected HeLa-JVM cells in a T-175 tissue culture flask were collected as previously described for northern blots. The resultant RNAs were subjected to targeted reverse transcriptase-PCR (RT-PCR) using SuperScript® III One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase (Thermo Fisher Scientific) following the protocol provided by the manufacturer. The REVLUC primer was used to synthesize first strand cDNA. The FWD5'UTR and REVLUC primers then were used to amplify the resultant cDNAs (see sequences below). The RT-PCR products were separated on a 1.2% agarose gel, excised from the gel using QIAQuik gel extraction (Qiagen), and were cloned using the TOPO TA Cloning Kit (Thermo Fisher Scientific). Sanger DNA sequencing performed at the University of Michigan DNA Sequencing Core verified the cDNA sequences in the resultant plasmids. Biological triplicates of this experiment yielded similar results.

Oligonucleotide sequences used in the RT-PCR experiments:

FWD5'UTR: 5'-GGAACAGCTCCGGTCTACAGCTCCC-3'

REVLUC: 5'-CCCTTCTTAATGTTTTTGGCATCTTCC-3'

Protein collection

The plating and transfection of HeLa-JVM cells in T-175 tissue culture flasks was performed as detailed above in the RNA isolation section except that HeLa-JVM cells were subjected to selection in DMEM-complete medium supplemented with 200μg/ml of hygromycin B (Thermo Fisher Scientific) 48 hours post-transfection and the selection medium was changed every other day for seven days. The hygromycin resistant HeLa-JVM cells were harvested nine days post-transfection as described in the RNA isolation section. The cell pellets were frozen overnight. The following day, pellets were lysed for 15 minutes on ice by incubation in 0.5 mL of lysis buffer: 10% glycerol, 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% NP-40 (IGPAL) (Sigma-Aldrich, St. Louis, MO), and 1X Complete Mini EDTA-free Protease Inhibitor Cocktail (Roche Applied Science, Germany). The resultant protein lysates then were centrifuged at 15,000 x g for 30 minutes to clear the lysate. The resultant supernatant (approximately 0.4mls) was designated as the whole cell lysate (WCL). Alternatively, the supernatant fraction was subject to RNP collection as previously described (Kulpa and Moran 2005). Briefly, 200μl of the WCL was layered onto a sucrose solution cushion (6 mL of 17% sucrose, bottom layer, followed by 4 mL of 8.5% sucrose, top layer, overlaid by 200 μl of the WCL) and ultracentrifuged at 178,000 x g for two hours at 4^{o}C. Following ultracentrifugation the supernatant was aspirated and the resultant RNP pellet was suspended in 100 μl of water supplemented with 1X Complete Mini EDTA-free Protease Inhibitor Cocktail (Roche Applied Science). Bradford assays (Bio-Rad Laboratories, Hercules, CA) were used to determine protein concentrations. WCLs generally yielded 15-19 μg/μl of protein. RNP preparations yielded 6-10 μg/μl of protein. The protein samples were stored at -80^{o}C.

Western blots

Protein samples were collected as described above and then were incubated with a 2X solution of NuPAGE reducing buffer (containing 1.75-3.25% lithium dodecyl sulfate and 50 mM dithiothreitol (DTT)) (ThermoFisher Scientific). An aliquot (20 μg) of the reduced proteins were incubated at 100^{o}C for 10 minutes and then were separated by

electrophoresis on 10% precast mini-PROTEAN® TGX gels (Bio-Rad Laboratories, Hercules, CA) run at 200V for 1 hour in 1X Tris/Glycine/SDS (25 mM Tris-HCL, 192 mM glycine, 0.1% SDS, pH 8.3) buffer (Bio-Rad Laboratories). Transfer was performed using the Trans-Blot® Turbo™ Mini PVDF Transfer Packs (BioRad Laboratories) with the Trans-Blot® Turbo™ Transfer System (BioRad Laboratories) at 25V for 7 minutes. The resultant membranes then were cut at the 75 kDa marker using the Precision Plus Protein™ Kaleidoscope™ marker (Bio-Rad Laboratories) as a guide. The membranes then were incubated at room temperature in blocking solution (containing 1X PBS and 5% dry low-fat milk) (Kroger, Cincinnati, OH). The eIF3 antibody (Santa Cruz Biotechnology Inc. (SC-28858)) was used at a 1:1,000 dilution to probe membranes containing to detect eIF3 at 110 kDa as a loading control. The α-N-ORF1p (Moldovan and Moran 2015) antibody (directed against ORF1p amino acids 31-49; EQSWMENDFDELREEGFRR), α-C-ORF1p (directed against ORF2p amino acids 319-338; EALNMERNNRYQPLQNHAKM), and anti-T7 (Merck Millipore 69048 T7•Tag® Antibody HRP Conjugate) antibodies were used at 1:10,000, 1:2,000 and 1:5,000 dilutions, respectively to probe membranes for ORF1p. Antibody hybridizations were carried out overnight at 4°C in blocking solution. The blots were washed three times with 1X PBS, 0.1% Tween-20 (Sigma Aldrich) and then were incubated with a 1:5,000 dilution of secondary Amersham ECL HRP Conjugated Donkey anti-rabbit IgG Antibodies (GE Healthcare Life Sciences) for 60 minutes at room temperature blocking solution. The membranes were washed three times with 1X PBS, 0.1% Tween-20 (Sigma Aldrich). The signals then were visualized using the SuperSignal™ West Pico Chemiluminescent Substrate reagent (ThermoFisher Scientific) according to the protocol provided by the manufacturer. The membranes exposed to Amersham Hyperfilm ECL (GE Healtchare) for a time that spanned five seconds to five minutes and were developed using a JP-33 X-Ray Processor (JPI America Inc.).

L1 Retrotransposition Assay

The cultured cell retrotransposition assay was conducted as described previously (Moran et al. 1996; Wei et al. 2001; Kopera et al. 2016). Briefly, 2×10^3 HeLa-JVM cells/well were plated in 6-well tissue culture dishes (BD Biosciences). Approximately 24

hours post-plating, transfections were performed using a mixture containing 100 µl Opti-MEM® (Life Technologies), 3 µl FuGENE6 (Promega) transfection reagent, and 1 µg L1 plasmid DNA per well of a 6-well plate. Approximately 24 hours post-transfection, the media was replaced with DMEM-complete medium to stop the transfection. Three days post-transfection, the tissue culture medium was replaced and the cells were grown in DMEM-complete medium supplemented with 400 µg/mL of G418 (Life Technologies) to select for retrotransposition events. After approximately 12 days of G418 selection, the resultant G418-resistant foci were washed with ice cold 1X Phosphate-Buffered Saline (PBS), fixed to the tissue culture plate by treating them for 10 minutes at room temperature in a 1X PBS solution containing 2% paraformaldehyde (Sigma Aldrich) and 0.4% glutaraldehyde (Sigma Aldrich), and stained with a 0.1% crystal violet solution for 30 minutes at room temperature to visualize the G418-resistant foci. As a transfection control, parallel 6-well tissue culture dishes of HeLa-JVM cells were co-transfected with 0.5 µg of an L1 expression plasmid and 0.5 µg of a pCEP/GFP expression plasmid (Stratagene). Three days post-transfection, the transfected HeLa-JVM cells were subjected to fluorescence detection on an Accuri C6 Flow Cytometer (BD Biosciences) to determine the transfection efficiencies (*i.e.*, the percentage of GFP-positive cells) for each experiment (Kopera et al. 2016).

The *trans*-complementation retrotransposition assay was modified slightly from a previously described protocol (Wei et al. 2001). Briefly, $2X10^5$ HeLa-JVM cells were plated into 60 mm dishes (BD Biosciences). Approximately 24 hours post-plating, transfections were performed using a mixture containing 93 µl of Opti-MEM® (Life Technologies), 6 µl of FuGeneHD (Promega), and 2µg plasmid DNA (*i.e.*, 1 µg of the L1 “reporter” plasmid and 1 µg of the L1 “driver” plasmid). Subsequent steps of the retrotransposition assay were carried out as described above. As a transfection control, parallel 60mm tissue culture dishes of HeLa-JVM cells were co-transfected with 0.5µg of an L1 “reporter” plasmid, 0.5 µg of an L1 “driver” plasmid, and 1 µg of a pCEP/GFP expression plasmid (Stratagene). Three days post-transfection, the transfected HeLa-JVM cells were subjected to fluorescence detection on an Accuri C6 Flow Cytometer (BD Biosciences) to determine the transfection efficiencies (*i.e.*, the percentage of GFP-positive cells) for each experiment (Kopera et al. 2016). The transfection efficiencies

were used to control for variability and normalize the retrotransposition efficiencies in individual transfections. At least three biological replicates were performed for each retrotransposition assay. Error bars on all retrotransposition assays represent standard deviation of technical triplicates from the indicated experiment.

Figure 2.1: LINE-1 RNA contains potential splice donor and splice acceptor sites. *A) Schematic of a full-length retrotransposition competent L1:* Top: the 5' and 3' UTRs (grey rectangles), ORF1 (yellow rectangle), and ORF2 (blue rectangle) are indicated in the cartoon. The 3'UTR ends in a poly(A) tract (A_N). The L1 is flanked by target-site duplications (black arrow heads) in genomic DNA (black helical lines). Bottom: a magnified schematic of the 5'UTR and 5' end of ORF1. The functional splice donor (SD, red) and splice acceptor (SA, green) sequences used to generate SpIREs are indicated above the gray rectangle. The position of the SD and SA sequences relative to L1.3 are indicated with superscript numbers. The relative positions of *cis*-acting transcription factor binding sequences are indicated in the 5'UTR. *B-D) Schematics of the splicing events generating* $SpIRE_{97/622}$, $SpIRE_{97/790}$, and $SpIRE_{97/976}$*:* The SD (bold red underlined GT nucleotides) and SA (bold green underlined AG nucleotides) demark the intron boundaries used to generate each class of SpIRE.

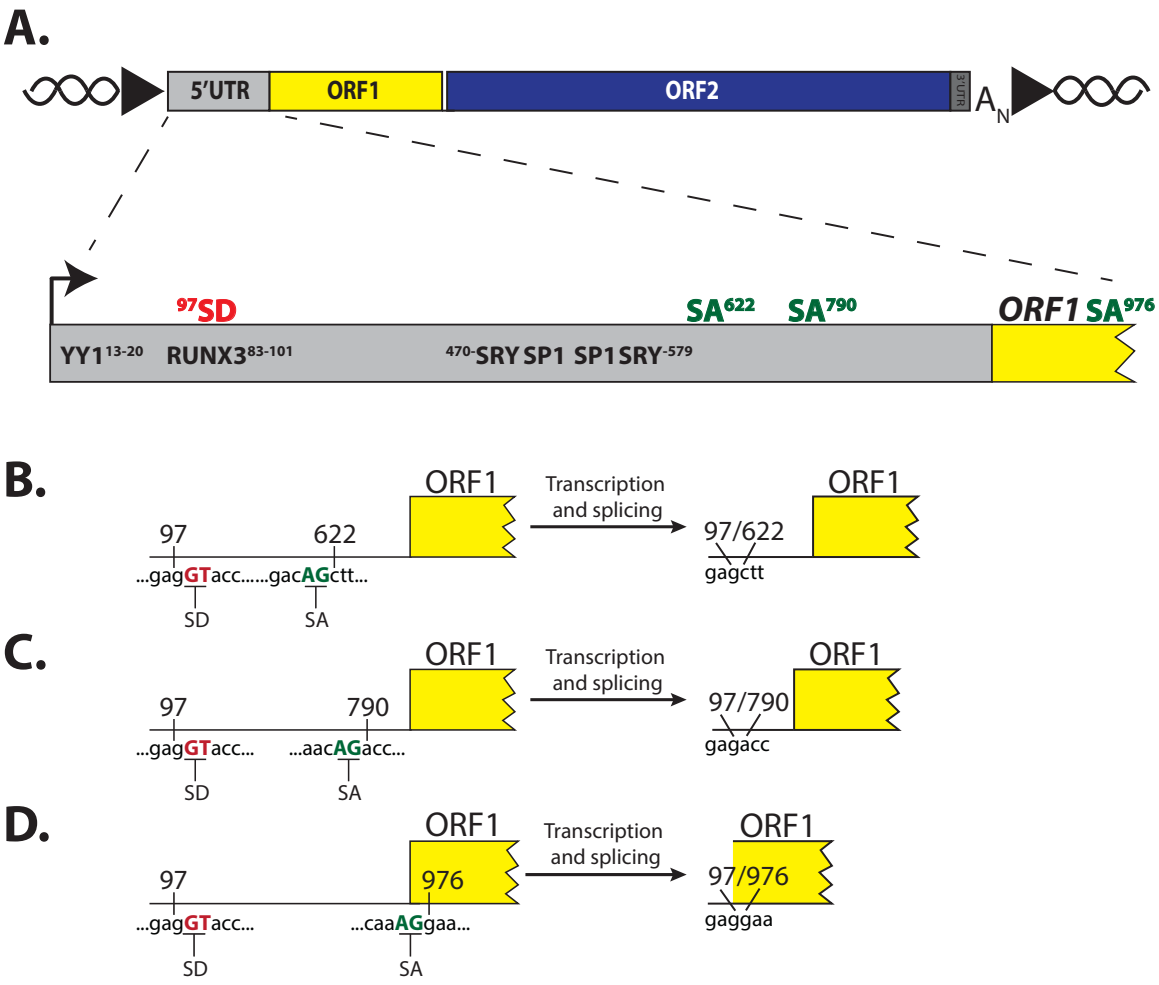


Figure 2.1: LINE-1 RNA contains potential splice donor and splice acceptor sites.

Figure 2.2: Intra-5'UTR splicing drastically reduces L1 promoter activity. *A) Schematic of the luciferase constructs and the relative position of northern blot probes:* The L1.3 5'UTR (grey bar) was used to drive the transcription of the promoterless firefly luciferase reporter gene (green bar) present in plasmid pGL4.11. The following plasmids were created: $pPL_{WT}LUC$ contains the full length L1.3 5'UTR; $pPL_{97\text{-}622}LUC$ contains the $SpIRE_{97/622}$ 5'UTR; $pPL_{SDm}LUC$ contains a $U_{99}C$ splice donor mutation (red asterisk) in the L1.3 5'UTR; $pPL_{SAm}LUC$ contains an $A_{620}C$ splice acceptor mutation (blue asterisk) in the L1.3 5'UTR. The relative positions of complementary riboprobes used in the Northern blot experiments (ribonucleotides 1-100 (purple line), ribonucleotides 103-330 (red line), and the 3' end of the luciferase gene (blue line)) are indicated below the schematic. *B) Representative northern blots:* The black arrowhead indicates the predicted size of full-length L1/luciferase RNA (~2.7 kb). Construct names are indicated above the gel lanes; UTF=untransfected HeLa-JVM cells. The probe used in the northern blot experiment is indicated below the autoradiograph. Actin served as an RNA loading control (2.1 kb). Molecular weight standards using Millenium™ RNA Markers (ThermoFisher Scientific) (kb) are indicated to the left of the autoradiograph panels. *C) Results from the luciferase assays:* The x-axis indicates the name of the luciferase expression plasmid. The y-axis indicates the relative firefly luciferase units normalized to a co-transfected Renilla luciferase control. These data represent the averages of three biological replicates. Each biological replicate contained six technical replicates. Error bars indicate the standard deviation. P-values were determined using a Student's one-tailed t-test. *D) Results from RT-PCR Assays:* Top: the relative positions of the oligonucleotide primers used to reverse transcribe (REV-LUC) and then amplify (FWD-5'UTR and REV-LUC) the L1/firefly luciferase cDNA products. Middle: a 1.2% agarose gel depicting the results from a representative qualitative RT-PCR experiment. DNA size markers (1 kb Plus DNA Ladder (Life Technologies)) are shown at the right and left of the gel. Construct names are indicated above the gel; UTF=untransfected HeLa-JVM cells, H_20=water control. The inset to the right of the gel indicates the major (* and #) and minor (+) cDNA products detected in the experiments. Bottom: a 1.2% agarose gel depicting the results from a representative experiment conducted without the addition of reverse transcriptase. DNA from $pPL_{97/622}LUC$ served as a positive control for PCR amplification. The final lane contains reverse transcriptase but no mRNA (-mRNA). RT-PCR assays were conducted at least 3 independent times.

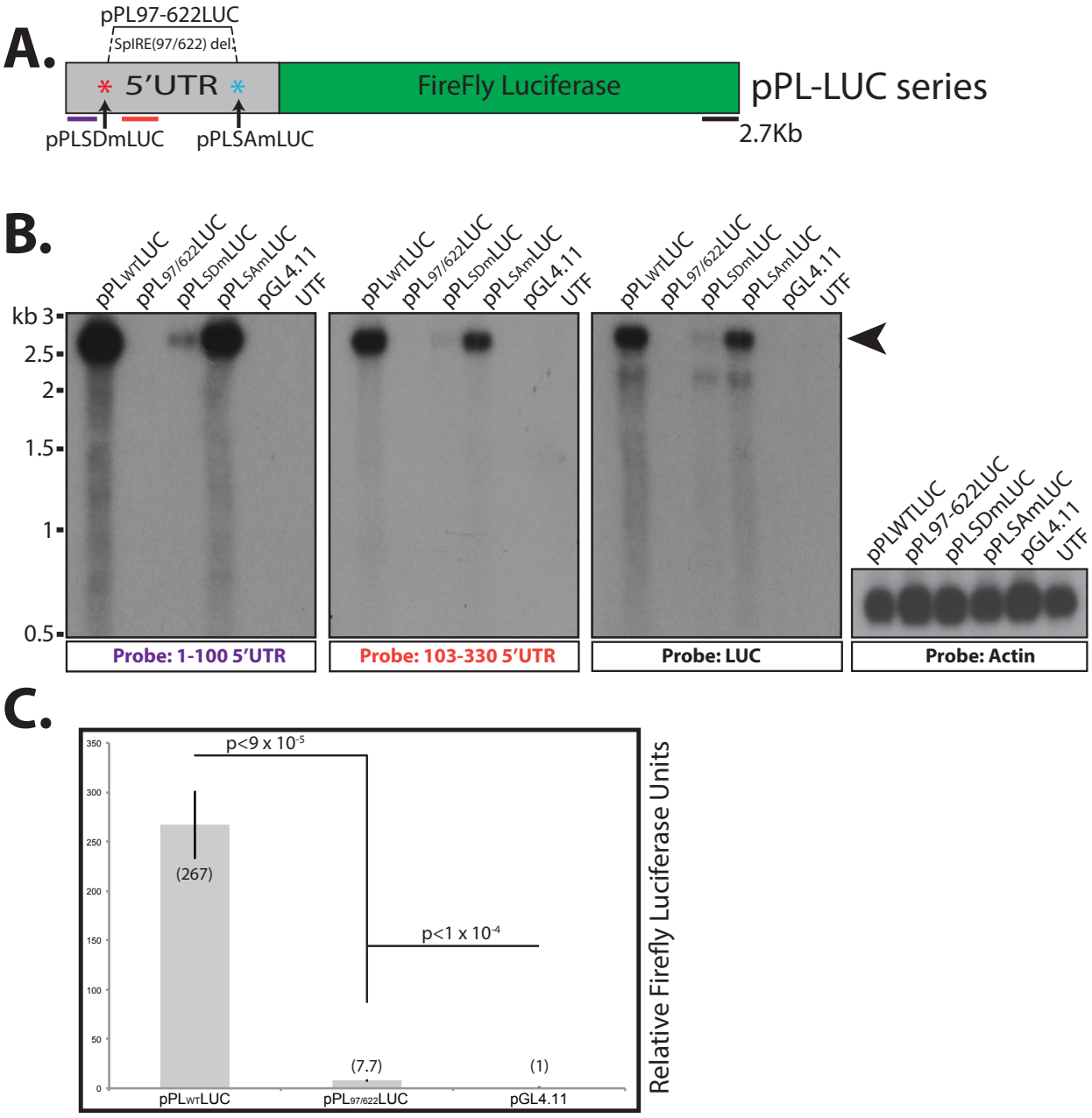


Figure 2.2: Intra-5'UTR splicing drastically reduces L1 promoter activity.

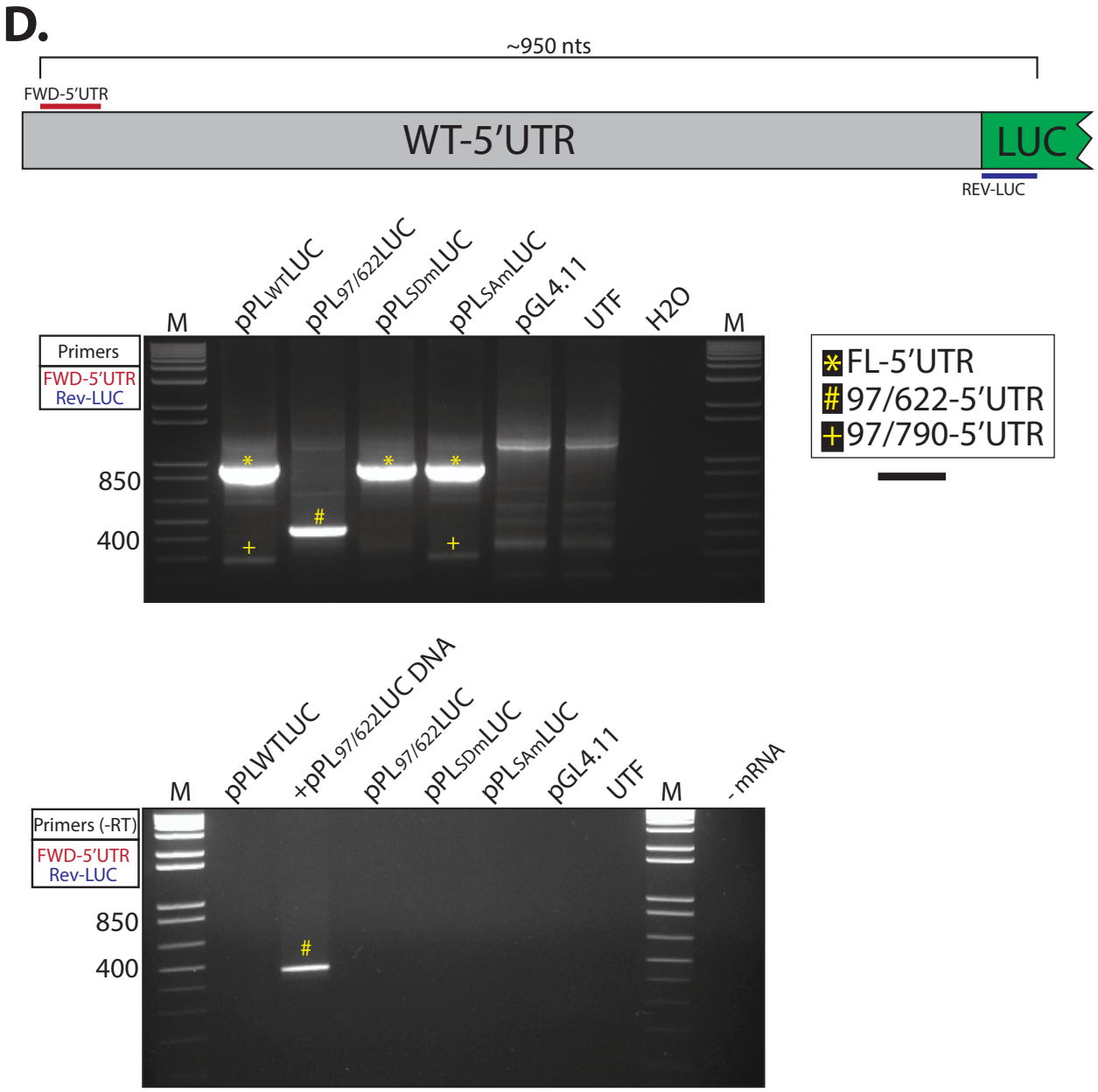


Figure 2.2: Intra-5'UTR splicing drastically reduces L1 promoter activity.

Figure 2.3: ORF1p expression from intra-5'UTR and 5'UTR/ORF1 SpIREs. *A) Schematics of the engineered L1 constructs:* The L1 5' and 3' UTRs (grey rectangles), ORF1 (yellow rectangle), and ORF2 (blue rectangle) are indicated in the cartoon. The CMV promoter (white arrowhead, left of 5'UTR), *mneoI* retrotransposition indicator cassette (green rectangle=*neo* gene sequences; black "v" line=intron), and relative positions of the $SpIRE_{97/622}$ and $SpIRE_{97/976}$ deletions (red triangles) are indicated in the figure. *B) Representative ORF1p western blots:* Molecular weight standards (Precision Plus Protein™ Kaleidoscope™(Bio-Rad)) are indicated (kDa) to the left of the gel. The black arrowhead indicates the predicted size of full-length ORF1p (~40 kD). Construct names are indicated above the gel; UTF=untransfected cells. The antibodies used in the western blot experiments are indicated to the right of the gel. The eIF3 (110 kDa) western blot served as a lysate loading control. Western blots were performed three times yielding similar results. *C) Schematic of ORF1 and representative western blots from $SpIRE_{97/976}$:* Top: the relative positions of the splice acceptor sequence at nucleotide 976 (SA, green), the canonical ORF1 initiator methionine (AUG, black, 37 kDa), the two in-frame putative initiator methionine codons (AUG, red, 33 kDa; AUG, blue, 27 kDa), and the N- and C-terminal epitopes recognized by the ORF1p antibody (Ab) are indicated in the figure. Bottom: Molecular weight standards (Precision Plus Protein™ Kaleidoscope™(Bio-Rad)) are indicated (kDa) to the left of the gels. The predicted sizes of full-length ORF1p (black arrowhead), and the N-terminal truncated ORF1p variants (orange and blue arrows, respectively) are highlighted on the gel. Construct names are indicated above the gel; pCEP/GFP=negative control. The antibodies used in the western blot experiments are indicated to the left (α-N-ORF1p) and right (α-C-ORF1p) of the gel images, respectively. The eIF3 (110 kDa) western blots served as lysate loading controls. The unlabeled band at ~25 kDa in the α-C-ORF1p experiment is a cross-reacting product. Western blots were performed three times yielding similar results.

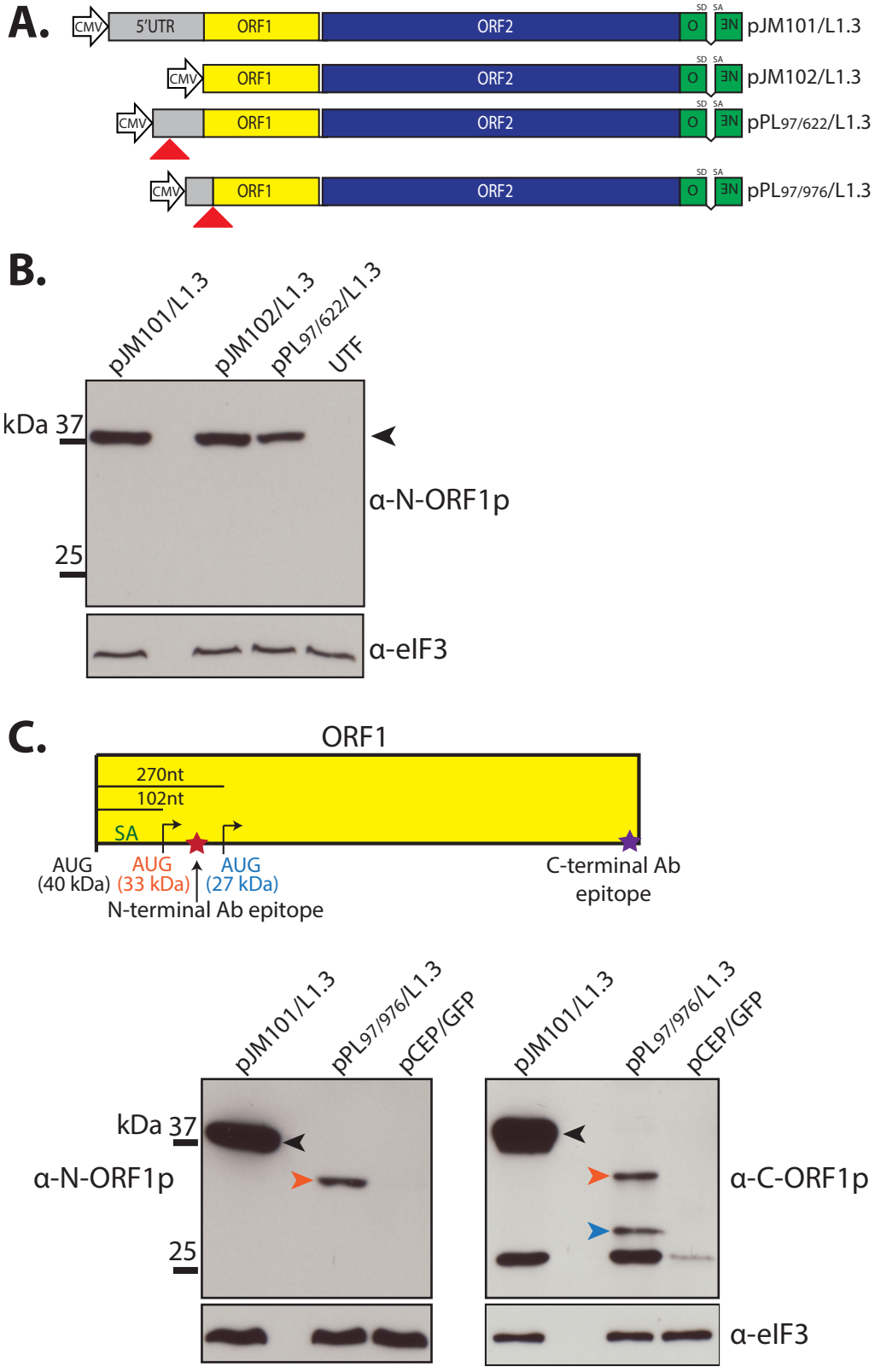


Figure 2.3: ORF1p expression from intra-5'UTR and 5'UTR/ORF1 SpIREs.

Figure 2.4: Intra-5'UTR and 5'UTR/ORF1 SpIREs are retrotransposition-defective. *A) Results from the $SpIRE_{97/622}$ retrotransposition assay:* The x-axis indicates the construct names. The y-axis indicates the relative retrotransposition efficiency (%). The CMV promoter either augments L1 expression (+CMV, black bars) or is absent from the L1 expression construct (ΔCMV, gray bars). The relative retrotransposition efficiencies are normalized to pJM101/L1.3 (set at 100%). The pJM105/L1.3 plasmid served as a negative control. The images and data are from one representative experiment. Error bars represent the standard deviations of technical triplicates for the depicted assay. Each assay was repeated three times yielding similar results. *B) Results from the $SpIRE_{97/976}$ retrotransposition assay:* The x-axis indicates the construct names. The y-axis indicates the relative retrotransposition efficiency (%). A CMV promoter augments L1 expression (+CMV, black bars). The relative retrotransposition efficiencies are normalized to pJM101/L1.3 (set at 100%). The pJM105/L1.3 plasmid served as a negative control. The images and data are from one representative experiment. Error bars represent the standard deviations of technical triplicates for the depicted assay. Each assay was repeated three times yielding similar results. *C) Results from the $SpIRE_{97/976}$* trans*-complementation assay:* The x-axis indicates the "reporter" (top text) and the "driver" (bottom text) construct names. The y-axis indicates the relative *trans*-complementation efficiency (%). The results of each assay were normalized to the $pPL_{97\text{-}976}$/L1.3 "reporter" plasmid + pJBM561 "driver plasmid" co-transfection experiment, which was set at 100%. The image at the bottom right hand side of the figure represents the efficiency of pJM101/L1.3 retrotransposition in *cis*. The $pPL_{97\text{-}976}$/L1.3 "reporter" plasmid + pCEP4 "driver plasmid" co-transfection experiment served as a negative control. The images and data are from one representative experiment. Each assay was repeated four times. Error bars represent standard deviations of technical triplicates for the depicted experiment.

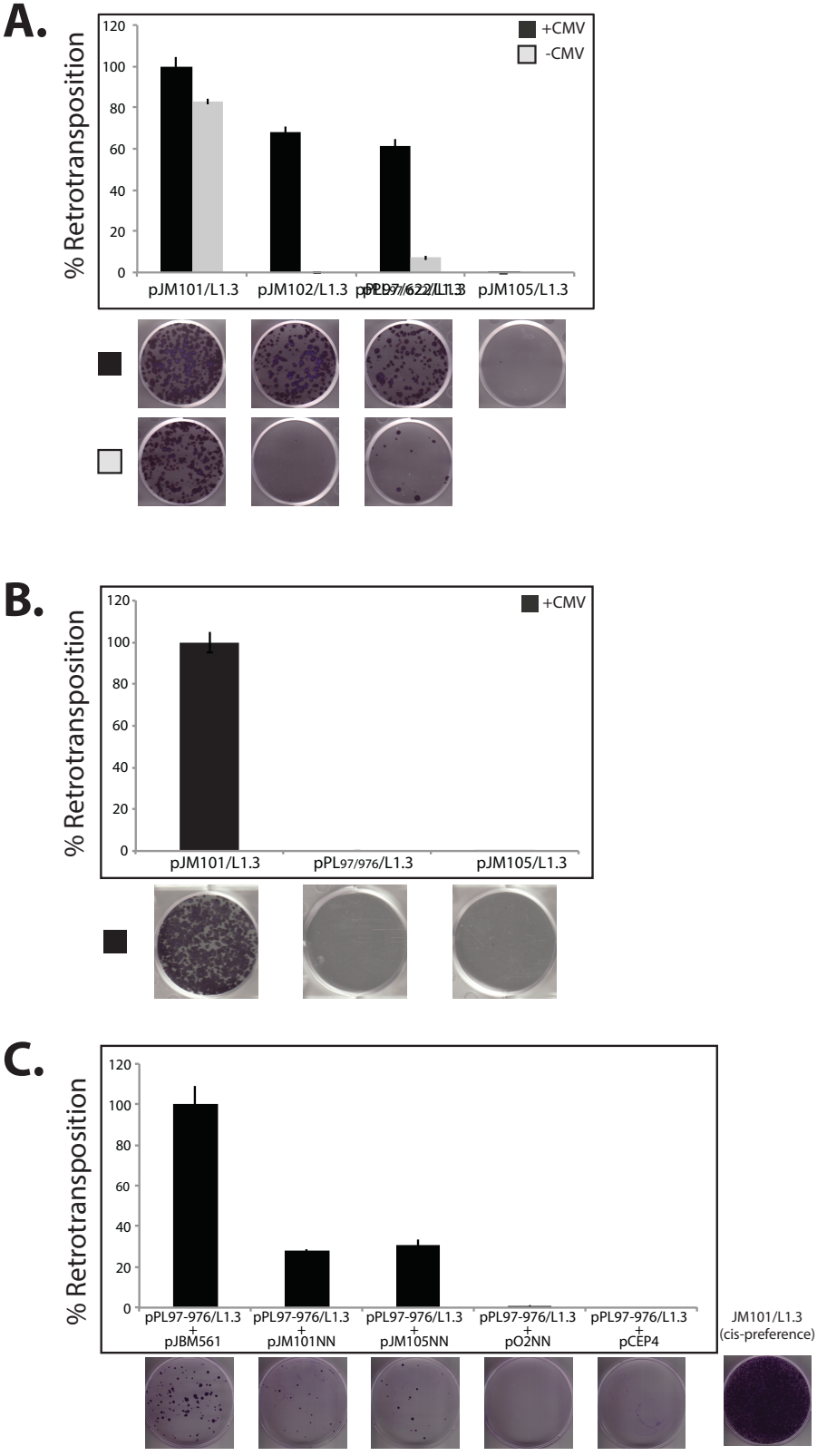


Figure 2.4: Intra-5'UTR and 5'UTR/ORF1 SpIREs are retrotransposition-defective.

Figure 2.5: A working model for how SpIREs are generated. *A) Canonical L1 Retrotransposition:* An L1 is transcribed from a genomic location (red chromosome). Translation of the mRNA (multi-colored wavy line) occurs in the cytoplasm and ORF1p (yellow circles) and ORF2p (blue oval) bind back onto their respective mRNA (*cis*-preference) to form an RNP. The L1 RNP then enters the nucleus and a *de novo* L1 insertion occurs at a new genomic location (green chromosome) by TPRT. This insertion, if full-length, could act as a source element, giving rise to new insertions (green arrow) at a new genomic location (grey chromosome). *B) Retrotransposition of intra-5'UTR spliced L1 isoform:* A full-length L1 element is transcribed from its genomic location (red chromosome) and undergoes intra-5'UTR splicing. Translation of the mRNA (multi-colored wavy line) occurs in the cytoplasm and ORF1p (yellow circles) and ORF2p (blue oval) bind back onto their respective mRNA (*cis*-preference) to form an RNP. The L1 RNP then enters the nucleus and L1 RNAs subject to intra-5'UTR splicing can undergo a single round of retrotransposition (green chromosome) by TPRT. However, since the intra-5'UTR splicing event deletes sequences required for L1 promoter activity, the resultant insertion is unlikely to undergo subsequent rounds of retrotransposition in future generations (dashed green arrow). *C.) Retrotransposition of 5'UTR/ORF1 spliced L1 isoform:* An L1 is transcribed from its genomic location (red chromosome) and is subject to 5'UTR/ORF1 splicing. Translation of the mRNA (multi-colored wavy line) occurs in the cytoplasm; however, since translation occurs at downstream AUG codons, ORF1p (yellow circles) is non-functional, the 5'UTR/ORF1 spliced L1 mRNA relies on a wild type source of ORF1p to be supplied from another L1 in *trans*. In the rare instance that *trans*-complementation occurs (dotted arrow), it is highly unlikely that the resultant SpIRE will generate RNAs that can undergo retrotransposition in future generations (dashed thin green arrow).

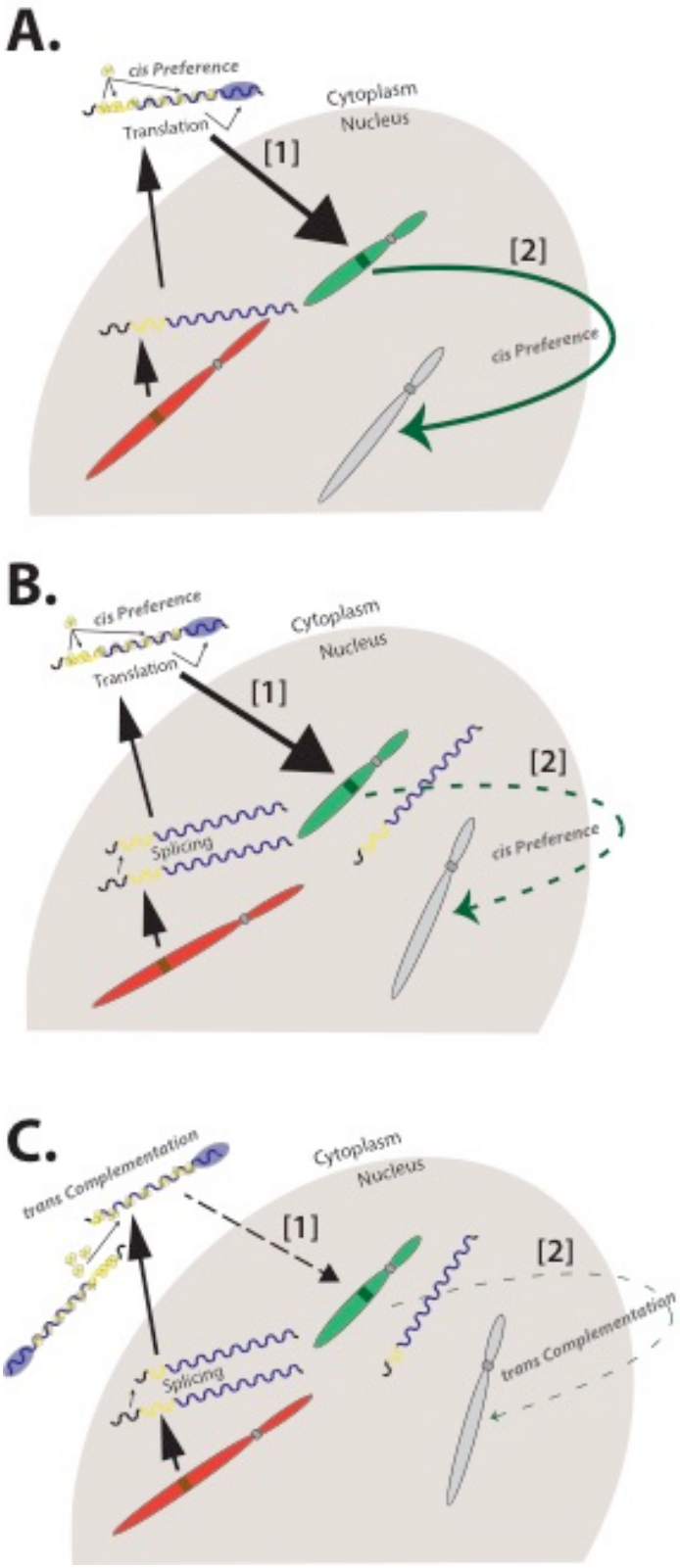


Figure 2.5: A working model for how SplREs are generated.

Figure 2.6: SpIREs in the human genome reference sequence (Supporting Figure 2.1). *A) SpIREs are present in the HGR at varying copy numbers:* The class of SpIRE ($SpIRE_{97/622}$, $SpIRE_{97/790}$ and $SpIRE_{97/976}$) is indicated at the top of each table. Column 1 indicates the L1 subfamily. Column 2 indicates the number of SpIREs present in each subfamily. Column 3 indicates the number of full-length L1s in each subfamily. Column 4 indicates the percentage (%) of SpIREs in each subfamily compared to full-length L1s. *B) Evolutionary conservation of L1 splice sites:* The panels shows the conservation of the splice donor site in the L1 5' UTR (column 1, SD, red box), as well as splice acceptor sites in the L1 5'UTR (columns 2 and 3, SA, green box) and ORF1 (column 4, SA, green box). Consensus sequences and alignment of those sequences that span the L1PA1(L1Hs)-L1PA16 subfamilies were downloaded (Khan et al. 2006) and manually inspected to determine conservation of splicing sequences (See methods).

A.

SpIRE97/622

Family	# Spliced	Full Length	Spliced/FL (%)
PA1(L1Hs)	6	296	2.0
PA2	10	1088	0.9
PA3	53	1548	3.4
PA4	22	1452	1.6
PA5	17	1136	1.5
PA6	1	1089	0.1
PA2-PA3	5	*	*
PA4-PA6	2	*	*

SpIRE97/790

Family	# Spliced	Full Length	Spliced/FL (%)
PA1(L1Hs)	1	296	0.33
PA2	3	1088	0.27

SpIRE97/976

Family	# Spliced	Full Length	Spliced/FL (%)
PA1(L1Hs)	0	296	0
PA2	2	1088	0.18
PA3	4	1548	0.26
PA4	5	1452	0.34

B.

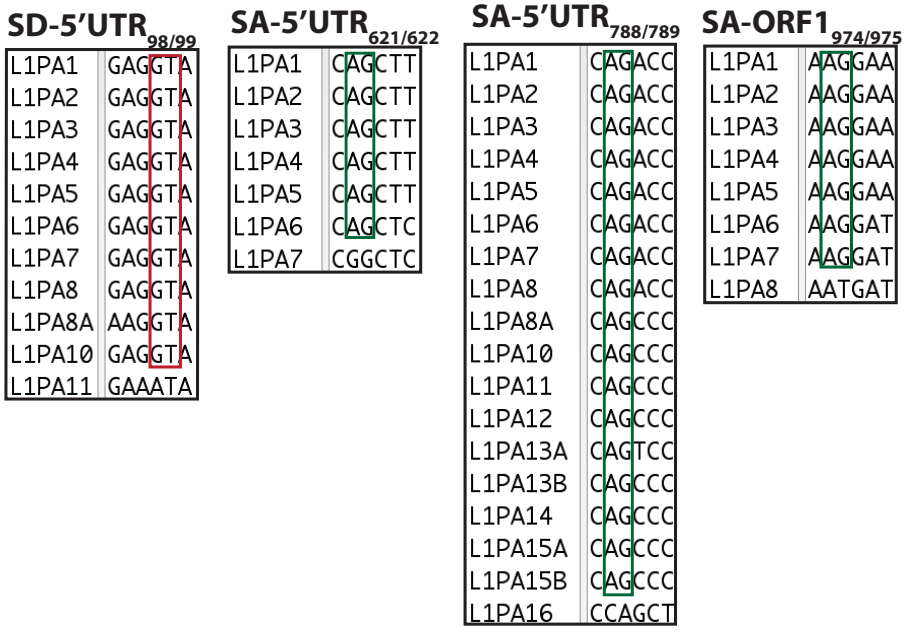


Figure 2.6: SpIREs in the human genome reference sequence (Supporting Figure 2.1).

Figure 2.7: Intra-5'UTR splicing drastically reduces L1 promoter activity (Supporting Figure 2.2). *A longer exposure of the northern blots depicted in Figure 2B.* Molecular weight standards (kb) are indicated to the left of the autoradiograph panels. The predicted sizes of the full-length 2.7 kb L1/Luciferase RNA from $pPL_{WT}LUC$ and $pPL_{SAm}LUC$ (black arrowhead) and 2.2 kb L1/Luciferase RNA from $pPL_{97\text{-}622}LUC$ (red arrowhead) are indicated at the right side of the autoradiographs. Construct names are indicated above the gel; UTF=untransfected cells. The probe used in the northern blot experiment is indicated below the autoradiograph.

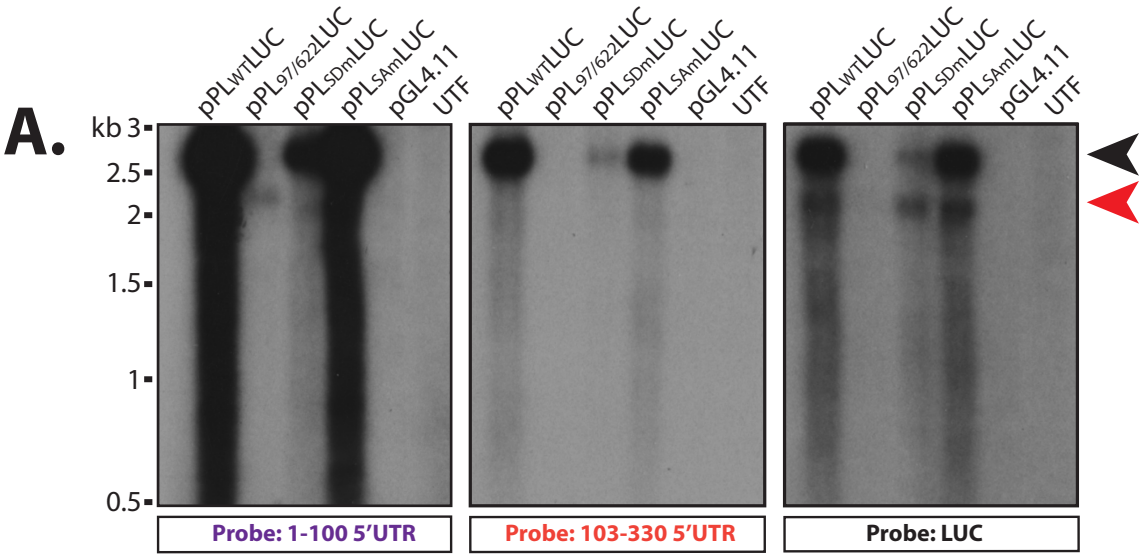


Figure 2.7: Intra-5'UTR splicing drastically reduces L1 promoter activity (Supporting Figure 2.2):

Figure 2.8: ORF1p expression from 5'UTR/ORF1 SplREs (Supporting Figure 2.3). *A) Amino terminal truncated ORF1 proteins are detected in RNPs preparations:* Molecular weight standards (Precision Plus Protein™ Kaleidoscope™(Bio-Rad)) are indicated (kDa) to the left of the gels. The predicted sizes of full-length ORF1p (black arrowhead), and the N-terminal truncated ORF1p variants (orange and blue arrows) are highlighted in the gel. Construct names are indicated above the gel; pCEP/GFP=negative control. The antibodies used in the western blot experiments are indicated to the left (α-N-ORF1p) and right (α-C-ORF1p) of the gel images. The eIF3 (110 kDa) western blots served as loading controls. Western blots were performed three times yielding similar results. B) *Amino terminal truncated ORF1 proteins are detected in WCL and RNPs preparations using an antibody to a carboxyl-terminal T7gene10 epitope tag:* Molecular weight standards (Precision Plus Protein™ Kaleidoscope™ (Bio-Rad)) are indicated by the "M" lanes (kDa). The predicted sizes of full-length ORF1p (black arrowhead), and the N-terminal truncated ORF1p variants (orange and blue arrows) detected by the anti-T7 gene10 antibody (α-T7) in WCLs (left) and RNP preparations (right) are highlighted on the gel. Construct names are indicated above the gel. The eIF3 (110 kDa) western blots served as loading controls. The untagged pJM101/L1.3 and pCEP/GFP samples served as negative controls. The bands at ~30 and ~45 kDa indicates a cross-reacting protein. Western blots were performed three times yielding similar results.

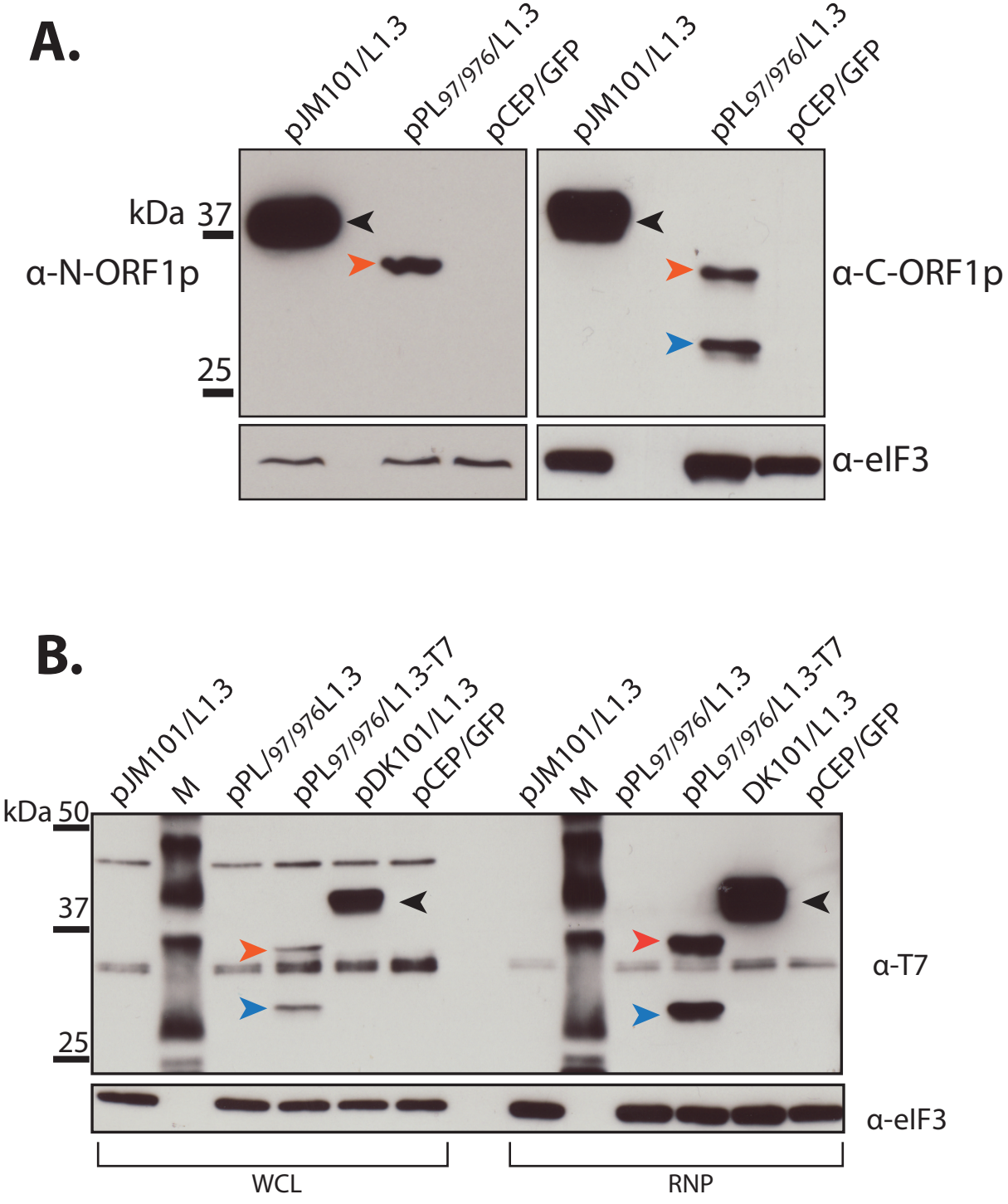
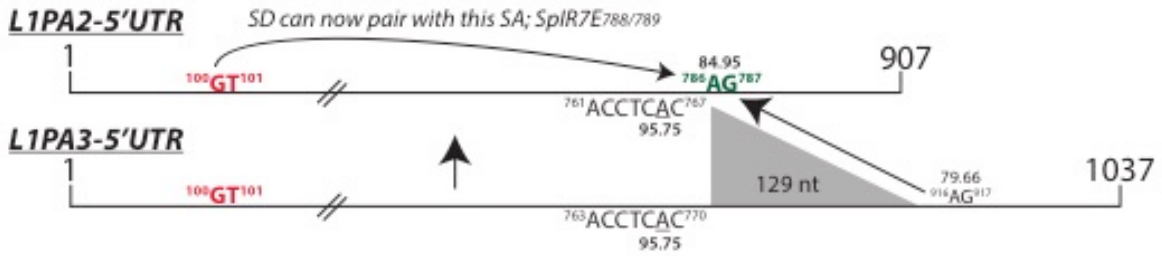


Figure 2.8: ORF1p expression from 5'UTR/ORF1 SpIREs (Supporting Figure 2.3).

Figure 2.9: Sequence changes within the L1 5'UTR may alter L1 RNA splicing profiles. *A) Schematic of the L1PA2 and L1PA3 5'UTRs:* Top, the relative positions of the splice donor (SD, red lettering), splice acceptor (SA, green lettering), and putative branch point sequence (ACCTCAC, black lettering) in the PA2 5'UTR that led to the formation of the $SpIRE_{97/790}$ are indicated in the schematic. Superscript numbers indicate the first and last nucleotide of the indicated sequence. Note nucleotide positions in the L1PA2 and L1PA3 differ slightly from those in L1PA1. Superscript numbers indicate the position of the splice sites in that subfamily. Numbers below the branch point (underlined A) (95.75) and SA (84.95) indicate the predicted strength of those sequences for utilization in a splicing reaction as determined using Human Splicing Finder v.3.0 (http://www.umd.be/HSF3/) (Desmet et al. 2009). Bottom, the relative positions of the splice donor (SD, red lettering), splice acceptor (SA, black lettering), and putative branch point sequence (ACCTCAC, black lettering) in the L1PA3 5'UTR are indicated in the schematic. Superscript numbers indicate the first and last nucleotide of the indicated sequence. Numbers above the branch point (95.75) and SA (79.66) indicate the predicted strength of those sequences for utilization in a splicing reaction as determined using Human Splicing Finder v.3.0 (http://www.umd.be/HSF3/) (Desmet et al. 2009). Notably, the PA3 5'UTR contains a 129 bp insertion that was lost in the transition from L1PA3 to L1PA2 subfamilies (gray triangle). The deletion results in moving the SA closer to the branch point in the PA2 5'UTR, leading to a higher predicted strength score (84.95 in PA2 compared to 79.66 in PA3) as determined using the Human Splicing Finder v.3.0 program (see above). *B) Table of SD, SA, and branch point sequence scores.* Consensus value scores were determined using Human Splicing Finder v.3.0 (http://www.umd.be/HSF3/)(Desmet et al. 2009). The score of the SD does not change from L1PA3 to L1PA1 as its sequence is identical and its position changes nominally. In contrast, the consensus value score of the SA increases from L1PA3 (79.66) to L1PA2 (84.95) (see methods). The table also shows the score of the strongest potential branch within 50 bp upstream of the SA (see methods).

A.



B.

Family	SD	SA	Top BP
PA1(Hs)	75.49	84.95	95.75
PA2	75.49	84.95	95.75
PA3	75.49	79.66	85.05

Figure 2.9: Sequence changes within the L1 5'UTR may alter L1 RNA splicing profiles.

Table 2.1: Additional information for each SpIRE. Column 1 indicates the clone number. Column 2 indicates the L1 subfamily. Column 3 indicates the chromosomal location. Column 4 indicates the length (bp). Column 5 indicates whether the insertion resides within a gene and the name of that gene. Column 6 indicates the transcriptional orientation of the insertion within the gene (same orientation="same", opposite orientation="Opp.". Column 7 indicates the starting location of the insertion in the HGR. Column 8 indicates the calculated L1 EN cleavage sequence of the insertion. Column 9 indicates the calculated size of the target site duplication. Column 10 indicates whether the insertion contains an L1-mediated sequence transduction. Column 11 indicates other structural features associated with the insertion. Column 12 indicates additional major deletions within the SpIRE. Column 13 described the type of deletion.

Name - Clone number	L1 Subfamily	Chromosome	Length	Gene	Orientation	Location Start	Consensus Site	TSD Length	Transduction	Extra	Additional deletions	Type of deletion
SplRE(97/622)-9	PA1(L1Hs)	1	3398			104464623	TATT/G	7	Yes	3' 391 untemp	3' truncation (3925-end)	In transposon
SplRE(97/622)-38	PA1(L1Hs)	4	5653			167187141	AGTA/A	13	No	5' 3 mismatch		
SplRE(97/622)-52	PA1(L1Hs)	6	2398			27911128	TTTA/A	16	No	5' 2 mismatch, 1 missing	del (1390-3692, 5225-end)	Inversion/deletion
SplRE(97/622)-62	PA1(L1Hs)	7	3822			62521121	NONE	NONE	No		3' truncation (4349-end)	3' truncation
SplRE(97/622)-90	PA1(L1Hs)	12	5484			10502541	CTTT/C	18	No	5' 1 mismatch		
SplRE(97/622)-104	PA1(L1Hs)	15	5507	UACA	Opp.	71022057	TTTT/C	12	No	5' 2 mismatch		
SplRE(97/622)-6	PA2	1	5505			206506662	CATT/C	16	No			
SplRE(97/622)-28	PA2	3	5344	SYN2	Opp.	12074791	AATT/C	14	No	5' one untemp		
SplRE(97/622)-35	PA2	4	5500	RAP1GDS	Same	99241583	GAAT/C	11	No			
SplRE(97/622)-64	PA2	7	5500	DGKB	Opp.	14744105	TTTT/G	32	No	5' one mismatch		
SplRE(97/622)-70	PA2	8	5500	C8orf34	Same	69274647	TTTC/C	13	No	5' 20 untemp		
SplRE(97/622)-72	PA2	8	5141			129724828	CTTT/G	7	Yes	5' 130 untemp	del(4381-4751)	Internal deletion
SplRE(97/622)-75	PA2	10	5486			7105433	CTTT/G	16	No	5' 2 untemp		
SplRE(97/622)-78	PA2	10	5505			7179414	TTTT/A	17	No	5' 23 untemp		
SplRE(97/622)-79	PA2	10	5501			44983702	ATTT/A	15	No			
SplRE(97/622)-105	PA2	15	5509			51465698	TATT/A	7	No			
SplRE(97/622)-1	PA3	X	5507			100795847	TCCT/A	19	No	5' one untemp		
SplRE(97/622)-2	PA3	X	5514			68683795	ATTT/C	15	No	5' one untemp		
SplRE(97/622)-4	PA3	X	5500			112689223	TTTA/A	16	No			
SplRE(97/622)-8	PA3	1	5613	C1orf146	Opp.	92703286	TATA/T	11	No			
SplRE(97/622)-10	PA3	1	5644			101148126	TTTT/G	14	No	5' one mismatch		
SplRE(97/622)-11	PA3	1	5494	RHOU	Same	228843077	TTTA/T	16	No	5' 12 untemp		
SplRE(97/622)-12	PA3	1	5497	BC054887	Same	71681513	TCTT/A	14	No			
SplRE(97/622)-13	PA3	1	5676			37673543	TCTT/A	13	Yes	3' 705 untemp		
SplRE(97/622)-14	PA3	1	5496			81411743	CTTT/A	16	No	5' 4 untemp		
SplRE(97/622)-18	PA3	2	5499			193902734	TTCA/G	17	No			
SplRE(97/622)-19	PA3	2	5482			162392723	TTTT/A	17	No	5' one mismatch		
SplRE(97/622)-20	PA3	2	5655			132131384	TTTA/A	13	No	5' A untemp		
SplRE(97/622)-23	PA3	2	5628	LRRTM4	Opp	77132008	TTTC/A	17	No			
SplRE(97/622)-25	PA3	2	5503			97106840	TCTT/A	15	No	5 one' mismatch, 3 untemp		
SplRE(97/622)-27	PA3	3	5514	IQCJ/SCH	Opp.	158808219	TTCT/G	15	No	5' 4 untemp		
SplRE(97/622)-30	PA3	3	5636			135541350	TTTC/T	14	No	5' 8 untemp		
SplRE(97/622)-34	PA3	3	5501	THRB	Opp.	24321589	GCTT/A	13	No	5' one mismatch		
SplRE(97/622)-36	PA3	4	5498	CXCL13	Opp.	78440573	CTTT/G	18	No	5' 2 mismatch		
SplRE(97/622)-39	PA3	4	5498			116093532	CTTT/C	11	No	5' 6 untemp, one mismatch		
SplRE(97/622)-42	PA3	4	5628			128781215	CTTT/C	7	No	5' one mismatch		
SplRE(97/622)-43	PA3	5	5511			13521914	ATTT/C	17	No	5' 2 mismatch, 3 untemp, 1		
SplRE(97/622)-44	PA3	5	5491	ANKHD1	Same	139895037	TTTT/C	9	No			
SplRE(97/622)-45	PA3	5	5527			8159176	TTAA/C	14	Yes	5' 3 mismatch, 3' 1036 untemp		
SplRE(97/622)-46	PA3	5	5504	HTR4	Opp.	147989751	CTTT/C	15	No			

Table 2.1: Additional information for each SplRE.

Name - Clone number	L1 Subfamily	Chromosome	Length	Gene	Orientation	Location Start	Consensus Site	TSD Length	Transduction	Extra	Additional deletions	Type of deletion
SplRE(97/622)-47	PA3	5	5444			108962938	TTTA/G	15	No	5' 2 mismatch, 2 untemp		
SplRE(97/622)-48	PA3	5	5463	U28131	Same	104053455	TCTT/A	12	No			
SplRE(97/622)-53	PA3	6	5483			77455377	TAAA/G	9	No	5' 1 mismatch		
SplRE(97/622)-54	PA3	6	5500			115330140	TTCT/A	18	No	5' 2 mismatch, 8 untemp		
SplRE(97/622)-56	PA3	6	5431			83342492	ATTT/A	14	No	5' 2 mismatch, 1 untemp		
SplRE(97/622)-57	PA3	6	5382			81530481	TTTT/A	13	No	5' 3 untemp		
SplRE(97/622)-58	PA3	6	5512	AF086303	Same	75194248	AATT/A	17	No	5' 2 missing		
SplRE(97/622)-59	PA3	6	5638	SLC35F1	Same	118317486	TTTA/G	16	No	5' 2 mismatch, 3' 17 untemp		
SplRE(97/622)-61	PA3	7	5503			86867995	CTCT/C	7	No	5' 1 mismatch, 6 untemp		
SplRE(97/622)-63	PA3	7	5479			89531713	TCTT/A	14	No			
SplRE(97/622)-66	PA3	7	5628			83378686	NONE	NONE	No			
SplRE(97/622)-71	PA3	8	5494	CSMD1	Same	3133943	TCTT/A	14	No	5' 4 untemp		
SplRE(97/622)-73	PA3	8	5525			130336552	TTTT/A	15	No			
SplRE(97/622)-76	PA3	10	5496			55557605	TTTC/G	10	No			
SplRE(97/622)-77	PA3	10	5528			11365205	TCAT/C	18	No	5' 2 mismatch		
SplRE(97/622)-82	PA3	10	5502			50454181	TTTT/A	15	No	5' 1 untemp		
SplRE(97/622)-83	PA3	11	5532			13820879	TTTT/A	11	No			
SplRE(97/622)-84	PA3	11	5499			100523754	CTTT/C	10	No	5' 1 mismatch, 1 untemp		
SplRE(97/622)-85	PA3	11	5503			29437673	TTTT/A	10	No			
SplRE(97/622)-89	PA3	12	5487			61160740	TTTT/A	9	No	5' 1 mismatch, 2 untemp		
SplRE(97/622)-98	PA3	13	5621			47952791	TTTT/A	17	No	5' 3 mismatch, 10 untemp		
SplRE(97/622)-100	PA3	14	5610			66327000	ACCC/C	8	Yes	5' one mismatch, 3' 199 untemp		
SplRE(97/622)-106	PA3	15	5618	BC04142	Opp.	72090979	TGTT/A	14	No	5' 1 mismatch, 2 untemp		
SplRE(97/622)-107	PA3	15	5600	Pcak1	Opp.	77553007	TTTT/C	7	No			
SplRE(97/622)-108	PA3	16	5581			63181610	TTTT/A	7	Yes	5' 560 untemp		
SplRE(97/622)-111	PA3	18	5484			47302933	TATT/C	7	No	5' 9 untemp		
SplRE(97/622)-114	PA3	20	5622	TMEM90B	Opp.	2453871	CTTT/A	13	No			
SplRE(97/622)-115	PA3	20	5573			12643330	TTTT/C	15	No	5' 1 untemp		
SplRE(97/622)-116	PA3	21	5489			18535862	AAAT/T	13	No	5' 7 untemp		
SplRE(97/622)-17	PA4	1	5624			197798790	TTTT/G	16	No	5' 2 mismatch		
SplRE(97/622)-21	PA4	2	4100	FSHR	Opp.	49309153	TTTA/A	16	No	5' 1 mismatch, 1 missing	del (4025-5465)	Internal deletion
SplRE(97/622)-29	PA4	3	5588	UBE2E2	Opp.	23343396	TCTT/C	12	No	5' 1 mismatch, 4 untemp		
SplRE(97/622)-32	PA4	3	5626			43986936	TTTT/A	10	No	5' 1 mismatch, 5 untemp		
SplRE(97/622)-33	PA4	3	5619			24126062	TCTT/A	17	No	5' 1 mismatch		
SplRE(97/622)-37	PA4	4	5576			53371143	CCAA/A	9	Yes	5' 1 mismatch, 3' 61 untemp		
SplRE(97/622)-40	PA4	4	5629			64819421	NONE	NONE	No			
SplRE(97/622)-49	PA4	5	5627	RNU5E	Same	80572990	TTTT/C	13	No	5' 3 mismatch		
SplRE(97/622)-60	PA4	6	5639	SMAP1	Same	71492534	TTCC/A	8	No	5' 8 untemp, 3' 12 untemp		
SplRE(97/622)-67	PA4	7	5612			140414028	TTTT/G	10	No	5' 1 mismatch, 16 untemp		
SplRE(97/622)-68	PA4	7	5618	RELN	Opp.	103382253	AATA/T	6	No	5' 3 untep, 3' 3 untemp		

Table 2.1: Additional information for each SplRE.

Name - Clone number	L1 Subfamily	Chromosome	Length	Gene	Orientation	Location Start	Consensus Site	TSD Length	Transduction	Extra	Additional deletions	Type of deletion
SplRE(97/622)-74	PA4	9	5585			129755879	TGTA/A	15	Yes	5' 1 missing, 1 mismatch, 20 untemp, 3' 250 untemp		
SplRE(97/622)-80	PA4	10	5614			36302656	TCTT/A	15	No	5' 2 mismatch, 5 untemp		
SplRE(97/622)-81	PA4	10	5612	ITGA8	Same	15712129	TCTT/A	11	No	5' 6 untemp		
SplRE(97/622)-86	PA4	11	5621			48588752	TATT/G	11	No	5' 1 untemp		
SplRE(97/622)-88	PA4	11	5628			87662913	CTGG/C	6	Yes	3' 200 untemp		
SplRE(97/622)-93	PA4	12	5625			23674212	TTTT/A	16	No			
SplRE(97/622)-94	PA4	12	5644	RAb3IP	Opp	70204301	TTTT/G	14	No	5' 13 untemp		
SplRE(97/622)-95	PA4	12	5631			11467678	TGTA/A	8	No	5' 4 untemp		
SplRE(97/622)-99	PA4	13	5590			82039335	TTTT/A	16	No	5' 8 untemp		
SplRE(97/622)-102	PA4	14	5577			44945977	TTTT/C	16	No	5' 3 mismatch		
SplRE(97/622)-112	PA4	18	5641			44508670	TTTT/C	14	No	5' 3 untemp		
SplRE(97/622)-3	PA5	X	5668	SPANX	Opp.	140536398	GTTT/C	11	Yes	5' 3 untemp, 2 mismatch, 3' 150 untemp		
SplRE(97/622)-5	PA5	X	5624	C1orf146	Opp.	151493710	TTTC/A	14	No			
SplRE(97/622)-15	PA5	1	1641	HHAT	Opp.	210789405	TTTT/A	10	No	5' 5 untemp	3' deletion (1294-end)	In transposon
SplRE(97/622)-16	PA5	1	5622			114730873	TTTG/T	18	Yes	5' 1 untemp, 2 mismatch, 3' 112 untemp		
SplRE(97/622)-22	PA5	2	5630			48285651	CTTT/G	23	No	5' 1 untmp, 7 mismatch		
SplRE(97/622)-24	PA5	2	5602	LRRTM4	Opp.	77363150	CTTT/A	12	No	5' 8 untemp, 3 mismatch		
SplRE(97/622)-31	PA5	3	5607			57592523	CTTT/A	13	No	5' 13 untemp		
SplRE(97/622)-41	PA5	4	5620	Mir5684	Same	165570450	GATA/A	15	Yes	5' 1 untemp, 3 mismatch, 3' 214 untemp		
SplRE(97/622)-51	PA5	6	5488			30215660	ATTT/G	30	No	5' 2 untemp, 9 mismatch, 4 missing		
SplRE(97/622)-69	PA5	7	5621			152819737	TTTC/T	12	No	5' 1 untemp		
SplRE(97/622)-92	PA5	12	5633			55939783	TTTT/A	15	No	5' 5 untemp		
SplRE(97/622)-96	PA5	12	5475			4450452	TTGT/G	20	No	5' 7 mismatch, 17 untemp		
SplRE(97/622)-97	PA5	12	5618			108116424	TTTT/A	40	No	5' 4 mismatch, 7 missing		
SplRE(97/622)-101	PA5	14	5638			50539565	TTTC/A	14	No	5' 22 untemp		
SplRE(97/622)-103	PA5	14	5577	KIAA0391	Opp.	35649534	TCTT/A	14	No	5' 4 untemp, 1 mismatch		
SplRE(97/622)-109	PA5	17	5639	ACCN1	Same	31391824	CTTT/G	17	No	5' 12 untemp, 1 mismatch		
SplRE(97/622)-110	PA5	17	5589			15258532	GGTT/A	7	Yes	5' 11 untemp, 3' 239 untemp		
SplRE(97/622)-50	PA6	5	5607			125422191	TTTT/G	15	No	5' 2 untemp, 2 mismatch	3' truncation (3309-end)	3' truncation
SplRE(97/622)-55	PA2-3	6	2765	LACE1	Opp.	108712189	NONE	NONE	No			
SplRE(97/622)-7	PA2-3	7	5509	SRGAP2P	Opp.	148086250	NONE	NONE	No		del (4693-end)	3' truncation
SplRE(97/622)-91	PA2-3	12	4125			77658384	NONE	NONE	No		3'truncation (3634-end)	3' truncation
SplRE(97/622)-87	PA4-6	11	3236	ANO3	Opp.	26314602	NONE	NONE	No		3' deletion (2842-end)	In transposon
SplRE(97/622)-113	PA4-6	19	2282			35348802	NONE	NONE	No		del (1111-5730)	internal deletion
SplRE(97/622)-65	PA2-3	7	1015	CNTNAP2	Opp.	148086130	CTTT/A	18	No		3' truncation (783-end)	In transposon
SplRE(97/622)-26	PA2-3	2	240			21071818	NONE	NONE	No			

Table 2.1: Additional information for each SplRE.

Name - Clone Number	L1 Subfamily	Chromosome	Length	Gene	Orientation	Location Start	Concensus Site	TSD Length	Transduction	Extra	Additional deletions	Type of deletion
SpIRE(97-790)-1	PA2	1	5342			40831200	TTCT/A	18	NO	5' 8 untemp		
SpIRE(97-790)-2	PA2	9	5335	MIR31HG	Opp.	21536613	CCTT/G	12	NO			
SpIRE(97-790)-3	PA1	3	5331	COL6A6	Same	130347579	TCTT/A	11	NO	5' 1 untemp		
SpIRE(97-790)-4	PA2	11	5334	RNF169	Opp.	74473673	TTTA/A	14	NO	5' 1 mismatch		

Table 2.1: Additional information for each SpIRE.

Name-Clone Number	L1 Subfamily	Chromosome	Length	Gene	Orientation	Location Start	Consensus Site	TSD	Transduction	extra	Additional deletions	Type of deletion
SpIRE(97/976)-1	PA2	14	5157			56460057	GCCT/G	27	No	5' 1 untemp		
SpIRE(97/976)-2	PA2	3	5146	CADPS	Same	62838102	CTTT/G	14	No	5' 1 untemp		
SpIRE(97/976)-3	PA3	12	5181			28161108	AAAT/A	18	No	5' 2 untemp, 2 mismatch		
SpIRE(97/976)-4	PA3	1	5136			157315230	TTTT/A	32	No	5' 8 mismatch		
SpIRE(97/976)-5	PA3	6	5123			85079405	TTTT/G	26	No	5' 3 mismatch		
SpIRE(97/976)-6	PA3	6	5161	FIG4	Opp.	110078291	TTTT/A	18	No	5' 1 mismatch		
SpIRE(97/976)-7	PA4-PA6	4	4058			150049502	TTTC/A	8	No	5' 1 mismatch		
SpIRE(97/976)-8	PA4	X	5232	KLHL13	Opp.	117227851	TTTT/G	16	No	5' 1 mismatch, 5 untemp		
SpIRE(97/976)-9	PA4	12	5101			22499733	TTAT/C	19	No	5' 1 missing, 3 untemp		
SpIRE(97/976)-10	PA4	2	5064	UNC13C	Same	162385954	TCTT/A	11	No			
SpIRE(97/976)-11	PA4	15	3229			54347959	TCTT/G	14	No			

Table 2.1: Additional information for each SpIRE.

References

Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. 2006. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes & development* 20(2): 210-224.

Athanikar JN, Badge RM, Moran JV. 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic acids research* 32(13): 3846-3855.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* 141(7): 1159-1170.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annual review of genomics and human genetics* 12: 187-215.

Becker KG, Swergold GD, Ozato K, Thayer RE. 1993. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Human molecular genetics* 2(10): 1697-1702.

Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic acids research* 34(5): 1512-1521.

Belancio VP, Roy-Engel AM, Deininger P. 2008. The impact of multiple splice sites in human L1 elements. *Gene* 411(1-2): 38-45.

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P. 2010. Somatic expression of LINE-1 elements in human tissues. *Nucleic acids research* 38(12): 3909-3922.

Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. 2006. Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences of the United States of America* 103(25): 9590-9594.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* 100(9): 5280-5285.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. *Molecular and cellular biology* 24(24): 10505-10514.

Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, Sverdlov E. 2002. A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. *Genomics* 80(4): 402-406.

Chen J, Rattner A, Nathans J. 2006a. Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements. *Human molecular genetics* 15(13): 2146-2156.

Chen JM, Ferec C, Cooper DN. 2006b. LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. *J Biomed Biotechnol* 2006(1): 56182.

Christensen SM, Bibillo A, Eickbush TH. 2005. Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic acids research* 33(20): 6461-6468.

Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *The EMBO journal* 21(21): 5899-5910.

Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW. 1981. Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *Journal of molecular biology* 151(1): 17-33.

Denli AM, Narvaiza I, Kerman BE, Pena M, Benner C, Marchetto MC, Diedrich JK, Aslanian A, Ma J, Moresco JJ et al. 2015. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* 163(3): 583-593.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* 37(9): e67.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35(1): 41-48.

Dmitriev SE, Andreev DE, Terenin IM, Olovnikov IA, Prassolov VS, Merrick WC, Shatsky IN. 2007. Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. *Molecular and cellular biology* 27(13): 4685-4697.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH, Jr. 1991. Isolation of an active human transposable element. *Science* 254(5039): 1805-1808.

Dombroski BA, Scott AF, Kazazian HH, Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proceedings of the National Academy of Sciences of the United States of America* 90(14): 6513-6517.

Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, Kopera HC, Athanikar JN, Hasnaoui M, Bucheton A, Moran JV et al. 2010. Characterization of LINE-1 ribonucleoprotein particles. *PLoS genetics* 6(10).

Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. 2015. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Molecular cell* 60(5): 728-741.

Ergun S, Buschmann C, Heukeshoven J, Dammann K, Schnieders F, Lauke H, Chalajour F, Kilic N, Stratling WH, Schumann GG. 2004. Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. *The Journal of biological chemistry* 279(26): 27753-27763.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 24(4): 363-367.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151(3): 483-496.

Fanning T, Singer M. 1987. The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic acids research* 15(5): 2251-2260.

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87(5): 905-916.

Freeman JD, Goodchild NL, Mager DL. 1994. A modified indicator gene for selection of retrotransposition events in mammalian cells. *BioTechniques* 17(1): 46, 48-49, 52.

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic acids research* 36(7): 2257-2267.

Garcia-Perez JL, Doucet AJ, Bucheton A, Moran JV, Gilbert N. 2007. Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome research* 17(5): 602-611.

Gilbert N, Lutz S, Morrish TA, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Molecular and cellular biology* 25(17): 7780-7795.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* 110(3): 315-325.

Goodier JL, Mandal PK, Zhang L, Kazazian HH, Jr. 2010. Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. *Human molecular genetics* 19(9): 1712-1725.

Goodier JL, Ostertag EM, Kazazian HH, Jr. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human molecular genetics* 9(4): 653-657.

Grimaldi G, Skowronski J, Singer MF. 1984. Defining the beginning and end of KpnI family segments. *The EMBO journal* 3(8): 1753-1759.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH, Jr. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human molecular genetics* 20(17): 3386-3400.

Hancks DC, Kazazian HH, Jr. 2016. Roles for retrotransposon insertions in human disease. *Mobile DNA* 7: 9.

Hastings ML, Krainer AR. 2001. Pre-mRNA splicing in the new millennium. *Current opinion in cell biology* 13(3): 302-309.

Hattori M, Kuhara S, Takenaka O, Sakaki Y. 1986. L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature* 321(6070): 625-628.

Hohjoh H, Minakami R, Sakaki Y. 1990. Selective cloning and sequence analysis of the human L1 (LINE-1) sequences which transposed in the relatively recent past. *Nucleic acids research* 18(14): 4099-4104.

Hohjoh H, Singer MF. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *The EMBO journal* 15(3): 630-639.

Hohjoh H, Singer MF. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *The EMBO journal* 16(19): 6034-6043.

Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH, Jr. 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat Genet* 7(2): 143-148.

Holmes SE, Singer MF, Swergold GD. 1992. Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *The Journal of biological chemistry* 267(28): 19765-19768.

Inoue H, Nojima H, Okayama H. 1990. High efficiency transformation of Escherichia coli with plasmids. *Gene* 96(1): 23-28.

Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*.

Januszyk K, Li PW, Villareal V, Branciforte D, Wu H, Xie Y, Feigon J, Loo JA, Martin SL, Clubb RT. 2007. Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. *The Journal of biological chemistry* 282(34): 24893-24904.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 110(1-4): 462-467.

Kazazian HH, Jr., Moran JV. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* 19(1): 19-24.

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 332(6160): 164-166.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* 12(4): 656-664.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome research* 16(1): 78-87.

Khazina E, Truffault V, Buttner R, Schmidt S, Coles M, Weichenrieder O. 2011. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nature structural & molecular biology* 18(9): 1006-1014.

Khazina E, Weichenrieder O. 2009. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proceedings of the National Academy of Sciences of the United States of America* 106(3): 731-736.

Kopera HC, Larson PA, Moldovan JB, Richardson SR, Liu Y, Moran JV. 2016. LINE-1 Cultured Cell Retrotransposition Assay. *Methods in molecular biology* 1400: 139-156.

Kozak M. 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. *Nature* 308(5956): 241-246.

Krawczak M, Reiss J, Cooper DN. 1992. The Mutational Spectrum of Single Base-Pair Substitutions in Messenger-Rna Splice Junctions of Human Genes - Causes and Consequences. *Human genetics* 90(1-2): 41-54.

Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J. 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC evolutionary biology* 7: 190.

Kubo S, Seleme MC, Soifer HS, Perez JL, Moran JV, Kazazian HH, Jr., Kasahara N. 2006. L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences of the United States of America* 103(21): 8036-8041.

Kulpa DA, Moran JV. 2005. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Human molecular genetics* 14(21): 3237-3248.

Kulpa DA, Moran JV. 2006. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nature structural & molecular biology* 13(7): 655-660.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921.

Leibold DM, Swergold GD, Singer MF, Thayer RE, Dombroski BA, Fanning TG. 1990. Translation of LINE-1 DNA elements in vitro and in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 87(18): 6990-6994.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72(4): 595-605.

Macfarlane CM, Collier P, Rahbari R, Beck CR, Wagstaff JF, Igoe S, Moran JV, Badge RM. 2013. Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. *Human mutation* 34(7): 974-985.

Martin F, Maranon C, Olivares M, Alonso C, Lopez MC. 1995. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. *Journal of molecular biology* 247(1): 49-59.

Martin SL. 1991. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. *Molecular and cellular biology* 11(9): 4804-4807.

Martin SL, Branciforte D, Keller D, Bain DL. 2003. Trimeric structure for an essential protein in L1 retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America* 100(24): 13815-13820.

Martin SL, Bushman FD. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Molecular and cellular biology* 21(2): 467-475.

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* 254(5039): 1808-1810.

McMillan JP, Singer MF. 1993. Translation of the human LINE-1 element, L1Hs. *Proceedings of the National Academy of Sciences of the United States of America* 90(24): 11533-11537.

Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y. 1992. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic acids research* 20(12): 3139-3145.

Moldovan JB, Moran JV. 2015. The Zinc-Finger Antiviral Protein ZAP Inhibits LINE and Alu Retrotransposition. *PLoS genetics* 11(5): e1005121.

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* 283(5407): 1530-1534.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* 87(5): 917-927.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* 31(2): 159-165.

Naufer MN, Callahan KE, Cook PR, Perez-Gonzalez CE, Williams MC, Furano AV. 2016. L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent property conserved despite extensive remodeling. *Nucleic acids research* 44(1): 281-293.

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *American journal of human genetics* 73(6): 1444-1451.

Perepelitsa-Belancio V, Deininger P. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* 35(4): 363-366.

Pickeral OK, Makalowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome research* 10(4): 411-415.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 35(Database issue): D61-65.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Lower J, Stratling WH, Lower R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic acids research* 40(4): 1666-1683.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology spectrum* 3(2): MDNA3-0061-2014.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH, Jr. 1997. Many human L1 elements are capable of retrotransposition. *Nat Genet* 16(1): 37-43.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature methods* 9(7): 671-675.

Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, Cooley T, Heath P, Smith KD, Margolet L. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* 1(2): 113-125.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development* 9(6): 657-663.

Smit AF, Toth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of molecular biology* 246(3): 401-417.

Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* 390(1-2): 206-213.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and cellular biology* 21(6): 1973-1985.

Swergold GD. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and cellular biology* 10(12): 6718-6729.

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* 110(3): 327-338.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. *Genome biology* 3(10): research0052.

Tchenio T, Casella JF, Heidmann T. 2000. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic acids research* 28(2): 411-415.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* Chapter 2: Unit 2 3.

Usdin K, Furano AV. 1989. The structure of the guanine-rich polypurine:polypyrimidine sequence at the right end of the rat L1 (LINE) element. *The Journal of biological chemistry* 264(26): 15681-15687.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Molecular and cellular biology* 21(4): 1429-1439.

Wei W, Morrish TA, Alisch RS, Moran JV. 2000. A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. *Anal Biochem* 284(2): 435-438.

Wimmer K, Callens T, Wernstedt A, Messiaen L. 2011. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS genetics* 7(11): e1002371.

Yang N, Zhang L, Zhang Y, Kazazian HH, Jr. 2003. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic acids research* 31(16): 4929-4940.

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R. 1998. Apolipoprotein(a) gene enhancer resides within a LINE element. *The Journal of biological chemistry* 273(2): 891-897.

Chapter 3

A Genome Wide Screen to Identify Factors Inhibiting Expression of Retrotransposed LINE-1 Elements in Embryonic Carcinoma Cells

Experiments and results discussed in this chapter were performed in collaboration with Dr. Jacob Kitzman and Mr. Trenton Frisbie. Mr. Trenton Frisbie will continue this line of inquiry and investigation.

Justification for Study

Long INterspersed Element-1 (LINE-1 or L1) is an endogenous non-LTR retrotransposon that comprises ~17% of the human genome (Lander et al., 2001). L1s move in the genome via a copy-paste mechanism called retrotransposition (Boeke et al., 1985). The mechanism of L1 genomic integration is termed target-site primed reverse transcription (TPRT) (Luan et al., 1993; Feng et al., 1996; Cost et al., 2002; Kulpa and Moran 2006). Importantly, TPRT is a mechanism unique to non-LTR retrotransposition, and differs from integration mechanisms used by LTR retrotransposons, DNA transposons, and retroviruses. L1s have been demonstrated to retrotranspose in various human cell lines, as well as in certain somatic cells, germ line cells, and in early embryonic development (Evrony et al., 2010; Upton et al., 2015; Muotri et al., 2005; Garcia-Perez et al., 2007b; Ostertag et al., 2002). Previous experimentation in human embryonic carcinoma-derived cells (hECs) revealed that engineered L1s (herein referred to as an L1-reporter) successfully retrotranspose into hEC genomic DNA (Garcia-Perez et al., 2010). However, either during, or immediately after genomic integration, expression of the L1-reporter is silenced in hECs. Interestingly, hEC cells treated with histone deacetylase (HDAC) inhibitors rapidly reverse L1-reporter silencing (Garcia-Perez et al., 2010). Here, we performed proof-of-

principle experiments and designed a CRISPR/Cas9 genetic knockout screen in an attempt to identify factors that may be involved with L1-reporter silencing.

Introduction

Previous work from our lab demonstrated that a human embryonic carcinoma (hEC) cell line (PA-1) permits retrotransposition of an engineered L1 tagged with an enhanced green fluorescent protein (EGFP) retrotransposition indicator cassette (L1-EGFP). However, the resultant integrated *L1-retro-EGFP* is silenced either during or immediately after retrotransposition and, as a result, does not express EGFP (Garcia-Perez et al., 2010). A small number of cells containing *L1-retro-EGFP* events (<0.3%) express EGFP, suggesting that silencing is not absolute in PA-1 cells (Garcia-Perez et al., 2010). Notably, treatment of PA-1 cell populations harboring *L1-retro-EGFP* events with pan-histone deacetylase (HDAC) inhibitors rapidly reversed *L1-retro-EGFP* silencing and resulted in a ~20 fold increase of EGFP expressing cells when compared untreated PA-1 cells. Thus, these data suggest *L1-retro-EGFP* is silenced during or immediately after integration (Garcia-Perez et al., 2010).

Isolation of a clonal PA-1 cell line (called pk-5 cells) containing a single, stably integrated, full-length, *L1-retro-EGFP* event faithfully did not express EGFP. Addition of HDAC inhibitors to pk-5 cells also reactivated EGFP expression (Garcia-Perez et al., 2010). Subsequent removal of the HDAC inhibitors from pk-5 cells re-established silencing of *L1-retro-EGFP*. These data suggest the presence of a memory mechanism in PA-1 cells that can re-establish *L1-retro-EGFP* silencing. Thus, evidence suggests *L1-retro-EGFP* silencing is a two-step process that requires: 1) an initiation step that is mediated by host factor(s); and 2) a maintenance step that uses the same or an additional host factors(s) (Garcia-Perez et al., 2010).

Additional experiments demonstrated that an *EGFP* reporter delivered by natural or synthetic mouse L1s, or zebrafish L2 retrotransposons were readily silenced in PA-1 cells (Garcia-Perez et al., 2010). Thus, sequence composition of the retrotransposon did not affect silencing. In contrast, EGFP delivered by HIV or MMLV retroviral vectors or by stable transfection were not efficiently silenced in PA-1 cells, although the treatment of retroviral-infected cells with HDAC inhibitors resulted in a modest (2-3 fold) increase of EGFP expression (Garcia-Perez et al., 2010). These data suggest reporter

genes delivered by retrotransposon-mediated TPRT may be specifically recognized and targeted for silencing.

PA-1 cells are diploid, but contain a reciprocal translocation between chromosomes 15 and 20 (Garcia-Perez et al., 2010; Sarraf et al., 2005) and preferentially differentiate into an ectodermal-like lineage (Garcia-Perez et al., 2010). PA-1 cells transfected with L1-EGFP and grown in differentiation media exhibited a ~30 fold increase in EGFP expressing cells compared to L1-EGFP transfected PA-1 cells grown in normal media (Garcia-Perez et al., 2010). Additionally, L1-EGFP transfected PA-1 cells grown in differentiation media exhibited a notable increase in EGFP expressing cells upon treatment with HDAC inhibitors (Garcia-Perez et al., 2010). Together, these data suggest *L1-retro-EGFP* silencing is more efficient in PA-1 cells than in actively differentiating PA-1 cells; however, addition of HDAC inhibitors to differentiating cells activates expression from additional *L1-retro-EGFP* events (Garcia-Perez et al., 2010).

Interestingly, differentiation is not sufficient to reactivate previously silenced *L1-retro-EGFP* events (Garcia-Perez et al., 2010). Culture of the clonal pk-5 cell line in differentiation media resulted in only a minor reactivation of EGFP expressing cells (Garcia-Perez et al., 2010). This phenomenon was recapitulated in populations of *L1-retro-EGFP* containing PA-1 cells that demonstrated a very low level of *L1-retro-EGFP* reactivation when grown in differentiation media (Garcia-Perez et al., 2010).

Here, we wanted to further explore *L1-retro-EGFP* silencing in PA-1 cells. We hypothesized that there are additional proteins that participate in *L1-retro-EGFP* silencing. We sought to develop a forward genetic screen that would potentially identify candidate genes involved with *L1-retro-EGFP* silencing. We employed CRISPR/Cas9 genome editing technology for this purpose (Barrangou et al., 2007; Ishino et al., 1987; Jinek et al., 2012). We used a commercially available platform called Genome wide CRISPR Cas9 Knock Out (GeCKOv2) (Shalem et al., 2014). This platform includes a lentiviral plasmid (lentiCRISPRv2 or LCv2) containing a single-guide RNA (sgRNA) sequence and a human codon-optimized Cas9 protein that can be packaged in a single lentivirus (Figure 3.1B) (Shalem et al., 2014). Using this platform we attempted to knock out nearly every gene in the PA-1 genome. Following gene knockout we performed

retrotransposition assays in knockout cell populations and assed their ability to express a *neomycin phosphotransferase* resistance gene delivered by an engineered L1 construct (L1-mneoI).

We determined the efficiency of genome editing over time and found that twenty-one days after transduction, PA-1 cells were sufficiently edited for use in our retrotransposition experiments. Using the same logic as in Garcia-Perez et al., 2010, we transfected edited PA-1 cells with pJM101/L1.3. (L1-mneoI). We reasoned that knockout of a gene involved in *L1-retro-mneoI* silencing would permit expression of the integrated *neomycin phosphotransferase*, rendering that cell resistant to the drug G418. We observed ~5X more G418 resistant colonies in edited PA-1 cells than in unedited PA-1 cells. These data suggest that in some edited PA-1 cells *L1-retro-mneoI* events escape silencing. Edited PA-1 cells resistant to G418 were then collected and inspected to determine guide sequence representation. Our preliminary screen uncovered potential candidate genes involved in *L1-retro-mneoI* silencing and we further investigated one of those here.

Results

PA-1 cells silence an L1-reporter

We first sought to verify that a *neomycin phosphotransferase* resistance gene delivered by an engineered L1 construct pJM101/L1.3 (L1-mneoI) is silenced in PA-1 cells. We employed a transient cell culture-based L1 retrotransposition assay (Moran et al., 1996; Wei et al., 2001). The L1 retrotransposition assay utilizes an episomal, engineered L1 expression construct containing a retrotransposition indicator cassette within the L1 3'UTR. The retrotransposition indicator cassette consists of an anti-sense copy of the *neomycin phosphotransferase* gene (mneoI*)* (Freeman et al., 1994; Moran et al., 1996*).* The coding sequence of mneoI is interrupted by an intron residing in the same transcriptional orientation as L1. This arrangement ensures that functional a *neomycin phosphotransferase* gene will only be activated upon L1 retrotransposition into genomic DNA (Freeman et al., 1994; Moran et al., 1996). Retrotransposition of *mneoI* confers resistance to the drug G418. If *mneoI* is delivered into the genome and

expressed, the resultant drug resistant foci provide a quantitative readout of L1 retrotransposition.

We transiently transfected PA-1 cells with a retrotransposition-competent L1 (RC-L1; pJM101/L1.3). In agreement with previous reports (Garcia-Perez et al., 2010), we found that PA-1 cells transfected with an RC-L1 were not resistant to the drug G418 and thus concluded *L1-retro-mneoI* events are efficiently silenced in PA-1 cells (Figure 3.1A). An additional control confirmed that these cells acquire G418 resistance when transfected with a plasmid that constitutively expresses the *neomycin phosphotransferase* resistance gene (pCDNA3) (Figure 3.1A). Given these results, and those from previous publications, we hypothesized that a genetic component was responsible for the silencing of *L1-retro-mneoI* (Garcia-Perez et al., 2010). We sought to investigate this hypothesis by taking advantage of CRISPR/Cas9 mediated genome editing technology. Our goal was to knock out genes individually in PA-1 cells and assess the effect of gene knockout on *L1-retro-mneoI* silencing.

PA-1 cells support transduction of lentivirus

We took advantage of a CRISPR/Cas9 editing platform called Genome wide CRISPR Cas9 Knock Out (GeCKO) (Shalem et al., 2014). This platform packages a plasmid (lentiCRISPRv2 or LCv2) containing a single-guide RNA (sgRNA) sequence and a human codon-optimized triple FLAG-tagged Cas9 protein in a single lentiviral particle (Figure 3.1B) (Shalem et al., 2014). The LCv2 plasmid also contains the *puromycin N-acetyl-transferase* (*pac)* gene that confers resistance to the drug puromycin (Shalem et al., 2014). Thus, successful transduction and integration of LCv2 into genomic DNA confers resistance to the drug puromycin.

We utilized the GeCKOv2 human sgRNA library (Sanjana et al., 2014; Shalem et al., 2014). The GeCKOv2 library contains 123,411 guide sequences (sgRNAs) divided equally between two libraries (A and B). Guide sequences in the GeCKOv2 library target 19,050 genes with six sgRNAs targeting each gene, 1,864 miRNAs with four sgRNAs targeting each miRNA, and 1000 control non-targeting sgRNAs (Sanjana et al. 2014).

Libraries A and B were first independently bulk-amplified using PCR, digested, and ligated into digested LCv2 vector. The resultant LCv2 plasmids contain a single guide sequence. To ensure that PCR amplification and cloning did not result in an artificial dropout of guide sequences, we submitted the LCv2 "A" library plasmids for deep sequencing on an Illumina Hi-Seq 2000. Of the 63,317 guide sequences in library "A", 62,764 (>99%) were represented in our final plasmid pool (performed in the Kitzman Laboratory) (Figure 3.2A). These results suggest we efficiently amplified and cloned guide sequences into the LCv2 plasmid. LCv2 "A" and "B" plasmids were separately packaged into a self-inactivating lentivirus expressing the vesicular stomatitis virus glycoprotein (VSV-G) with the aid of the University of Michigan Vector Core (Directed by Dr. Thomas Lanigan). Subsequent viral supernatants contained LCv2 "A" or LCv2 "B" lentivirus.

We first determined the viral titer of the "A" and "B" libraries to be ~$1.62x10^7$ infectious particles per milliliter of supernatant. Next we wanted to determine the kinetics of gene editing in PA-1 cells. We reasoned that guide sequences targeting critical cell survival genes would be lost over time (Blomen et al., 2015; Hart et al., 2014; Hart et al., 2015). In order to ensure that we did not artificially lose guide sequences because of under-transduction, which could thus confound our downstream analyses, we initially transduced $1.2x10^8$ million PA-1 cells at a multiplicity of infection (MOI) of 0.35 to ensure that each cell was infected by only one viral particle. Thus, each guide sequence in the transduced population was represented with ~295X coverage. Separate cultures of cells transduced with the "A" and "B" libraries were maintained.

Transduced PA-1 cells were cultured in puromycin selection media for 31 days and were passaged only upon reaching confluency (Figure 3.1C). Transduced PA-1 cells were collected at 8, 14, 21, 24, 28, and 31 days post-transduction (Figure 3.1C). At each passage cells were counted (~$1X10^9$ cells at each passage) and one third (~$3X10^8$ cells) of the cells were re-plated and cultured for continued growth, one third were retained for genomic DNA (gDNA) extraction, and one third were cryopreserved as cell stocks. The high density of re-plating was to ensure we retained a robust representation of guide sequences.

Cas9 is efficiently expressed from a transduced PA-1 cell line

We first wanted to determine the profile of Cas9 expression in our transduced PA-1 cells. Western blot analyses were conducted using whole cell lysates (WCLs) derived from LCv2 transduced PA-1 cells, from each of the six time points (collected 8, 14, 21, 24, 28, 31 days post-transduction). An anti-FLAG mouse monoclonal antibody (Sigma-Aldrich, F1804) was used for detection of triple FLAG-tagged Cas9 expression. We detected robust Cas9 expression from WCLs at each of the six time points (Figure 3.2A). These data verify that expression of Cas9 delivered by a lentivirus is maintained for at least 31 days in PA-1 cells.

The above result is important for two reasons. Firstly, sustained expression of Cas9 suggests that Cas9 mediated editing of genomic loci could continue throughout the course of our experiment. Sustained expression of Cas9 increases the likelihood that any given genomic loci is edited until it can no longer be recognized by the targeting sgRNA. Secondly, sustained expression of Cas9, coupled with the fact that transduced PA-1 cells remain puromycin resistant after 31 days, indicates that lentiviral delivered sequences are not silenced in PA-1 cells. These data are in agreement with a previous publication and support our hypothesis that genomic silencing of reporter genes delivered by L1 is a potentially distinct mechanism from silencing of lentiviral delivered sequences (Garcia-Perez et al., 2010).

CRISPR/Cas9 efficiently knocks out genes in the PA-1 cell line

To determine the editing status of PA-1 cells, we analyzed the complement of guide sequences present from cells collected 8 (T8), 21 (T21), and 31 (T31) days post-transduction (Figure 3.1C). After gDNA collection, guide sequences were bulk PCR amplified from both the “A” and “B” libraries using two rounds of PCR. The first round of PCR uses primers that flank the guide sequence of the integrated LCv2 and results in a product ~280 nt long. The second round of PCR adds eight nucleotide Illumina barcodes that allow multiplex sequencing, and complimentary sequences that bind to the flow cell oligonucleotide, resulting in products of ~330 nt. Amplified second round PCR products were size selected for products larger than 300 nt using Solid Phase Reversible Immobilization (SPRI). The 62,764 guide sequences from the LCv2 “A”

plasmid pool were deep sequenced on an Illumina Hi-Seq 2000 and were represented by ~55 million reads with a median coverage of 664 reads and 877 reads per guide sequence (performed by Dr. Jacob Kitzman) (Figure 3.2A). As we only had access to the sequencing data from the LCv2 "A" plasmid pool, we only analyzed guide sequences corresponding to the "A" library in our time course experiment.

We submitted our PCR products for sequencing on an Illumina MiSeq. At T8, 45,599 guide sequences were represented by 208,499 reads at a median coverage of 2 reads and a mean of 4.5 reads per guide (Figure 3.2B). At T21, 49,017 guide sequences were represented by 310,942 reads at a median coverage of 3 reads and a mean of 6.3 reads per guide (Figure 3.2B). At T31, 51,124 guide sequences were represented by 443,982 reads at a median coverage of 4 reads and a mean of 8.6 reads per guide (Figure 3.2B). The increase in represented guide sequences over time is likely due to increased sequencing depth at each time point.

We reasoned that the LCv2 "A" plasmid pool represented the starting population of guide sequences available for viral packaging and subsequent infection. To determine loss of guide sequences, we compared guides represented at T8, T21, and T31 to the guides represented in the original LCv2 "A" plasmid pool (Figure 3.2B). We used the Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout (MAGeCK) pipeline (Li et al., 2014) to analyze guide drop-out (Figure 3.2C). Additionally, MAGeCK performs a pathway analysis to determine if genes corresponding to pathways are over- or underrepresented in T8, T21, and T31 cells compared to LCv2 "A" plasmid pool.

MAGeCK pathway analysis uncovered that at T8 only genes associated with the spliceosome, 40/125 genes (KEGG id: hsa03040) and proteasome, 18/44 genes (KEGG id: hsa03050) were significantly ($p<1X10^{-4}$) depleted when compared to the LCv2 "A" plasmid pool (Figure 3.2D). In contrast, at T21, 60/125 spliceosome genes were depleted as well as 23/44 proteosome genes when compared to LCv2 "A" plasmid pool (Figure 3.2D). Additionally at T21, the ribosome, 39/86 genes (KEGG id: hsa03008), aminoacyl tRNA biosynthesis, 24/41 genes (KEGG id: hsa00970), RNA polymerase, 16/29 genes KEGG id: hsa03020), and RNA degradation pathway, 25/56

genes (KEGG id: hsa03018) pathways exhibited significant gene depletion ($p<1X10^{-4}$) when compared to LCv2 “A” plasmid pool (Figure 3.2D).

At T31, we observed a similar number of genes depleted in the same KEGG pathways as at T21 (Figure 3.2D). However, genes in two additional pathways, the pyrimidine metabolism, 46/96 genes (KEGG id: hsa00240) and valine, leucine, and isoleucine biosynthesis, 8/11 genes (KEGG id: hsa00290) pathways were also depleted ($p<1X10^{-4}$) when compared to LCv2 “A” plasmid pool (Figure 3.2D). These data are suggestive that 21 days post-transduction some genes representing pathways critical for cell survival are edited and effectively knocked out. Due to low read counts, our data set is likely underrepresenting the actual efficiency of our experimental design and deeper sequencing of our experimental populations should more accurately describe knockout efficiency.

PA-1 cells transduced with LCv2 support increased expression of *L1-retro-mneoI* compared to untransduced PA-1 cells

Our data suggested that T21 transduced PA-1 cells are efficiently edited. We reasoned that using the T21 population for our retrotransposition experiments would decrease the number of potential off-target effects mediated by sgRNA/Cas9 editing while also maintaining a high representation of guide sequences (Sanjana et al., 2014; Shalem et al., 2014). We combined T21 cells from both the “A” and “B” libraries and seeded ~$1x10^7$ cells in media containing puromycin. Roughly 72 hours after seeding we recovered ~$9X10^7$ T21 cells. Approximately $8.4x10^6$ T21 cells were seeded and subsequently transfected (see methods) in eight separate 15 cm plates; seven were transfected with the RC-L1, pJM101/L1.3, and one plate was left untransfected (Figure 3.3A). The seven individual transfections served as technical replicates. Previous experiments demonstrated that a small number of PA-1 cells (<.3%) do not silence *L1-retro-EGFP* (Garcia-Perez et al., 2010). To determine the background level of *L1-retro-mneoI* events that escape silencing in wild-type PA-1 cells, we transfected ~$8.4x10^6$ wild-type PA-1 cells with pJM101/L1.3.

Approximately seventy-two hours post transfection, cells were selected with 200ug/ml of the drug G418 (Figure 3.3A). By 14 days post transfection no un-

transfected T21 cells remained, indicating efficient drug selection. In contrast, 15 cm plates containing T21 cells transfected with pJM101/L1.3 contained ~150-300 visible G418 resistant colonies per plate. These colonies represent possible *L1-retro-mneoI* events that escaped silencing. In contrast the wild-type PA-1 cells transfected with the RC-L1 contained ~35 G418 resistant colonies, indicating the level of background of *L1-retro-mneoI* events that escape silencing (Figure 3.4A). To maximize our recovery of guide sequences we did not fix and stain any of these replicates and we are likely underestimating the true number of G418 resistant colonies (discussed further below) (Figure 3.4A). Thus, we observed that T21 cells exhibited at least a 5-fold increase in G418 resistant colonies compared to wild-type PA-1 cells.

We allowed the transfected plates containing colonies to grow for an additional two days in G418 selection media to increase cell number and quantity of genomic DNA. Roughly $4x10^6$ -$1.2x10^7$ G418 resistant T21 cells were collected from each of the seven 15 cm plates. We collected gDNA from each of the 15 cm plates keeping the samples separate. Guide sequences present in collected gDNA were PCR amplified using two rounds of amplification as described in the above section. Each of the seven samples was amplified with a different bar code allowing us to maintain independent replicates. PCR amplified guide sequences were then subject to deep sequencing on an Illumina MiSEq, as described above.

Sequencing these 7 samples yielded ~2 million total mappable reads with an average of ~$2.8x10^5$ reads per replicate plate. In total, 18,478 unique guide sequences were recovered for an average of 2,639 guides per replicate, and an average of 108 reads per guide. Closer inspection of the data revealed that the minority of guides represented the majority of sequencing reads. The top 100 most abundant guides represented ~15% of all reads, the top 200 most abundant guides represented ~23%, the top 1000 guides represented ~48%, and the top 5000 guides represented ~83% of all reads. Thus ~1% of the guide sequences (200/18,478) represented ~23% of all sequencing reads, and ~27% (5000/18478) of the guide sequences identified represented ~83% of all sequencing reads.

Analysis of potential candidate genes that silence *L1-retro-mneoI* in PA-1 cells

We generated a list of potential candidate genes using a very simple metric. We reasoned that if knockout of a gene resulted in expression of *L1-retro-mneoI,* then the guide sequence targeting that gene should be present in more than one of the seven replicates. Notably, this analysis was blind to the depth of sequencing reads represented for each guide. These criteria whittled our list of 18,478 guide sequences down to 2,850. These data suggest that most of the guide sequences we recovered are "one-offs" and may represent background in our system. We then determined which of the 2,850 guide sequences corresponded to the same gene. To be considered a potential candidate gene, we required that at least two different guides targeting a gene had to each be represented on at least two different replicates. Thus, our final list included 161 candidate genes, represented by at least two guides in at least two replicates (Figure 3.3B).

It could be that of the six guides targeting each gene, only one is highly efficient (HE-guide). To safeguard against the possibility that our selection criteria was too stringent, we added an additional criteria. We observed that 91 guides were present in four or more replicates and considered that these guides were likely highly HE-guides. We determined which genes the HE-guides targeted and then asked if at least one other single replicate guide targeted an HE-guide gene. As an example, an HE-guide targeting the *Tumor Protein 53* gene, *TP53,* was present in six replicates. Two other guides targeting *TP53* were also recovered, but each was present in only a single replicate. These criteria added an additional 27 potential candidate genes (Figure 3.3C).

An analysis of the top candidate gene, *NF2*

We further inspected the top candidate gene *Neurofibromin 2 (NF2* or *MERLIN*) which was targeted by five of six guide sequences (Figure 3.3B). Notably, two of the five guides were HE-guides represented in 7, and 6, replicates, respectively. The additional three guides were each represented in two replicates. *NF2* is generally considered to be a tumor suppressor gene and loss of *NF2* results in the formation of tumors composed of Schwann cells (schwannomas) on cranial and peripheral nerves (Welling et al., 2007).

In an initial attempt to validate *NF2* as a potential factor that silences *L1-retro-menoI* we sought to knock out *NF2* in wild-type PA-1 cells using a transient transfection assay. The JKP116 plasmid, (generously provided by Dr. Jacob Kitzman) contains restriction sites for guide sequence cloning, a human codon optimized Cas9 gene, and a *puromycin N-acetyl-transferase* gene, and can be used as an empty vector control (Ran et al., 2013). We designed three oligonucleotides containing a unique guide sequence targeting *NF2.* The guide sequences are identical to three of the five guide sequences recovered in our screen. Oligonucleotides were designed with additional sequences at their 5' and 3' ends that facilitate cloning into JKP116 (see methods). Successful cloning resulted in three plasmids, pPL_NF2_31760, pPL_NF2_31761, and pPL_NF2_31718, each targeting a unique sequence in the *NF2* gene. Wild-type PA-1 cells were seeded and concomitantly transiently transfected with pPL_NF2_31760, pPL_NF2_31761, pPL_NF2_31718, JKP116, or left untransfected. Approximately 48 hours after transfection cells were selected with 2ug/ml of the drug puromycin (Figure 3.4A). Approximately 72 hours post-selection no untransfected cells remained. PA-1 cells containing pPL_NF2_31760, pPL_NF2_31761, pPL_NF2_31718 or JKP116 were passaged into T-75 tissue culture flasks, and then into T-175 tissue culture flasks to increase cell number.

Once sufficient cell numbers were attained, we performed L1 retrotransposition assays. Wild-type PA-1 cells, or pPL_NF2_31760, pPL_NF2_31761, pPL_NF2_31718 or JKP116 PA-1 cells were transfected with pJM101/L1.3 and subject to selection exactly as described for the above retrotransposition assay (Figure 3.4A). As a control, we included untransfected pPL_NF2_31760, pPL_NF2_31761, and pPL_NF2_31718 PA-1 cells. Approximately 14 days after selection with the drug G418, cells were fixed and stained (Figure 3.4A). In pPL_NF2_31760, pPL_NF2_31761, and pPL_NF2_31718 PA-1 cells, we readily observed colonies resistant to the drug G418 (Figure 3.4B). This observation suggests that in some pPL_NF2_31760, pPL_NF2_31761, and pPL_NF2_31718 PA-1 cells, *L1-retro-mneoI* events escaped silencing. In JKP116 cells, and wild-type transfected PA-1 cells no such G418 resistant colonies were observed (Figure 3.4B). Crucially, untransfected pPL_NF2_31760, pPL_NF2_31761, and pPL_NF2_31718 PA-1 cells were susceptible to G418 treatment, which strongly

suggests knockout of NF2 does not simply confer drug resistance to PA-1 cells (Figure 3.4C). These preliminary data suggest that *NF2* may act directly or indirectly to silence expression of *L1-retro-mneoI* in PA-1 cells. Additional experiments are necessary to validate editing and knockout of *NF2.*

Discussion

PA-1 cells are edited by a CRISPR/Cas9 based genome-editing platform

We verified previous reports that PA-1 cells efficiently silence *L1-retro-mneoI* (Figure 3.1A) (Garcia-Perez et al., 2010). In an effort to identify factors that might be involved in *L1-retro-mneoI* silencing in PA-1 cells, we developed a forward genetic screen. We utilized a CRISPR/Cas9 based platform that packages an LCv2 plasmid into a lentivirus. The LCv2 plasmid contains a single-guide RNA (sgRNA) sequence, a human codon-optimized triple flag-tagged Cas9 protein, and *puromycin N-acetyl-transferase* gene that confers resistance to the drug puromycin (Shalem et al., 2014). Our first goal was to validate that the platform worked in our hands. We developed a scheme where we transduced PA-1 cells with LCv2 and analyzed genome editing at eight (T8), twenty-one (T21), and thirty-one (T31) days post-transduction (Figure 3.1C).

We first determined Cas9 protein expression at T8, T21, and T31. Western blot experiments demonstrated that PA-1 cells transduced with LCv2 robustly express Cas9 at least thirty-one days post-transduction (Figure 3.2A). In agreement with previous results, these data demonstrate that, unlike sequences delivered into the genomes of PA-1 cells by TPRT; sequences delivered by a lentivirus are not efficiently silenced in PA-1 cells (Garcia-Perez et al., 2010).

We next sought to determine the kinetics of gene editing in PA-1 cells. We reasoned that guide sequences targeting genes essential for cell survival would be lost in our population over time. We compared the complement of guide sequences in our T8, T21, and T31 populations to the complement of guide sequences present in the initial LCv2 plasmid pool using the MAGeCK software (Li et al., 2014). Analysis revealed that at T8, only genes in the proteasome, and spliceosome pathways were depleted, whereas at T21, genes in the spliceosome, proteasome, aminoacyl tRNA

biosynthesis, ribosome, RNA polyermase, and RNA degradtaon pathways were depleted (Figure 3.2D). Gene depletion was only modestly increased by T31 compared to T21 (Figure 3.2D). Thus, we decided to use T21 cells for our further analysis.

Detection of *L1-retro-mneoI* in PA-1 cells

Our goal was to identify genes that potentially silence *L1-retro-mneoI events* in PA-1 cells. We subjected T21 cells, and wild-type PA-1 cells to retrotransposition assays in an effort to satisfy this goal (Figure 3.3A). Successful retrotransposition and expression of *L1-retro-mneoI* results in resistance to the drug G418. We observed a ~5 fold increase in G418 resistant colonies in T21 cells compared to wild-type PA-1 cells. These data suggest a population of T21 cells were unable to silence *L1-retro-mneoI* events (Figure 3.A). We subsequently determined the complement of guide sequences present in the G418 resistant cells.

Our strategy for candidate gene selection was based on the seven technical replicates built within our experiment. We leveraged the idea that if a single guide was represented on more than one replicate it was more likely to be a true positive. We reasoned that any guides that are present on only a single replicate likely represented background. Surprisingly, 18,478 unique guide sequences were recovered across the seven replicates. We estimated that each replicate contained ~150-300 visible *L1-retro-mneoI* colonies. If each colony contains only one guide, and we assume the upper limit of colony number, then we would expect a maximum of 2,100 unique guides (300 colonies x 7 replicates). Thus, it appears that more guide sequences are present in our population than visible colonies. There are three possible explanations for this observation.

The first explanation is that we drastically underestimated the viral titer of the LCv2 lentivirus, resulting in multiple infections in each cell. This explanation is possible, but unlikely. Viral titer was determined independently three times by a colony-forming assay. Other assays for determining viral titer include qPCR as well as ELISA. Both qPCR and ELISA are reported to be more sensitive and reproducible than a colony-forming assay. It may be beneficial to subject our viral lysates to an additional assay to

more accurately determine titer and bolster our confidence that we accurately transduced PA-1 cells with an MOI of 0.35

The second explanation is that there are many more G418 resistant cells than what we observe visually. Estimation of colony number was based on observing live colonies as opposed to fixing and staining. It is very likely that we drastically underestimated the number of colonies present in each replicate. It could be that each replicate contained thousands of colonies represented by a small number of cells too small to see with the naked eye. Our two-step PCR could potentially amplify even the most rare guide sequences represented by only a few cells. It will be beneficial to fix and stain some 15 cm plates in future experiments to better determine the level of background G418 resistant cells.

The third explanation is coupled with the second explanation. Though we estimated a ~5 fold increase in G418 resistant colonies in T21 cells compared to wild-type PA-1 cells, the background in wild-type PA-1 cells may be higher. This notion is bolstered by the observation that of the 18,478 recovered guides in the G418 resistant T21 population, 1000 guides represented almost half of our total number of reads. Thus, it seems possible that there is a higher level of *L1-retro-mneoI* events that escape silencing than we initially determined. Future investigation will likely need to adjust drug selection parameters to increase the signal to noise ratio.

Additional modifications to the retrotransposition assay should also be considered to optimize and clarify downstream analysis. It is likely there are many single cells or small clusters of cells that are G418 resistant independent of an L1 retrotransposition event. Some of these cells could be differentiated cells that have stopped dividing, whereas others could stochastically remain affixed to the culture vessel. After completion of drug selection, resistant cells could be trypsinized and re-plated in a new culture vessel in the presence of G418. Differentiated cells are unlikely to adhere to the vessel after re-plating and G418 enforces maintained expression of the *L1-retro-mneoI.* Alternatively, cells could be transfected with an L1 containing an EGFP retrotransposition indicator cassette and subsequently be subject to flow cytometry. In

this way, only *L1-retro-EGFP* expressing cells that escape silencing due to a genetic mechanism will be identified.

Analysis of *NF2*

In this preliminary screen we identified 188 candidate genes that may silence *L1-retro-mneoI* in PA-1 cells (Figure 3.3B, C). Our investigation determined that, when knocked out, *L1-retro-mneoI* events were not silenced and thus conferred resistance to the drug G418. We investigated the top ranked gene, *NF2*, in an attempt to determine if we could recapitulate the result observed in the screen. We used a previously published method that relies on transient transfection of a plasmid as opposed to viral integration, to knock out *NF2* (Ran et al., 2013). We targeted *NF2* for knockout in wild-type PA-1 cells using three different guide sequences. We then subsequently subjected those cells to retrotransposition assays to determine if *L1-retro-mneoI* events escaped silencing (Figure 3.4A).

In all three knockout conditions we readily observed G418 resistant foci suggesting *L1-retro-mneoI* events escaped silencing in these cells. Control experiments yielded no such foci and importantly demonstrate that knockout of *NF2* does not result in the conferral of drug resistance (Figure 3.4B, C). These preliminary data suggest that *NF2* and its protein product, MERLIN, silence expression of *L1-retro-mneoI* in PA-1 cells. Additional experiments are required to validate that the PA-1 cells used for these experiments contain appropriately edited *NF2* loci. *NF2* knockout cells should also be transfected with an L1 that delivers an EGFP or *blasticidin* resistant gene. The expectation is that transfected *NF2* knockout cells should robustly express EGFP, or be resistant to the drug blasticidin, respectively. These experiments would further validate that *NF2* directly, or indirectly silences L1-reporter in PA-1 cells. Finally, only three of the five guide sequences targeting *NF2* were tested and it will be important to determine if the remaining two guide sequences similarly effect silencing of L1-reporter genes in PA-1 cells.

Here, we described the design and implementation of a genome wide screen to identify potential candidate genes that silence *L1-retro-mneoI* events in PA-1 cells. The 188 candidate genes identified here provide a starting point for further investigation.

Certainly, the parameters of the screen require additional optimization. However, through preliminary experimentation, we began to validate our top candidate gene as a potential factor that silences *L1-retro-mneoI* in PA-1 cells. These data are promising and suggest that with some minor improvements we will be able to confidently identify factors that may play a role in silencing sequences delivered by an engineered L1.

MATERIALS and METHODS

E. coli and the Propagation of Plasmids

All plasmids were propagated in DH5α *E. coli* (genotype: F- φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rk-, mk+) *pho*A *sup*E44 λ- *thi*-1 *gyr*A96 *rel*A1) (Invitrogen, Carlsbad, CA). Competent cells were generated as described previously (Inoue et al., 1990). Plasmids were prepared using the Plasmid Midi Kit (Qiagen, Germany) according to the protocol provided by the manufacturer.

Cell Lines and Cell Culture Conditions

PA-1 cells were cultured as previously described (Garcia-Perez et al., 2010). Briefly cells were grown in Minimum Essential Media (MEM) supplemented with 10% heat inactivated fetal bovine serum, 0.1 mM non-essential amino acids, 2mM L-Glutamine, 100U/ml penicillin, and 0.1 mg/ml streptomycin and subsequently filtered with a 0.22 uM filter (PA-1 Complete media). Cells were grown at 37ºC in a humidified incubator in the presence of 7% CO_2.

Plasmids

pJM101/L1.3: contains a full-length version of L1.3 in the pCEP4 backbone. The 3'UTR of L1.3 contains the *mneoI* retrotransposition indicator cassette (Dombroski et al., 1993; Moran et al., 1996; Sassaman et al., 1997).

pCDNA3: expresses a neomycin resistance gene and was acquired from Invitrogen.

LentiCRISPRv2: the packaging vector that containing the targeting guide sequence and Cas9 protein (Sanjana et al., 2014). The plasmid contains a psi+ packaging signal, a *rev* response element (RRE), central polypurine tract (cPPT), and a human U6 snRNA promoter driving expression of the guide sequence (Sanjana et al., 2014). The elongation factor 1α (EFS) short promoter drives expression of a *Streptococcus pyogenes* human codon-optomized Cas9/C-terminal FLAG octapeptide fusion protein as well as the puromycin resistance gene (Sanjana et al., 2014). A P2A self-cleaving peptide separates the Cas9/FLAG fusion protein from the puromycin resistance gene

(Sanjana et al., 2014). A woodchuck hepatitis virus posttranscriptional regulatory element (WPRE) is at the 3' end of the EFS transcriptional unit to enhance expression. Following transcription and translation the P2A self-cleaving peptide mediates the cleavage of the Cas9/FLAG tag peptide from the puromycin resistance peptide.

sgRNA libraries: the GeCKOv2 sgRNA libraries have been previously described (Sanjana et al., 2014; Shalem et al., 2014). Briefly, the GeCKOv2 library contains 123,411 guide sequences (sgRNAs) divided equally between two libraries (A and B). Guide sequences in the GeCKOv2 library target 19,050 genes with six guides targeting each gene, 1,864 miRNAs with four guides targeting each miRNA, and 1000 control non-targeting guides (Sanjana et al. 2014). The libraries are separated into two sub-libraires (A and B) each containing half of the total complement of guides. GeCKOv2 library was purchased from Addgene.

Cloning sgRNAs into lentiCRISPRv2 (note: these experiments performed in the Kitzman laboratory):

The GeCKOv2 plasmid library was first digested with *Bsm*BI to release the guide sequences. The *Bsm*BI insert sequences were then PCR amplified using custom primers to increase guide copy number. PCR products were then digested with *Bsm*BI and ligated into *Bsm*BI digested LentiCRISPRv2 plasmids (Sanjana et al., 2014; Shalem et al., 2014). The resultant LentiCRISPRv2 plasmid contains a single guide sequence and can be subsequently packaged into lentivirus.

Viral production and packaging of LentiCRISPRv2

Lentivirus packaging was performed at the University of Michigan Vector Core (Dr. Thomas Lannigan). For each sub library 650 µg of purified proviral plasmid were packaged into lentiviral particles using the psPAX2 plasmid containing the *gag* and *pol* genes and the vesicular stomatitis virus G glycoprotein (VSV-G) envelope protein expressing plasmid. Human embryonic kidney A293T cells were used for lentiviral production. Supernatants were subsequently collected and contained the mature viral particles. Mature viral particles contain the lentiCRISPRv2 proviral sequence and can

be used to transduce cells. Note viral supernatants from sub-libraries A and B were maintained separately.

Viral transduction and growth of PA-1 cells

PA-1 cells were plated in 20 15 cm dishes (BD Biosciences) at a density of $6X10^6$ cells/plate in 20 plates—yielding $1.2X10^8$ cells total. Twenty-four hours post-plating 10, 15 cm dishes were transduced by replacing PA1-Complete media with PA1-Complete media that contained the viral supernatant from either sub-library A, or sub-library B at an MOI of 0.3 and 8 µg/ml Polybrene (Sigma). Transduced PA-1 cells from each sub-library were maintained separately throughout the duration of selection. Twenty-four hours post transduction media was replaced with fresh PA-1 Complete media. Forty-eight hours post transduction cells were selected with PA-1 Complaete media containing (1 ug/ml) of the drug puromycin daily for the remainder of the experiment. Cells were collected and reseeded on days 8, 14, 21, 24, 28, and 31. At each time point over $1X10^9$ cells were collected; ~$4X10^7$ cells from each sub-library were reseeded into 10, 15 cm dishes and allowed to grow until they reached confluency, $5X10^7$ cells were cryofrozen to be maintained as cell stocks; and $2X10^8$ cells were frozen at -30 C and saved for gDNA collection.

gDNA collection

Cell pellets were thawed and genomic DNA was extracted using the QIAGEN DNeasy Blood and Tissue Kit per the manufacturers recommendations. DNA was extracted from $2X10^8$ cells in supplied elution buffer and stored at 4C for further analysis. We typically recovered ~150 ug gDNA from each time point. Genomic DNA collected from PA-1 cells subject to retrotransposition assays was performed as above. Genomic DNA from each of the 7 replicate plates were collected independently yielding ~7.5 ug gDNA per plate.

PCR amplification of guide sequences recovered from genomic DNA for drop-out analysis

Genomic DNA collected drop-out analysis was subjected to two rounds of PCR amplification. PCRs were performed as previously described (Sanjana et al., 2014; Shalem et al., 2014). To ensure complete coverage guide sequences and to avoid

artificial dropout multiple PCR reactions from the same sample were performed. Eight PCR reactions from each time point (T8, T21, T31) were performed with the KAPABIOSYSTEMS HiFi PCR kit (KapaBiosystems). Reactions were performed as recommended by the manufacturer with the following exception that 2.5ug of DNA was used per reaction. An initial denaturation time of 5 minutes at 95 °C was followed by 27 cycles of amplification, using a 20 second annealing step at 60 °C, and a 30 second elongation step at 72 °C (Sanjana et al., 2014; Shalem et al., 2014). Primers used for PCR1 added sequences necessary for MiSEQ sequencing and are available upon request. Round 2 PCR amplification was performed using the same protocol as PCR 1 with the following exceptions: 5 ul of PCR 1 product was used in the reaction, initial denaturation was 2 minutes at 95 °C and the was cycled seven times. Primers used in PCR 2 are P7 and P5 barcode adaptor primers provided by Illumina. All concentrations of reagents with the exception of starting gDNA material were followed as recommended by the manufacturer. Oligos for round 2 PCR add 8 bp barcodes onto the PCR 1 product allowing multiplexing and sequencing of numerous samples on a single MiSEQ run. Each of the 3 time point samples received their own unique combination of P5 and P7 barcodes. PCR products from PCR 2 were size selected to enrich for products greater than 300 bp using the SPRI-cleanup based Beckman Coulter Agencourt AMPure XP purification system following the manufacturers recommendations.

PCR amplification of guide sequences recovered from genomic DNA collected from retrotransposition assays

Genomic DNA from each of the seven replicates was collected as described above and was subjected to two rounds of PCR amplification. PCRs were performed as previously described (Sanjana et al., 2014; Shalem et al., 2014). A single PCR reaction was performed independently for each of the 7 replicates using the KAPABIOSYSTEMS HiFi PCR kit. Reactions were performed as described for time point analysis except 0.5 ug of gDNA was used. In addition to amplifying the guide sequence, primers added sequences used for MiSEQ sequencing and are available upon request. Round 2 PCR amplification was performed using the same protocol as PCR 1 with the following

exceptions: 5 ul of PCR 1 product was used in the reaction, initial denaturation was 2 minutes at 95 °C and the reaction was cycled seven times. Primers used in PCR 2 are P7 and P5 barcode adaptor primers provided by Illumina. All concentrations of reagents with the exception of starting gDNA material were followed as recommended by the manufacturer. Primers for round 2 PCR add 8 bp barcodes onto the PCR 1 product allowing multiplexing and sequencing of numerous samples on a single MiSEQ run. Each of the seven technical replicates received their own unique combination of P5 and P7 barcodes. PCR products from PCR 2 were size selected to enrich for products greater than 300 bp using the SPRI-cleanup based Beckman Coulter Agencourt AMPure XP purification system following the manufacturers recommendations.

MiSEQ sequencing of PCR amplified guide sequences from gDNA

Sequencing of T8, T21, and T31 PCR were performed on a MiSEQ (Illumina) using MiSEQ Reagent Kits V3 with a 2 X 75 output. The final concentration of the the T8, T21, and T31 PCR products together was 12 pmol total. Cleaned, size selected products from all three time points were multiplexed on one MiSeq run. Sequencing of PCR products from the seven technical replicates was performed as above maintaining 12 pmol final concentration and all 7 samples were multiplexed on the same sequencing run.

MAGeCK data analysis

Raw FASTQ files were downloaded and demultiplexed and trimmed to contain only the raw guide sequence. Alignment of reads required construction of an indexed reference of all guide sequences. Sequenced guides were then mapped to that reference index. For time T8, T21, and T31 time point analysis the number of reads for each unique guide sequence was normalized to the total number of reads for that timepoint. As each different time point produced a different number of reads, guide sequence read normalization compared to the plasmid pool was normalized independently for each timepoint. To determine enrichment and depletion of genes we compared guide sequences present in the plasmid pool to guide sequences present at T8, T21, and T31. To accurately perform the analysis reads from the plasmid pool and from the various time points were median normalized. MAGeCK (Li et al., 2014) was downloaded from

(https://sourceforge.net/p/mageck/wiki/Home/) using versioin 0.5.4. to analyze gene enrichment and depletion. Analysis was performed as described in Li et al. 2014.

Protein collection

The portion of the cells that were collected from the time course experiment and frozen at -30 °C were thawed and used for protein collection. After thawing, ~5X10^6 cells were lysed for 15 minutes on ice and incubated in 0.5 mL of lysis buffer: 10% glycerol, 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% NP-40 (IGPAL) (Sigma-Aldrich, St. Louis, MO), and 1X Complete Mini EDTA-free Protease Inhibitor Cocktail (Roche Applied Science, Germany). The resultant protein lysates then were centrifuged at 15,000 x g for 30 minutes to clear the lysate. The resultant supernatant (approximately 0.4mls) was designated as the whole cell lysate (WCL). Bradford assays (Bio-Rad Laboratories, Hercules, CA) were used to determine protein concentrations. WCLs generally yielded ~10 µg/µl of protein. The protein samples were stored at -80 ^{o}C.

Western blots

Protein samples were collected as described above and then were incubated with a 2X solution of NuPAGE reducing buffer (containing 1.75-3.25% lithium dodecyl sulfate and 50 mM dithiothreitol (DTT)) (ThermoFisher Scientific). An aliquot (20 µg) of the reduced proteins were incubated at 100^{o}C for 10 minutes and then were separated by electrophoresis on 10% precast mini-PROTEAN® TGX gels (Bio-Rad Laboratories, Hercules, CA) run at 200V for 1 hour in 1X Tris/Glycine/SDS (25 mM Tris-HCL, 192 mM glycine, 0.1% SDS, pH 8.3) buffer (Bio-Rad Laboratories). Transfer was performed using the Trans-Blot® Turbo™ Mini PVDF Transfer Packs (BioRad Laboratories) with the Trans-Blot® Turbo™ Transfer System (BioRad Laboratories) at 25V for 7 minutes. The membranes then were incubated at room temperature in LI-COR blocking solution (LI-COR Biosciences, Lincoln, NE) for 30 minutes. Following blocking, fresh blocking solution was added with the following antibodies: FLAGM2 antibody (Agilent Technologies, Santa Clara, CA, cat # 200472) was diluted at 1:2,000; actin antibody (ThermoFisher Scientific, MA1-744) was diluted at 1:2,000. The membrane was incubated overnight at 4 C. After incubation the membrane was washed 3X with 1X PBS and secondary antibody was applied. Fresh blocking solution was added with the

following secondary antibodies: Anti-Mouse IRDye 680LT (LICOR Biosciences, Lincoln, NE, cat # 925-68022) diluted at 1:10,000 was used to visualize the actin loading control, and anti-Rabbit IRDye 800CW (LI-COR Biosciences, cat # 925-32213) was diluted at 1:10,000 and used to visualize FLAGM2. The membrane was incubated for 2 hours at room temperature with secondary antibody. Following incubation the membrane was washed 3X with 1X PBS and fluorescent signals were detected with an Odyssey CLx (LI-COR Biosciences).

L1 Retrotransposition Assay

T21 cryofrozen cells from both sub-library A and B were thawed and cultured together in three T-175 flasks (BD Biosciences) and allowed to grow to confluency.. Seven, 15 cm dishes (BD Biosciences) were seeded with 8.4×10^6 T21 cells/dish in 20 ml of PA-1 Complete media containing the transfection cocktail. The transfection cocktail is as follows; 1.68 ml of Opti-MEM® (Life Technologies), 67.2 µl FuGENE6 (Promega) transfection reagent, and 16.8 µg pJM101/L1.3 plasmid DNA per 15 cm dish. Approximately 24 hours post-transfection, the media was replaced with PA-1 Complete media to stop the transfection. Three days post-transfection, the tissue culture medium was replaced and the cells were grown in PA-1 Complete media supplemented with 200 µg/mL of G418 (Life Technologies) to select for retrotransposition events. After 16 days of G418 selection, the resultant G418-resistant foci were washed with ice cold 1X Phosphate-Buffered Saline (PBS) and collected using a cell scraper. Cells from each plate were collected separately and subsequently frozen at -30 C to be used for future gDNA collection and analysis.

Cloning of NF2 knockout plasmids

Cloning of oligos into JKP116 was performed as described in Ran et al., 2013. Oligos used in cloning reactions were ordered from Integrated DNA Technologies (IDT), see below for oligo sequences. Sense and antisense oligos were used at a concentration of 100 uM and were first phosphorylated using T4 PNK (NEB) following the manufacturers recommendation and subsequently annealed at 37 °C for 30 minutes. Enzyme was subsequently heat killed at 95 °C for five minutes, and temperature was ramped down to 25 °C at 5 °C /min. Annealed oligos were stored at 4

°C until future use. Plasmid JKP116 was generously provided by the Kitzman laboratory. ~ 100 ng of JKP116 plasmid was digested with the enzyme *BbsI* (NEB) following manufacturers recommendations. Following digestion annealed oligos at a concentration of 0.25 uM were incubated with the digested JKP116 plasmid at 37 °C for 5 minutes, followed by 23 °C for five minutes and repeated 6 times. Resultant plasmids were transformed into competent bacteria and plated on ampicillin containing plates to select for drug resistant bacteria. Individual bacterial clones were picked and expanded. Plasmid DNA extracted from clones was sequenced to ensure it contained the proper guide sequence insert.

Oligonucleotide guide sequences

Oligonucleotides are designed so they can easily be cloned into digested *BbsI* JKP116 plasmid (Ran et al., 2013). Underline indicates the actual sequence targeting the *NF2* gene.

NF2_31718_TOP: 5'-CACCGATTCCACGGGAAGGAGATCT-3'

NF2_31718_BOT: 5'-AAACAGATCTCCTTCCCGTGGAATC-3'

NF2_31760_TOP: 5'-CACCGCCTGGCTTCTTACGCCGTCC-3'

NF2_31760_BOT: 5'-AAACGGTCGGCGTAAGAAGCCAGGC-3'

NF2_31761_TOP: 5'-CACCGAAACATCTCGTACAGTGACA-3'

NF2_31761_BOT: 5'-AAACTGTCACTGTACGAGATGTTTC-3'

Cloning of annealed oligonucleotides into JKP116 yielded pPL_NF2_31718, pPL_NF2_31760, and pPL_NF2_31761, resultant plasmids are resistant to the drug puromycin and contain a Cas9 expression sequence.

Knock out of NF2 in PA-1 cells

PA-1 cells were seeded in PA-1 Complete media at a density of $2x10^5$ cell/well in each well of a six-well dish. Five of six wells were seeded in the presence of the transfection cocktail. Transfection cocktail contained 100 µl Opti-MEM® (Life Technologies), 3 µl FuGENE6 (Promega) transfection reagent, and 1 µg of

pPL_NF2_31718, pPL_NF2_31760, pPL_NF2_31761, or JKP116 plasmid DNA per well of a 6-well plate, the sixth well remained untransfected. Twenty-four post transfection media was changed with PA-1 complete media to stop the transfection. Fort-eight hours post transfection cells were selected with cells were selected with PA-1 Complete media containing 2 ug/ul of the drug puromycin. Seventy-two hours of drug selection resulted in no untransfected cells remaining. Resultant cells were grown in PA-1 Complete media lacking puromycin. Cells from all five remaining wells were passaged twice to increase cell number.

L1 Retrotransposition Assay in knockout cell lines

The cultured cell retrotransposition assay was conducted as described previously (Garcia-Perez et al., 2010; Kopera et al., 2016; Moran et al., 1996; Wei et al., 2001). Briefly, 2×10^5 PA-1 pPL_NF2_31718, pPL_NF2_31760, pPL_NF2_31761, JKP116, or wild-type PA-1 cells/well were plated in 6-well tissue culture dishes (BD Biosciences). Approximately 24 hours post-plating, transfections were performed in three of the six wells using a mixture containing 100 µl Opti-MEM® (Life Technologies), 3 µl FuGENE6 (Promega) transfection reagent, and 1 µg pLM101/L1.3 DNA per well of a 6-well plate. Three wells were left untransfected. Approximately 24 hours post-transfection, the media was replaced with PA-1 Complete medium to stop the transfection. Three days post-transfection, the tissue culture medium was replaced and the cells were grown in PA-1 complete medium supplemented with 200 µg/mL of G418 (Life Technologies) to select for retrotransposition events. After approximately 14 days of G418 selection, the resultant G418-resistant foci were washed with ice cold 1X Phosphate-Buffered Saline (PBS), fixed to the tissue culture plate by treating them for 10 minutes at room temperature in a 1X PBS solution containing 2% paraformaldehyde (Sigma Aldrich) and 0.4% glutaraldehyde (Sigma Aldrich), and stained with a 0.1% crystal violet solution for 30 minutes at room temperature to visualize the G418-resistant foci.

Figure 3.1: Schematic design of PA-1 screen.

A) *L1-retro-mneoI events are silenced in PA-1 cells.* PA-1 cells were transfected with pJM101/L1.3, left image. The retrotransposition of pJM101/L1.3 delivers a *neomycin phosphotransferase* resistance gene (*mneoI*) conferring resistance to the drug G418 only if the integrated *mneoI* reporter (*L1-retro*-mneoI) is expressed. The pJM101/L1.3 transfected PA-1 cells are susceptible to the drug G418, indicating silencing of *L1-retro-mneoI*. PA-1 cells transfected with pCDNA3, which constitutively express *neomycin phosphotransferase,* are resistant to G418 B) *Cartoon of the LentiCrispr version two (LCv2) plasmid*. Light blue arrow indicates U6 promoter driving expression of the sgRNA sequence. Purple box is variable guide sequence. Blue arrow indicates elongation factor 1 (EFS) promoter driving expression of Cas9. Yellow box indicates Cas9, containing a red 3X-FLAG tag at its 3' end. The entire sequence is flanked by Long Terminal Repeates (LTRs) indicated by white boxes. C) *Design of PA-1 transduction experiment with LCv2 lentivirus.* On Day, 1 PA-1 cells are transduced with virus. On Day 8, PA-1 cells reached confluency and were trypsinized. Roughly 1/3 of all cells collected were then either reseeded into new plates (Reseed), cryofrozen as cell stocks (Cell Stocks), or collected for genomic DNA extractions (gDNA). That process occurred every day indicated on the schematic (Day 8, 14, 21, 24, 28, 31). The red box around Day 8, Day 21, and Day 31 (corresponding to T8, T21, and T31 in Figure 3.2B, D) depicts which gDNAs were used for time point analysis (see text).