LINE-1 Retrotransposition in Human Genomic Variation

by

Christine R. Beck

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2012

Doctoral Committee:

Professor John V. Moran, Chair
Professor Jeffrey W. Innis
Professor Jianzhi Zhang
Associate Professor JoAnn Sekiguchi

I dedicate this thesis to my husband, Tanner Beck

Acknowledgements

I would like to thank my mentor, Dr. John Moran, for his constant and infectious excitement about science, and for his critical analyses of my data, manuscripts, and presentations. He has helped the graduate school experience exceed my expectations at every turn, and to make me the scientist I am today. For his guidance, I will be forever grateful.

Many amazing scientists other than my advisor have helped me along the way. My thesis committee members, Dr. JoAnn Sekiguchi, Dr. Jeffrey Innis, and Dr. George Zhang have contributed greatly to the pages herein through their scientific input and criticism. I additionally wish to thank Dr. Jeffrey Long, who started as a member of my committee, but left Michigan to join the University of New Mexico department of Evolutionary Anthropology. Dr. Deanna Kulpa was the first person I met in the lab, and not only has she been a constant companion for scientific discussion and pondering, but is also a wonderful friend. The members of the Moran lab are intelligent and critical thinking individuals that have shaped my science and presentations through tough and imaginative lab meetings, lunches and conversations. The graduate students in the department of Human Genetics helped welcome me into a different aspect of science from my biochemistry stomping grounds, and were there in clutch moments of my

thesis work. Dr. Kate Barald warmly greeted me upon my arrival to the University of Michigan, and is a wonderful friend and mentor. Dr. Martin Arlt has been a helpful scientific and professional inspiration in the last few years, and I am lucky to have him in close proximity to our lab. Additionally, I greatly appreciate my collaborators, and they have played important roles in my research. Their work, and our conversations about data, made this thesis possible. Thanks especially to Dr. Richard Badge and Dr. Evan Eichler.

I would also like to thank my family and friends for supporting me throughout the course of graduate school. My parents, Jim and Susan Rowley, were tolerant of my curiosity from an early age, and have been keen on informing me when I'm being too hard on myself. My in laws, Teresa, Chuck, and Corey Beck, have kept me laughing and continually remind me of the important things in life. My sister, Ann, has helped me understand the benefit of having a tough skin, and has been there to talk through the trials and tribulations of every step of the Ph.D. ladder. My wonderful friends who have held me up from time to time know who they are, but especially to Aislinn and the dudes of the lab, thanks for everything.

Most of all, this thesis was possible because of my husband Tanner. He left Iowa to come with me on this journey, and we'll have fun on all of our future trips together, too. I could not have made it through some of the last ~2,000 days without him to come home to, or without his steadfast support of my intellectual pursuits.

Table of Contents

List of Figures

List of Tables

Abstract

LINE-1 Retrotransposition in Human Genomic Variation

by

Christine R. Beck

Chair: John V. Moran

Long interspersed element-1 (LINE-1 or L1) is a ubiquitous mobile element in mammalian genomes. There are ~500,000 copies of L1 throughout the human genome, comprising ~17% of our DNA. Interestingly, though the majority of these elements are inactive due to 5' truncation or mutations in the two L1-encoded open reading frames (ORF1 and ORF2), a small number of L1s (~80-100 per human genome) are capable of mobilization by the copy and paste mechanism of retrotransposition termed target-site primed reverse transcription (TPRT). As L1s rely upon their encoded proteins (ORF1p and ORF2p) for mobility, only intact, full-length L1s are potentially active. Previous analysis of the human genome reference sequence (HGR) showed that 90 L1s contained intact ORFs, and that 6 of these were responsible for >80% of the retrotransposition activity. These 6 highly active L1s were polymorphic in humans. Therefore, I

hypothesized that low allele frequency L1s comprise the majority of activity in human populations.

I used fosmid libraries developed from six geographically diverse individuals to identify 68 full-length L1s absent from the HGR. Approximately 55% (37/68) were highly active when tested in a cultured cell assay. To determine the allele frequency of the L1s in the population, 26 of the 68 elements were examined in genotyping panels. Four of the 26 were either private or African specific when typed on the H952 subset of the human genome diversity panel. The sequences of the 68 L1s showed that 53 (37 highly active and 16 low-level or inactive L1s) contained intact ORF1 and ORF2. The sixteen L1s were then examined for changes that led to their inability to retrotranspose with high efficiency in cell culture. Using sequence comparisons and functional assays, I identified a novel amino acid change in the ORF2-encoded protein of one element and a splicing mutation in the 5' untranslated region of another L1.

My examination of individual genomes readily identified highly active, rare L1s, and suggests that L1 activity in humans is more prevalent than previously appreciated. The sequences of these elements present future opportunities for the elucidation of L1 biology.

Chapter 1

LINE-1 Elements in Structural Variation and Disease

Overview

This thesis is centered on the role of L1 in generating human variation, and the insights into L1 biology that can be gained through the study of human-specific elements. Here, I present an introductory chapter (Chapter 1) followed by a published manuscript (Chapter 2), additional experimental data (Chapter 3), and a concluding chapter (Chapter 4).

The first chapter, entitled “LINE-1 Elements in Structural Variation and Disease” is an introduction to the various classes of mobile elements that exist in human genomes and the roles they play in inter-individual human variation and disease. This chapter was published as a review in the 2011 volume of *The Annual Review of Genomics and Human Genetics*. The second chapter, “LINE-1 Retrotransposition Activity in Human Genomes”, examines the human-specific, polymorphic L1s present in 6 individuals. This publication appeared as an article in *Cell*, and was prepared in collaboration with the laboratories of Dr. Evan Eichler and Dr. Richard Badge. Both Chapters 1 and 2 were presented with

permission from the publishers. Chapter 3 discusses experimental data and other analysis regarding the mutations present in the L1s examined in Chapter 2. This Chapter includes experiments that are a part of a manuscript in preparation from the Badge lab regarding transduction-specific amplification typing of L1 active subfamilies (TS-ATLAS). Chapter 4 discusses conclusions and future directions of the studies presented herein.

Abstract

The completion of the human genome reference sequence ushered in a new era for the study and discovery of human transposable elements. It now is undeniable that transposable elements, historically dismissed as junk DNA, have had an instrumental role in sculpting the structure and function of our genomes. In particular, long interspersed element-1 (LINE-1 or L1) and short interspersed elements (SINEs) continue to affect our genome, and their movement can lead to sporadic cases of disease. Here, we briefly review the types of transposable elements present in the human genome and their mechanisms of mobility. We next highlight how advances in DNA sequencing and genomic technologies have enabled the discovery of novel retrotransposons in individual genomes. Finally, we discuss how L1-mediated retrotransposition events impact human genomes.

Introduction

Approximately 45% of the human genome is derived from transposable elements (Lander et al., 2001). These include DNA transposons, long terminal repeat (LTR) retrotransposons, and non-LTR retrotransposons. Although most

transposable elements have been rendered inactive through mutation, long interspersed element-1 (LINE-1 or L1) retrotransposition continues to diversify human genomes.

L1s comprise ~17% of human DNA (Lander et al., 2001). The L1-encoded proteins (ORF1p and ORF2p) can mobilize non-autonomous retrotransposons, other non-coding RNAs, and messenger RNAs, leading to the generation of processed pseudogenes (Buzdin et al., 2002; Dewannieux et al., 2003; Esnault et al., 2000; Garcia-Perez et al., 2007a; Gilbert et al., 2005; Hancks et al., 2011; Wei et al., 2001). Thus, in total, L1-mediated retrotransposition has generated a third of our genome.

In 1988, an examination of 240 unrelated males afflicted with hemophilia A revealed that independent mutagenic L1 insertions into exon 14 of the *Factor VIII* gene were responsible for the disease in two individuals (Kazazian et al., 1988). Heroic efforts to isolate an active progenitor L1 (Dombroski et al., 1991) and the development of a cultured cell retrotransposition assay (Moran et al., 1996) then helped elucidate the molecular mechanism of L1 retrotransposition. It now is apparent that L1s are alive and well in human populations, and that L1-mediated retrotransposition events account for approximately 1 of every 1,000 of spontaneous, disease-producing insertions in man (Chen et al., 2005a; Kazazian and Moran, 1998).

Several reviews have discussed aspects of human retrotransposon biology and how the host genome defends itself from retrotransposon activity (Babushok and Kazazian, 2007; Belancio et al., 2008a; Cordaux and Batzer, 2009; Goodier

and Kazazian, 2008; O'Donnell and Burns, 2010; Ostertag and Kazazian, 2001). Here, we discuss recent progress in understanding the mechanism of L1 retrotransposition. We then highlight how new genomic technologies have illuminated the impact of L1-mediated retrotransposition events on human genetic variation and genome structure.

Mobile Elements in Human Genomes

Transposable elements are classified by whether they mobilize via a DNA (DNA transposons) or an RNA (retrotransposons) intermediate (Figure 1.1). They further are distinguished by whether they encode proteins to mediate their own mobility (autonomous elements) or rely upon proteins encoded by other elements (nonautonomous elements).

DNA Transposons

DNA transposons generally move via a cut-and-paste mechanism. They tend to have a limited lifespan in higher eukaryotic genomes (Lander et al., 2001; Smit, 1996), which likely is due to the accumulation of non-autonomous deletion derivatives that compete for the transposase encoded by autonomous elements. In most cases, transposase binds at DNA transposon inverted repeat sequences, “cuts” the transposon from its existing location, and then “pastes” it into a new genomic location (reviewed in (Craig et al., 2002)).

DNA transposons comprise ~3% of the human genome reference sequence (HGR, NCBI 36/hg18). Sequence divergence among paralogous copies indicates that virtually all DNA transposons mobilized prior to the

eutherian radiation, whereas a composite method indicates they have been extinct in the primate lineage for at least 37 million years (Lander et al., 2001; Pace and Feschotte, 2007). Nevertheless, DNA transposons have had an enduring effect on the human genome. For example, the recombination activating genes, *RAG1* and *RAG2*, which are critical for V(D)J recombination and immune system development, likely were domesticated from the *Transib* family of DNA transposons ~500 million years ago (Kapitonov and Jurka, 2005; Zhou et al., 2004).

Engineered DNA transposons have practical applications, and can be exploited for useful purposes. For example, a reanimated salmon DNA transposon, Sleeping Beauty, has been used to discover genes implicated in cancer progression and shows promise as a delivery vehicle in gene therapy studies (Collier and Largaespada, 2005; Hackett et al., 2010; Ivics et al., 1997, 2009). Similarly, an insect DNA transposon, piggyBac, has been used to create gene-specific knockouts in mouse embryonic stem cells (Ding et al., 2005; Sun et al., 2008). Finally, a zebrafish transposon, Tol2, shows promise as a mutagen in both the mouse and zebrafish germ line (Kawakami, 2005; Keng et al., 2009).

Retrotransposons

Retrotransposons mobilize via an RNA intermediate by a copy-and-paste mechanism and remain active in most mammalian genomes. They can be subdivided into two general classes, depending on whether they contain or lack LTRs (Figure 1.1).

Long Terminal Repeat Retrotransposons: Human Endogenous Retroviruses

LTR-containing elements, such as human endogenous retroviruses (HERVs), resemble retroviruses in both their structure and mobility mechanism. Most HERVs contain a non-functional envelope (*ENV*) gene, which relegates them to an intracellular existence (Bannert and Kurth, 2006).

HERVs and their non-autonomous derivatives comprise ~8% of the human genome (Lander et al., 2001). Virtually all HERVs are retrotransposition-defective; however, a small number of HERV-K elements (where K denotes the host lysine transfer RNA (tRNA) that presumably initiates HERV-K (-) strand complementary DNA (cDNA) synthesis) are polymorphic with respect to presence/absence status in humans, indicating that they have retrotransposed since the human-chimpanzee divergence (Belshaw et al., 2005; Kidd et al., 2010; Macfarlane and Simmonds, 2004; Moyes et al., 2007). Moreover, some HERV-K elements contain intact open reading frames (ORFs) (Mayer et al., 1997), and a reanimated HERV-K virus is infectious in cultured cell assays (Dewannieux et al., 2006; Lee and Bieniasz, 2007). Thus, it is formally possible that rare HERV-K alleles retain the ability to move in modern humans.

Recent studies indicate that HERV-K retrotransposons are expressed in certain tumors, and chromosomal rearrangements involving HERV-K sequences have been implicated in prostate cancer (Moyes et al., 2007; Tomlins et al., 2007). HERV expression also has been implicated in the etiology of certain autoimmune and neurological diseases; however, these data remain

controversial and causal links between HERV expression and disease require additional experiments (reviewed in Moyes et al., 2007).

Although apparently immobile in humans, the remnants of endogenous retroviruses also can impact the function of mammalian genomes. A growing number of examples indicate that *cis*-acting sequences derived from endogenous retroviruses can play roles in the transcription, splicing, and/or epigenetic regulation of endogenous genes (Cohen et al., 2009; Maksakova et al., 2006; Michaud et al., 1994; Robins and Samuelson, 1992). Moreover, sequences derived from endogenous retroviruses have been exapted by mammalian genomes, and now play important roles in placental development (*e.g.,* Syncytin and peg10 (Mi et al., 2000; Ono et al., 2006)).

Non-Long Terminal Repeat Retrotransposons: LINE-1 Elements

L1s are the only known autonomously active human retrotransposons, and account for approximately one-sixth of our genome (Lander et al., 2001). Over 99.9% of L1s have been rendered inactive by 5' truncations, inversions, and/or point mutations within the two L1-encoded ORFs (Grimaldi et al., 1984; Lander et al., 2001; Ostertag et al., 2001b). However, a consensus sequence derived from human genomic L1s suggested the existence of full-length, retrotransposition-competent L1s (RC-L1s) (Scott et al., 1987). Indeed, the subsequent isolation of the progenitors of mutagenic L1 insertions into the *Factor VIII* and *dystrophin* genes revealed that a cohort of L1s continue to mobilize in our genome (Dombroski et al., 1991; Holmes et al., 1994).

The Structure and Mobility Mechanism of Retrotransposition-Competent LINE-1s

RC-L1s are ~6 kb in length and contain a 5' untranslated region (UTR), two ORFs, and a 3' UTR that is punctuated by a poly(A) tail (Scott et al., 1987) (Figure 1.1). The L1 5' UTR houses an internal RNA polymerase II promoter that directs transcription from the 5' end of the element (Swergold, 1990); it also contains *cis*-acting binding sites for multiple transcription factors (Athanikar et al., 2004; Becker et al., 1993; Kuwabara et al., 2009; Minakami et al., 1992; Tchenio et al., 2000; Yang et al., 2003).

Recent studies have demonstrated that the L1 5' UTR contains a potent anti-sense promoter (L1 ASP). Transcription from the L1 ASP can lead to chimeric transcripts that contain a portion of the L1 5' UTR and genomic sequences flanking the 5' end of the L1 (Nigumann et al., 2002; Speek, 2001). It is further speculated that these chimeric transcripts may either function in gene regulation or promote the formation of double-stranded L1 RNAs that regulate L1 retrotransposition by RNA interference-based mechanisms (Matlik et al., 2006; Yang and Kazazian, 2006). Interestingly, L1 ASP-derived chimeric transcripts have proven useful in the identification of expressed L1s from human embryonic stem cells, embryonic carcinoma cell lines, and somatic human tissues (Faulkner et al., 2009; Macia et al., 2011; Speek, 2001).

Human ORF1 encodes a ~40-kDa protein (ORF1p) required for L1 retrotransposition (Holmes et al., 1992; Moran et al., 1996). ORF1p has an amino-terminal coiled-coil domain (Holmes et al., 1992), a centrally located RNA recognition motif, and a basic carboxyl-terminal domain (Khazina and

Weichenrieder, 2009; Moran et al., 1996). Biochemical experiments with mouse and human ORF1p demonstrate that the amino-terminal coiled-coil domain facilitates ORF1p trimer formation (Khazina and Weichenrieder, 2009; Martin et al., 2003). Structural and biochemical analyses also suggest that the RNA recognition motif and carboxyl-terminal domain play critical roles in ORF1p binding to nucleic acids (Basame et al., 2006; Januszyk et al., 2007; Khazina and Weichenrieder, 2009; Kolosha and Martin, 1997). Finally, ORF1p has nucleic acid chaperone activity, which may be important for L1 integration (Martin and Bushman, 2001).

ORF2 encodes a ~150-kDa protein (ORF2p), which has endonuclease and reverse transcriptase activities that are critical for L1 retrotransposition (Ergun et al., 2004; Feng et al., 1996; Martin et al., 1995; Mathias et al., 1991; Moran et al., 1996). ORF2p also contains a cysteine-rich domain of unknown function near its carboxyl terminus that is required for retrotransposition (Fanning and Singer, 1987; Moran et al., 1996). Recent experiments suggest that human ORF2 translation occurs by an unconventional termination-reinitiation mechanism, and that the putative ORF2 AUG initiation codon is dispensable for translation, although it is conserved in all human L1s examined thus far (Alisch et al., 2006; Dmitriev et al., 2007; McMillan and Singer, 1993).

The analysis of mutagenic L1 insertions in conjunction with biochemical and genetic assays has demonstrated that human ORF1p and ORF2p preferentially associate with their encoding messenger RNA (mRNA) (Dombroski et al., 1991; Esnault et al., 2000; Kulpa and Moran, 2006; Wei et al., 2001). This association,

termed *cis*-preference, leads to the generation of an L1 ribonucleoprotein (RNP) particle (Hohjoh and Singer, 1996; Kulpa and Moran, 2005; Martin, 1991). ORF1p is detectable in cytoplasmic RNPs from embryonic stem cells, embryonic carcinoma cell lines, and HeLa cells that overexpress engineered L1 elements (Garcia-Perez et al., 2007b, 2010; Hohjoh and Singer, 1996; Kulpa and Moran, 2005; Martin, 1991). ORF1p and/or L1 RNA also have been detected in human oocytes, in some human somatic cells, and at select times during human and mouse germ cell development (Belancio et al., 2010b; Branciforte and Martin, 1994; Coufal et al., 2009; Georgiou et al., 2009; Trelogan and Martin, 1995). In contrast, ORF2p appears to be much less abundant than ORF1p in L1 RNPs, and its detection has relied largely upon an assay to detect L1 reverse transcriptase activity in RNPs (Kulpa and Moran, 2006). However, epitope-tagging strategies recently have allowed reliable ORF2p detection in RNPs derived from cells transfected with L1 expression constructs by both Western blotting and immunofluorescence (Doucet et al., 2010; Goodier et al., 2010). L1 RNPs also associate with stress granules, although determining whether this association is important for L1 retrotransposition requires further study (Doucet et al., 2010; Goodier et al., 2007).

Experiments using adenovirus-based vectors suggest that L1 retrotransposition can occur in the absence of cell division (Kubo et al., 2006). Thus, some components of L1 RNPs may enter the nucleus in the absence of nuclear envelope breakdown. L1 retrotransposition then likely occurs by target-site primed reverse transcription (TPRT) (Cost et al., 2002; Feng et al., 1996;

Luan et al., 1993). During TPRT, the L1-encoded endonuclease generates a single-strand endonucleolytic nick in genomic DNA to expose a 3'-OH (Feng et al., 1996). The liberated 3'-OH is then used as a primer by the L1 reverse transcriptase to initiate cDNA synthesis using the L1 mRNA as a template (Cost et al., 2002; Feng et al., 1996; Kulpa et al., 2006). Molecular details regarding second-strand target-site cleavage and second-strand L1 cDNA synthesis require elucidation, but insights about both these steps in TPRT have arisen from studying a related retrotransposon, R2, from *Bombyx mori* (Christensen and Eickbush, 2005). The process of TPRT generates a new L1 copy that generally is flanked by ~7-20 base pair (bp) target-site duplications (TSDs) (Kazazian and Moran, 1998) (Figure 1.2).

Nonautonomous Retrotransposons: Alu and SVA Elements

The human genome also contains numerous nonautonomous retrotransposons that rely upon activity of L1-encoded proteins to mediate their mobility. These non-autonomous elements consist primarily of the small interspersed element (SINE), Alu, which accounts for ~10% of sequence in the HGR (Lander et al., 2001).

Alu elements arose in mammalian genomes ~65 million years ago and contain two monomeric sequences derived from the signal recognition particle (SRP) 7SL RNA (Batzer and Deininger, 2002; Ullu et al., 1982). Active Alu elements are ~280 bp in length and end in an A-rich tail. The left monomer contains an internal RNA polymerase III promoter (Batzer and Deininger, 2002) and is separated from the right monomer by an adenosine-rich sequence (Figure

1.1). However, flanking genomic sequences also can influence Alu transcriptional initiation and termination (Chu et al., 1995; Comeaux et al., 2009; Dewannieux and Heidmann, 2005; Goodier and Maraia, 1998; Liu and Schmid, 1993; Ullu and Weiner, 1985). Like L1s, Alu elements can be stratified into subfamilies (Deininger et al., 1992). Recent computational analyses and studies in cultured cells suggest that there may be thousands of active Alu "core" elements in the HGR (Bennett et al., 2008; Cordaux et al., 2004). However, Alu Y elements, most notably Ya5 and Yb8 subfamily members, account for the vast majority of disease-producing insertions in humans (Carroll et al., 2001).

Alu elements rely upon L1 ORF2p to facilitate their retrotransposition in *trans* (Dewannieux et al., 2003). Mutations that either block Alu transcription or interfere with SRP9/14 protein binding adversely affect Alu retrotransposition *in vitro* (Bennett et al., 2008). Indeed, the ability of the SRP9/14 proteins to interact with Alu RNA may be intimately tied to the evolution of active Alu subfamilies (Bennett et al., 2008; Sarrowa et al., 1997).

Alu elements can also be co-opted to play roles in gene expression. For example, inverted Alu elements within the 3' UTRs of some cellular mRNAs can result in adenosine-to-inosine RNA editing and preferential nuclear retention of the resultant edited mRNAs in the nuclei of differentiated cells (Chen and Carmichael, 2009). Additionally, Alu elements can be incorporated into existing transcription units by a process known as exonization, which may serve to enhance transcriptome diversification (Lev-Maor et al., 2003; Sela et al., 2010; Shen et al., 2011). Finally, the poly(A) tails flanking Alu elements may serve as a

source for generating microsatellite sequences in human DNA (Arcot et al., 1995).

SINE-R/VNTR/Alu (SVA) elements arose in primate lineages ~25 million years ago and are present at ~2,700 copies in the human genome (Wang et al., 2005). They have a composite structure consisting of a variable-length hexameric repeat $(CCCTCT)_n$ that is sequentially followed by an inverted Alu-like sequence, a variable number of tandem repeats (VNTRs) region, a sequence derived from the 3' end of a HERV-K10 element (SINE-R), and a poly(A) tail (Ono et al., 1987; Ostertag et al., 2003; Shen et al., 1994) (Figure 1.1). Whether SVA elements contain functional promoter sequences remains an open question; however, it is likely that SVA elements are transcribed by RNA polymerase II and that the resultant SVA RNAs are mobilized to new genomic locations by the L1-encoded proteins (Damert et al., 2009; Hancks et al., 2009; Hancks et al., 2011; Ostertag et al., 2003) (Figure 1.2). Indeed, the recent development of a cell-based assay for SVA mobilization should be instrumental in deciphering mechanistic details of SVA *trans*-mobilization (Hancks et al., 2011).

Nonautonomous Retrotransposons: Mobilization of Cellular RNAs

Cellular messenger RNAs can occasionally use the L1-encoded proteins to mobilize to new genomic locations, thereby generating processed pseudogenes (Esnault et al., 2000; Maestre et al., 1995; Wei et al., 2001) (Figure 1.1). There are ~8,000-15,000 processed pseudogene copies in the HGR, and most are derived from genes that are highly expressed in the germline, such as housekeeping genes and ribosomal protein genes (Torrents et al., 2003; Zhang

et al., 2002, 2003). Interestingly, some ribosomal protein processed pseudogenes (*e.g.,* RPL21) occur at a relatively high copy number, suggesting that some property of these mRNAs allows them to recruit the L1-encoded proteins more effectively than other mRNAs (Zhang et al., 2002).

Most processed pseudogenes are dead on arrival because they lack a functional promoter (Vanin, 1985; Weiner et al., 1986). Thus, they can be used as molecular clocks to estimate mutational rates between species (Graur et al., 1989). However, some human processed pseudogenes are expressed, and a small number may encode functional genes or serve as sources of small interfering RNAs (siRNAs) with gene regulatory functions (Harrison et al., 2005; Tam et al., 2008; Vanin, 1985).

To date, there are no examples of *de novo* processed pseudogene retrotransposition events causing human disease. However, the expression of a processed pseudogene has been implicated in human facioscapulohumeral dystrophy patients (Lemmers et al., 2010; Snider et al., 2010). Similarly, expression of an *FGF4* processed pseudogene is associated with chondrodysplasia in 19 dog breeds, consistent with the idea that selective breeding can enrich for rare mutagenic L1-mediated insertion alleles (Parker et al., 2009).

The L1-encoded proteins also can mobilize other noncoding cellular RNAs, such as U6 small nuclear RNA (snRNA), to new genomic locations (Buzdin et al., 2002). Computational and experimental evidence suggests that the L1 reverse transcriptase can switch templates to the U6 snRNA during TPRT to generate

U6/L1, and less frequently U6/processed pseudogene chimeras (Buzdin et al., 2002, 2003a; Garcia-Perez et al., 2007a; Gilbert et al., 2005) (Figure 1.2). Interestingly, some U6 pseudogenes, small uracil-rich RNAs (*e.g.,* U1 snRNA), and small nucleolar RNAs (*e.g.,* U3 snoRNA) end in poly(A) tails and are flanked by target site duplications, suggesting that they also may have been mobilized by the L1-encoded proteins (Bennett et al., 2004; Denison et al., 1981; Garcia-Perez et al., 2007a; Van Arsdell et al., 1981; Weber, 2006).

Technologies to Identify Human-Specific LINE-1s

Overview

The ability to discriminate human-specific L1-mediated retrotransposition events from the large mass of defective retrotransposons in the genome is akin to finding a needle in a haystack. However, an elegant combination of phylogenetic, molecular biological, computational, and modern genomic technologies has revolutionized our ability to identify human-specific L1-mediated retrotransposition events in both reference sequences and individual genomes (Figure 1.3). Some of the seminal findings allowing these advances are discussed below.

A Brief Historical Perspective

L1s in the human genome have succeeded one another in a single lineage for the last ~40 million years (Boissinot et al., 2000; Boissinot et al., 2004; Boissinot and Furano, 2001; Khan et al., 2006). Thus, new L1 subfamilies are continuously replacing older ones to dominate the expanding lineage of active

elements (Boissinot et al., 2000; Boissinot and Furano, 2001; Deininger et al., 1992; Khan et al., 2006; Smit et al., 1995). A majority of L1s are shared between human and chimp genomes, and likely represent retrotransposition-defective molecular fossils. However, studies in human embryonic carcinoma cell lines revealed that a subset of expressed L1s contained diagnostic sequence variants within their 3' UTR (*e.g.,* an ACA instead of a GAG trinucleotide at positions 5930 to 5932 of L1.2; accession number M80343) (Skowronski et al., 1988). This subset of expressed L1s were designated the Ta (transcribed, subset a) L1 subfamily (Skowronski et al., 1988). Interestingly, all but one of the ~18 disease-producing L1 insertions identified to date are derived from the Ta subfamily (Goodier and Kazazian, 2008; Belancio et al., 2008). The remaining mutagenic insertion was derived from the slightly older, human-specific pre-Ta subfamily of L1 (which contains an ACG trinucleotide at positions 5930 to 5932 relative to L1.2) (Kazazian et al., 1988).

The identification of sequence variants peculiar to human-specific L1s and the subsequent development of a cultured cell retrotransposition assay were instrumental in allowing the identification of RC-L1s in the HGR (Figure 1.4) (Beck et al., 2010; Brouha et al., 2003; Lander et al., 2001; Moran et al., 1996; Myers et al., 2002). Comprehensive computational studies revealed that the HGR contains ~90 L1s with intact ORFs (Brouha et al., 2003; Myers et al., 2002). Polymerase chain reaction (PCR)-based cloning revealed that ~44 of these show a range of activity in cultured human HeLa cells (Brouha et al., 2003). Unexpectedly, six highly active, or “hot” L1s accounted for ~84% of the

retrotransposition activity in the HGR. Extrapolation of the ~44 RC-L1s contained in the haploid HGR working draft to the scale of a complete diploid genome suggested that the average human genome harbors ~80-100 active L1 elements (Brouha et al., 2003; Sassaman et al., 1997). Interestingly, limited genotyping analyses indicated that many active and retrotransposition-defective Ta subfamily L1s in the HGR are polymorphic with respect to presence or absence, indicating that many are recent insertions (Badge et al., 2003; Boissinot et al., 2004; Brouha et al., 2003; Myers et al., 2002).

Identification of Human-Specific LINE-1s by Mining Genome Sequences

Large-scale DNA sequencing projects have provided valuable resources to identify human L1-mediated retrotransposon polymorphisms. For example, comparative genomic analyses between the HGR and the draft chimpanzee genome allowed the identification of ~11,000 species-specific transposable elements, including 5,530 Alu, 1,174 L1, and 864 SVA elements specific to humans (Mills et al., 2006). Comparisons of DNA sequence trace files from 36 geographically diverse humans also enabled the identification of ~505 Alu elements, 65 L1s, 39 SVAs, 2 HERV-Ks, and 5 other polymorphic insertions (Bennett et al., 2004). Similarly, comparison of a complete human diploid sequence (Levy et al., 2007) to the HGR allowed the identification of ~706 mobile element-associated structural variants, including the insertion of 584 Alu elements, 52 L1s, and 14 SVA elements (Xing et al., 2009). Mining 8 human genome sequences generated by next-generation sequencing yielded 4,342 Alu insertions absent from the HGR, 3,432 of which were additionally absent from a

number of previous studies (Hormozdiari et al., 2011). Finally, a pilot analysis of sequence data obtained from ~200 – 300 individuals as part of the 1000 Genomes Project allowed the identification of over 1,000 L1 insertions that are absent from the HGR (Durbin et al., 2010; Ewing and Kazazian, 2011; Mills et al., 2011; Rouchka et al., 2010). However, due to the low fold sequence coverage (2-4 fold for many genomes) and the composite nature of the HGR, it is likely that L1s are underrepresented in these analyses.

Clearly, the exploitation of human DNA sequence resources and nonhuman primate comparative genomics has revealed extensive human-specific transposable element diversity. A union of the above data has shown that these insertions are much more common in the population than once thought**.**

Experimental Approaches to Identify Novel Retrotransposon Insertions

The genomics revolution has revealed how structural variation contributes to inter-individual genetic diversity. For example, representative oligonucleotide microarray analysis and array comparative genomic hybridization technologies have allowed the high-throughput identification of submicroscopic genetic differences of ~100kb among individuals (Iafrate et al., 2004; Sebat et al., 2004). However, the size resolution of those techniques was not sufficient to permit identification of L1-mediated retrotransposition events.

More recently, a number of methods have been developed that are capable of detecting small- and intermediate-scale human structural variants (Figure 1.3). These methods include transposon display, array-based hybridization, second-

generation DNA sequencing, and paired-end mapping of clone libraries (Badge et al., 2003; Bentley et al., 2008; Conrad et al., 2006; Durbin et al., 2010; Kidd et al., 2008; Korbel et al., 2007; McKernan et al., 2009; Mills et al., 2011; Ovchinnikov et al., 2001, 2002; Redon et al., 2006; Sheen et al., 2000; Tuzun et al., 2005; Wang et al., 2008; Wheeler et al., 2008). Together, these approaches have been instrumental in revealing transposon diversity in the genomes of geographically diverse individuals. A somewhat unexpected result is that mobile element dimorphisms account for a relatively large proportion of human genetic diversity (up to ~20%–25% of the interindividual genetic differences identified in some studies) (Kidd et al., 2010; Korbel et al., 2007). The development of approaches to specifically identify L1-mediated genetic variation is described below.

PCR-based display methods have allowed the identification of polymorphic L1 and Alu insertions in human genomes (Badge et al., 2003; Boissinot et al., 2004; Buzdin et al., 2003b; Ovchinnikov et al., 2001; Roy et al., 1999; Sheen et al., 2000). These methods exploit DNA sequence variants peculiar to human-specific retrotransposons (*e.g.,* the ACA character present in Ta subfamily L1s) in conjunction with a pool of short, arbitrary oligonucleotides or sequences complementary to ligated linkers. PCR reactions utilizing this combination of retrotransposon-specific and degenerate or linker sequences then are used to generate complex amplicon libraries that contain the candidate retrotransposon and its immediate 5' or 3' flanking sequences. Sequencing of the flanking DNA and subsequent searches for sequence similarity then determines whether the

candidate retrotransposon is present or absent in the HGR (Badge et al., 2003; Boissinot et al., 2004). Together, these methods have identified numerous L1 and Alu insertion polymorphisms in individual genomes.

A derivation of classical display methods that employs suppression PCR methodology [amplification typing of L1 active subfamilies (ATLAS)] enabled the identification of nine full-length human-specific L1s from individual genomes (Badge et al., 2003). Interestingly, three out of seven tested sequences were "hot" in the cultured cell retrotransposition assay, and these three L1s were present at a minor allele frequency of less than 24% when genotyped in a panel of 90 geographically diverse individuals. Thus, these data provided additional evidence to support the hypothesis that young, human-specific L1s are underrepresented in the HGR.

Deconvolution of complex amplicon libraries generated by retrotransposon display approaches traditionally relied on electrophoretic fractionation using agarose or acrylamide gel systems and the subsequent cloning and characterization of individual molecules to map insertions to the HGR. The advent of new genomic technologies, including high-throughput DNA sequencing, now offers a means to revolutionize the discovery of polymorphic retrotransposon insertions (see Figure 1.3).

One method to identify dimorphic retrotransposon insertions, transposon insertion profiling by microarray (TIP-chip), employs the principles of transposon display to specifically amplify retrotransposons and their associated flanking sequences (Gabriel et al., 2006; Huang et al., 2010; Wheelan et al., 2006). The

resultant amplicons are hybridized back to oligonucleotide arrays to identify sequences flanking the retrotransposons. TIP-chip allowed the discovery of numerous L1, Alu, and HERV-K insertion polymorphisms in genome-wide analyses. Application of TIP-chip to 69 unrelated individuals with X-linked intellectual disabilities also allowed the identification of L1 insertions within introns of the *NHS* and *DACH2* genes, and mutations in these genes are implicated in intellectual disability (Huang et al., 2010). However, future studies are needed to determine whether the L1 insertions play a causal role in the observed cognitive phenotypes in these patients.

Another method to identify dimorphic L1 and Alu insertion polymorphisms employs traditional Sanger capillary sequencing and 454 or Illumina-based second-generation DNA sequencing technologies to de-convolute transposon-derived amplicon libraries (Ewing and Kazazian, 2010; Iskow et al., 2010). Iskow *et al.* identified 152 L1s from 38 ethnically diverse humans and 8 cell lines using capillary sequencing. They also identified 650 L1 and 403 Alu insertions in 30 lung or brain tumors and their matched non-tumor controls via 454-based sequencing (Iskow et al., 2010). These insertions were absent from both the HGR and a database of retrotransposon insertion polymorphisms (dbRIP) (Lander et al., 2001; Wang et al., 2006), yielding 1,145 novel transposable elements. Ewing *et al.* used an Illumina-based sequencing approach to identify 367 polymorphic L1s from a cohort of 25 humans that contained 15 unrelated individuals, individuals from 6 trios, and 3 pairs of monozygotic twins (Ewing and Kazazian, 2010). Finally, Witherspoon *et al.* used an Illumina-based sequencing

approach to identify 487 Alu Yb8 and Yb9 insertions in 4 unrelated individuals that were absent from the HGR (Witherspoon et al., 2010).

A third method to identify full-length or near-full-length L1 insertion polymorphisms involves paired-end DNA sequencing of fosmid libraries derived from individuals belonging to geographically diverse populations, which successfully identified intermediate-sized structural variants in human DNA (Beck et al., 2010; Kidd et al., 2008; Tuzun et al., 2005). The screening of six individual libraries identified 68 L1s that were absent from the HGR. Remarkably, 37 of these 68 were "hot" L1s, two of which belonged to the pre-Ta L1 subfamily (Beck et al., 2010). Notably, unlike the methodologies described above, paired-end DNA sequencing of fosmid libraries does not involve PCR, allows identification of L1s in repetitive regions of the genome, and though labor intensive, allows a comprehensive and relatively unbiased snapshot of L1 diversity. Together, the above methodologies have uncovered a virtual treasure trove of natural retrotransposon diversity in human genomes.

Allelic Heterogeneity in LINE-1 Activity

In addition to the polymorphic status of an L1, allelic heterogeneity may cause different L1 insertion alleles to exhibit different retrotransposition efficiencies in cultured cells. For example, L1.2A and L1.2B, which are likely progenitor alleles of a mutagenic insertion in the *Factor VIII* gene, exhibit a ~16-fold difference in their retrotransposition efficiencies because of amino acid substitutions near the ORF2p carboxyl terminus (Dombroski et al., 1991; Farley et al., 2004; Kazazian et al., 1988; Lutz et al., 2003). Similarly, the examination of

three “hot” L1s from the HGR in a geographically diverse set of individuals revealed alleles with a wide range of retrotransposition activities (ranging from 0% to 390% of a reference L1) (Seleme et al., 2006). Finally, a recent study identified an RC-L1 allele in an individual genome which apparently is defective in the HGR due to a stop codon in ORF2p (Beck et al., 2010). Thus, both presence/absence of polymorphisms and allelic heterogeneity can influence L1 retrotransposition in an individual genome.

Impact of Mobile Elements on Mammalian Genomes

LINE-1 as a Mutagen

A wealth of data is revealing the consequences of L1-mediated retrotransposition events in human genomes. Since their original discovery, approximately 65 disease-causing mutations in man have been attributed to L1-mediated retrotransposition events (Belancio et al., 2008a; Goodier and Kazazian, 2008). L1-mediated retrotransposition events can act as mutagens by directly disrupting exons (Kazazian et al., 1988). Similarly, insertions into introns can induce missplicing or exon skipping, thereby generating hypomorphic or null expression alleles (Figure 1.5) (Belancio et al., 2008b; reviewed in Goodier and Kazazian, 2008; Ostertag and Kazazian, 2001a). Finally, recent reports suggest the L1 endonuclease may cause double strand breaks, which in principle could lead to genomic instability (Gasior et al., 2006; Lin et al., 2009). Clearly, advances in genomics and the evolution of DNA sequencing technologies should allow the rapid identification of other disease-producing insertions in the coming years.

Effects on Gene Expression

L1 insertions can impact gene expression by a variety of mechanisms (Figure 1.5). For example, experimental studies have revealed that the adenosine-rich nature of the L1 transcript can introduce premature polyadenylation and/or RNA polymerase II transcriptional pause sites into genes, thereby attenuating their expression (Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003). Interestingly, transcriptional pausing appears to depend on both the length of the L1 insertion and whether it is in the same transcriptional orientation as its resident gene (Chen et al., 2006; Han et al., 2004). The L1 ASP also can generate transcripts that, in principle, could affect gene expression (Matlik et al., 2006; Nigumann et al., 2002; Speek, 2001; Yang and Kazazian, 2006). Finally, in rare instances, L1 insertions may disrupt genes, leading to the generation of distinct transcription units through a phenomenon known as gene breaking (Figure 1.5) (Wheelan et al., 2005).

Approximate Rates of Heritable Retrotransposition Events

Determining the rate of germline retrotransposition in the human population remains an area of ongoing investigation. Estimates suggest that Alu elements are the most active retrotransposons in the human genome, with new insertions occurring in approximately 1 out of 20 live births (Cordaux and Batzer, 2009; Xing et al., 2009). L1 insertions follow, with estimates ranging between 1 out of 20 and 1 out of 200 births, depending upon the method used in the analysis (Cordaux and Batzer, 2009; Ewing and Kazazian, 2010; Huang et al., 2010; Li et al., 2001; Xing et al., 2009). SVA insertions may be the least frequent

retrotransposition events, occurring in approximately 1 out of 900 births (Cordaux and Batzer, 2009; Xing et al., 2009). In spite of disparate rates of retrotransposition, the different sizes and unique characteristics of L1, Alu, and SVA elements may pose distinctive challenges to the genome.

The above estimates are subject to ascertainment biases, and may represent only minimal estimates of the actual *de novo* retrotransposition frequency. For example, because some studies relied on comparisons with the HGR, they may underestimate the contribution of low allele-frequency insertions to human genetic diversity (Xing et al., 2009). In fact, Iskow *et al.* suggested that low allele-frequency L1-mediated retrotransposition events, representing either rare or perhaps private L1 and/or Alu alleles, may be present in "virtually all personal genomes in the human populations" (Iskow et al., 2010). Consistently, fosmid-based paired-end DNA sequencing, TIP-chip profiling, and the preliminary examination of the 1000 Genomes Project data have readily allowed the detection of rare L1 retrotransposition events in geographically diverse individuals (Beck et al., 2010; Ewing and Kazazian, 2011; Huang et al., 2010; Mills et al., 2011). Clearly, the exhaustive DNA sequencing of parent/offspring trios at high coverage should provide more accurate estimates of L1-mediated germline retrotransposition events in the near future.

Somatic LINE-1 Retrotransposition: Insights from Cancer Studies

In addition to acting as a germline mutagen, studies have revealed that L1 retrotransposition also occurs in certain somatic cells. Historically, the identification of a mutagenic L1 insertion into the *adenomatous polyposis coli*

(*APC*) gene in a colorectal tumor that was absent from adjacent non-tumor tissue established that L1 could retrotranspose in somatic cells (Miki et al., 1992).

Recent data generated using L1 display approaches combined with 454-based DNA sequencing led to the discovery of 9 *de novo* somatic L1 retrotransposition events in 6 of 20 non small-cell lung cancers that were absent from matched adjacent normal tissue samples (Iskow et al., 2010). Interestingly, tumors containing new L1 retrotransposition events also exhibited global patterns of hypomethylation, providing a correlative link between epigenetic changes and increased L1 retrotransposition in tumors. Hypomethylation of the L1 5' UTR also has been observed in malignant cells and cancer tissues, and is correlated with an increase in L1 mRNA and/or ORF1p expression (Alves et al., 1996; Asch et al., 1996; Belancio et al., 2010a; Goodier and Kazazian, 2008). Similarly, 5-azacytidine treatment leads to an elevation of L1-ASP driven chimeric transcripts in non-malignant breast epithelial cells (Cruickshanks and Tufarelli, 2009). Thus, the above data suggest that the rate of L1 retrotransposition, and by proxy Alu and SVA retrotransposition, may be elevated in some cancers. Further studies should allow a greater understanding of whether and/or how often L1-mediated retrotransposition occurs in cancer and whether these events play a causal role in tumorigenesis.

Somatic LINE-1 Retrotransposition During Normal Development

Human genetic approaches, experiments conducted in transgenic animals, and cell culture models further suggest that L1 retrotransposition can occur in normal cells at discrete times during development. For example, genetic

analyses have conclusively demonstrated both germline and somatic mosaicism of a mutagenic L1 retrotransposition event in the mother of a male patient afflicted with X-linked choroideremia (van den Hurk et al., 2007). Thus, the mutagenic L1 insertion must have occurred during early embryonic development of the mother prior to partitioning of the germline. Consistent with this hypothesis, engineered L1s can retrotranspose in human embryonic stem cells (Garcia-Perez et al., 2007b).

Engineered human L1s can also retrotranspose during the early stages of mouse and rat embryonic development; however, most of the resultant insertions are not heritable (Kano et al., 2009). These findings are remarkable and suggest that L1 mRNA may be transferred from the gametes to the zygote to undergo retrotransposition at a later time in development, generating somatic mosaicism in the resultant offspring.

Unexpectedly, recent studies also revealed that engineered human L1s can retrotranspose at discrete times during neuronal development (Coufal et al., 2009; Muotri et al., 2005, 2010). These retrotransposition events could, in principle, generate somatic mosaicism in the nervous system and have the potential to affect intra-individual neuronal variability (Coufal et al., 2009; Muotri et al., 2005, 2010). In fact, sensitive TaqMan PCR-based approaches suggest that certain regions of the brain have an increase in human-specific L1 content relative to heart and/or liver tissues isolated from the same individual (Coufal et al., 2009; Muotri et al., 2010). The observed increase in L1 DNA copy number in the brain suggests elevated levels of endogenous L1 retrotransposition in brain.

However, other non mutually exclusive mechanisms must be considered when accounting for these L1 copy number differences (Peterson et al., 2008; Westra et al., 2010). Indeed, formal proof of endogenous L1 retrotransposition in brain will require the characterization of new L1 insertions from individual neurons. Although this remains a daunting task, advances in DNA sequencing technologies–including the ability to comprehensively characterize genome sequences from single or small pools of cells–should allow rigorous testing of these interesting observations.

Together, the above findings overturned the long-held dogma that L1-mediated retrotransposition could only occur in germ cells, and has led to speculations about how L1-meditated retrotransposition events may affect intra-individual genetic variation (Kano et al., 2009; Martin, 2009; Muotri et al., 2005, 2010; Singer et al., 2010). Time and rigorous experimentation will tell whether somatic L1 retrotransposition events represent stochastic genetic noise that is tolerated by the host genome or whether these events have functional consequences in disease pathogenesis and/or neuronal development.

LINE-1 as an Agent of Genome Diversification

LINE-1 Retrotransposition by Target-Site Primed Reverse Transcription

Canonical TPRT generally results in the insertion of an L1 or non-autonomous retrotransposition event at a new genomic location flanked by short TSDs (Figures 1.2 and 1.5). On occasion, TPRT also can lead to small deletions of target-site DNA that vary from approximately 2 to 50 bp in length, and/or to the addition of non-templated or filler nucleotides at the 5' genomic DNA/L1 junction

sequence (Athanikar et al., 2004; Lavie et al., 2004; Narita et al., 1993). Similar target-site alterations were observed upon examining engineered L1 retrotransposition events in transformed cell lines, human embryonic stem cells, and neuronal progenitor cells (Coufal et al., 2009; Garcia-Perez et al., 2007b; Gilbert et al., 2002, 2005; Muotri et al., 2005; Symer et al., 2002).

Approximately 35% of Ta subfamily L1 retrotransposition events represent full-length insertions (Boissinot et al., 2000). The remaining L1s are 5' truncated (Grimaldi et al., 1984) and often contain short microhomologies at the 5' genomic DNA/L1 junction sequence (Babushok et al., 2006; Gilbert et al., 2002, 2005; Martin et al., 2005; Symer et al., 2002; Zingler et al., 2005). Additionally, approximately 25% of L1s contain L1 inversion/deletion events that likely are generated by a process termed twin priming (Figure 1.5) (Ostertag and Kazazian, 2001b). The high frequency of 5' truncation associated with new L1 retrotransposition events remains enigmatic and may reflect host defense or DNA repair processes that act to either dissociate the L1 reverse transcriptase from the nascent L1 cDNA or degrade the L1 mRNA template prior to the completion of reverse transcription.

LINE-1 Retrotransposition-Mediated Deletion Events

In addition to acting as an insertional mutagen, L1-mediated retrotransposition events can lead to various forms of human structural variation. For example, studies conducted in HeLa and HCT116 cells revealed that ~10% of retrotransposition events derived from engineered human L1s are associated with the formation of chimeric L1s accompanied by intrachromosomal deletions,

intrachromosomal duplication or inversions, and perhaps interchromosomal translocations (Gilbert et al., 2002, 2005; Symer et al., 2002). Comparisons of the pre- and post-integration sites of chimeric retrotransposition events suggested that DNA recombination processes such as single-strand annealing, synthesis-dependent strand annealing, and perhaps nonhomologous end joining are involved in the formation of these aberrant structures (Gilbert et al., 2002, 2005; Symer et al., 2002).

L1 retrotransposition-mediated genomic deletions are not peculiar to studies conducted in transformed human cells. For example, comparative biological approaches between the human and chimpanzee reference genomes enabled the discovery of 30 L1 and 19 Alu retrotransposition-mediated deletion events, which together account for a loss of ~26 kb of human genomic DNA in the past six million years (Callinan et al., 2005; Han et al., 2005; Salem et al., 2003).

L1 retrotransposition-mediated deletion events have also been observed in human genetic diseases (Chen et al., 2005b). For example, a full-length L1 insertion that was accompanied by a deletion of ~46 kb, including 7 exons of the *PDHX* gene, led to a sporadic case of pyruvate dehydrogenase complex deficiency (Mine et al., 2007). Similarly, a ~4-kb L1 insertion was accompanied by a ~17-kb deletion that disrupted 3 exons of the *EYA1* gene, leading to a sporadic case of branchio-oto-renal syndrome (Morisada et al., 2010); however, the structure of this event suggests that it may have occurred by an endonuclease-independent (ENi) L1 retrotransposition mechanism (see below).

Alu and SVA retrotransposition-mediated deletion events also have impacted the human genome. For example, an Alu retrotransposition-mediated deletion approximately one million years ago resulted in the loss of a 92-bp exon of the *CMP-Neu5Ac hydroxylase* gene, leading to a loss-of-function frameshift mutation in humans (Hayakawa et al., 2001). Similarly, an SVA retrotransposition-mediated deletion apparently resulted in the loss of a 14-kb region of genomic DNA encompassing the entire *HLA-A* gene in three Japanese families containing patients afflicted with leukemia (Takasu et al., 2007). Thus, although relatively rare, retrotransposition-mediated deletion continues to impact the human genome and represents a mechanism that can lead to both human structural variation and disease.

Postintegration LINE-1- and Alu-Mediated Recombination Processes

The abundance of L1 and Alu retrotransposons in the human genome provides numerous potential substrates for post-integration recombination events that can lead to disease and/or structural variation in human genomes (reviewed in Cordaux and Batzer, 2009; Goodier and Kazazian, 2008). For example, nonallelic homologous recombination (NAHR) events between two inverted Alu elements originally was observed in a ~5 kb deletion that included exons of the *Low-Density Lipoprotein Receptor* gene, resulting in a case of familial hypercholesterolemia (Lehrman et al., 1985). Similarly, NAHR between genomic L1s has been implicated in sporadic cases of phosphorylase kinase deficiency, Alport syndrome, and Ellis-van Creveld syndrome (Burwinkel and Kilimann, 1998; Segal et al., 1999; Temtamy et al., 2008). Finally, comparative biological

and genomic studies have revealed that NAHR events between L1s or Alus have generated structural variants in the chimpanzee and human genomes (Han et al., 2007, 2008; Kidd et al., 2010; Lupski and Stankiewicz, 2005; Sen et al., 2006). Notably, young Alu elements are significantly enriched in regions flanking segmental duplications, suggesting that Alu-mediated recombination events may be involved in some segmental duplication expansions observed in humans (Bailey et al., 2003). Personal genomics and advances in DNA sequencing likely will continue to reveal that inter-retrotransposon recombination events represent a significant portion of structural variation in human genomes.

Endonuclease-Independent (ENi) Insertions

Experiments in XRCC4-deficient and DNA protein kinase catalytic subunit (DNA-PKcs)-deficient Chinese hamster ovary cell lines, which are defective in the nonhomologous end-joining pathway of DNA repair, revealed an alternative mechanism of ENi L1 retrotransposition distinct from conventional TPRT (Morrish et al., 2007; Morrish et al., 2002). Characterization of ENi insertions revealed that they generally integrated at non-canonical L1 endonuclease cleavage sites, lacked TSDs, and frequently were truncated at both their 5' and 3' ends. In addition, some ENi retrotransposition events were associated with genomic DNA deletions and/or contained cDNA fragments derived from non-L1 cellular RNAs that likely were reverse transcribed during the integration process. L1 insertions bearing the hallmarks of ENi retrotransposition events have been identified in the human genome, again showing how cultured cell models can predict events that occur in nature (Sen et al., 2007; Srikanta et al., 2009). Together, these findings

suggested that ENi retrotransposition might bypass the requirement for endonuclease activity by initiating reverse transcription at endogenous lesions in genomic DNA, thereby acting as a molecular bandage (Morrish et al., 2002; Voliva et al., 1984). Consistent with the above hypothesis, some ENi retrotransposition events in DNA-PKcs-deficient Chinese hamster ovary cells integrate at dysfunctional telomeres (Morrish et al., 2007). Thus, ENi retrotransposition shows curious similarities to the action of telomerase and may represent a form of RNA-mediated DNA repair utilized by retrotransposons that lack an endonuclease domain (Curcio and Belfort, 2007; Gladyshev and Arkhipova, 2007; Morrish et al., 2007).

LINE-1-Mediated Transduction

The examination of disease-producing L1 insertions and experiments using the cultured cell retrotransposition assay revealed that L1s are able to mobilize genomic DNA sequences flanking their 3' (and less commonly their 5') ends by a process termed L1-mediated transduction (Figure 1.5). The transduction of 3' sequences is relatively common, and suggests that RNA polymerase II frequently bypasses the natural L1 polyadenylation site and instead uses a site in flanking genomic DNA (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996). Computational analyses and the examination of L1 structural variants in individual genomes have revealed that 3' transductions flank ~20% of human-specific L1s in the HGR and that some 3' transductions are over 1 kb in length (Beck et al., 2010; Goodier et al., 2000; Kidd et al., 2010; Pickeral et al., 2000). Extrapolations based on these data suggest that L1-mediated 3' transductions

may comprise as much as ~19-30 Mb of the human genome (Goodier et al., 2000; Pickeral et al., 2000).

L1 3' transductions create recognizable sequence tags, which can be used to infer parent/offspring relationships between full-length L1s and their progeny. For example, a ~489 bp region of genomic DNA flanking a mutagenic L1 insertion into the *dystrophin* gene was instrumental in identifying an active L1 progenitor allele (Holmes et al., 1994). Similarly, shared 3' transduction sequences were recently used to identify mini families of L1s that contain active, rare alleles in the human population (Beck et al., 2010; Kidd et al., 2010). L1s from two of these 3' transduction families, LRE3 and $L1_{RP}$, are responsible for disease-producing insertions in sporadic cases of chronic granulomatous disease and X-linked retinitis pigmentosa, respectively (Brouha et al., 2002; Kimberland et al., 1999; Schwahn et al., 1998). Thus, the examination of transduction families that include highly active elements may be used as a way to discover novel, rare, and highly active L1s in human populations.

The cultured cell L1 retrotransposition assay revealed that 3' transduction could, in principle, lead to the mobilization of exons and/or regulatory DNA sequences to new genomic locations, and may represent a mechanism of exon shuffling (Moran et al., 1996, 1999). To date, there are no *in vivo* examples of L1 3' transduction leading to the formation of a new gene. However, SVA elements are also frequently associated with 3' transductions (Ostertag et al., 2003), and SVA-mediated 3' transduction events prior to the divergence of humans and great apes led to the dispersion of the *AMAC* gene to three distinct places in the

human genome (Xing et al., 2006). Whether these copies of the *AMAC* gene are functional requires validation, but the potential for genomic diversification through mobile element-mediated dispersion is clear.

The ability of the L1-encoded proteins to mobilize non-L1 RNAs to new genomic locations in *trans* provides another potential mechanism for exon dispersal and/or the creation of new genes. For example, retrotransposition of a *cyclophilin A* cDNA into the *TRIM5-*α locus of New World monkeys led to the creation of a chimeric protein that confers HIV resistance to owl monkeys (Sayah et al., 2004). Similarly, a novel testes-specific hominoid gene, *PIPSL*, is a chimeric ubiquitin-binding protein derived from a fusion of the *PIP5K1A* and 26S proteasome subunit RNAs that underwent retrotransposition (Babushok et al., 2007; Ohshima and Igarashi, 2010). The *PIPSL* gene appears to have been subject to positive selection and is conserved among humans, suggesting a functional role in modern genomes. Finally, it appears that exons from the *CFTR* and *ATM* genes have been mobilized in *trans* to new genomic locations by the L1 retrotransposition machinery (Ejima and Yang, 2003; Rozmahel et al., 1997). However, it remains possible that some of these examples represent L1-mediated 3' transductions that were so severely 5' truncated they lack L1 sequences (Ejima and Yang, 2003; Moran et al., 1999).

In principle, L1 5' transduction can occur if a transcript initiating upstream of an L1 undergoes retrotransposition. Because 5' transduction can only be found by examining full-length L1s, they appear to be much less common than 3' transductions. However, potential examples of L1 5' transduction have been

identified in the HGR, in cultured cell experiments, and in a mutagenic mouse insertion (Chen et al., 2006; Lander et al., 2001; Symer et al., 2002; Wei et al., 2001). In contrast, 5' transduction events frequently accompany SVA retrotransposition events, which is consistent with the idea that some SVA elements rely on host promoter sequences for their transcription (Damert et al., 2009; Hancks et al., 2009). Indeed, a 5' transduction derived from the *MAST2* gene was used to identify a family of SVA elements (SVA_F) that likely are amplifying in the human genome (Damert et al., 2009; Hancks et al., 2009).

Epigenetic Phenomena Related to LINE-1s

L1s may also play critical roles in the epigenetic regulation of host genes (Figure 1.5). For example, recent research suggests that indicator cassettes delivered into the genome by engineered L1 retrotransposons can be epigenetically silenced either during or immediately after their integration (Coufal et al., 2009; Garcia-Perez et al., 2010; Muotri et al., 2005). It will be interesting to determine whether silencing is specific for the retrotransposed L1 sequence or whether it can affect the epigenetic status of adjacent genes.

The ability of L1 to affect the epigenetic regulation of genes may not be restricted to new retrotransposition events. For example, the ability of L1s to accumulate on sex chromosomes over evolutionary time led Mary Lyon to speculate that L1s might function as *cis*-acting booster elements to aid in the spreading of heterochromatin formation observed during X-inactivation (Lyon, 1998). Consistent with this notion, some genes that escape X-inactivation are in relatively L1-poor regions of the X chromosome (Bailey et al., 2000; Carrel et al.,

2006), and L1 density appears to correlate positively with heterochromatin spread in X/autosomal translocations (Tang et al., 2010). Indeed, recent experiments suggest that L1s might play an active role in nucleating heterochromatin formation on the inactive X chromosome (Chow et al., 2010). Clearly, although these findings are still in their early stages, further research should allow even more discoveries about how L1-associated epigenetic changes impact gene expression.

Closing Remarks

It is now clear that transposable elements are an integral part of our genomes. In the coming years, technological breakthroughs in DNA sequencing and genome annotation will undoubtedly uncover many more examples of how transposable elements impact biological processes. For example, high-coverage, longer-read sequencing of trios and monozygotic twins, and/or the ability to sequence the genomes of single or small populations of cells will yield more information about the actual rate of L1-mediated retrotransposition events in both germline and somatic cells, and should clarify the role of L1-mediated retrotransposition events in certain types of human cancers. It will be interesting to determine whether genetically distinct populations or particular individuals are more prone to L1-mediated retrotransposition events.

Because transposable elements can be considered intracellular genomic parasites, they also provide an ideal means to study host-parasite interactions. The availability of new experimental reagents, such as epitope-tagged engineered human L1s, combined with cell-culture based retrotransposition

assays now allows a way to identify host factors that act to combat the genomic effects of continued retrotransposition. We probably will discover even more examples where genetic differences in host factors correlate with variation in L1 retrotransposition rates (*e.g.,* (Chiu and Greene, 2008; Goodier and Kazazian, 2008; Kidd et al., 2007; OhAinle et al., 2008; Stetson et al., 2008)).

Finally, future studies will enlighten us about how transposable element-derived sequences serve as seeds for evolutionary change. There is an ever-growing number of examples of regulatory elements derived from transposable elements that are required for proper gene expression (*e.g.,* (Bejerano et al., 2006; Jurka, 2008; Sasaki et al., 2008; Wang et al., 2007), and reviewed in (Feschotte, 2008)), and we anticipate this list will expand. Indeed, transposable elements may provide a mechanism to amplify and distribute *cis*-acting DNA sequences to new genomic locations, which after selection may function in gene regulation — a potentially prescient hypothesis put forth by Davidson & Britten over 40 years ago (Davidson and Britten, 1979).

We have come a long way since Barbara McClintock's discovery of mobile genetic elements in maize over half a century ago (McClintock, 1950). In the past decade alone, we have witnessed an exponential growth in knowledge about how transposable elements have impacted the human as well as other mammalian genomes. Though much progress has been made, our work has just begun. The coming years likely will identify how transposable elements contribute to phenotypic variation and human-specific traits. Clearly, the "junk" in our genomes is stepping into the limelight. It is an exciting time for transposable

element research; the human genetics community should take greater notice of these dynamic elements.

Acknowledgments

This Chapter was previously published: Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 Elements in Structural Variation and Disease. Annu Rev Genomics Hum Genet *12*, 187-215, and is reproduced with the permission of the publisher, Annual Reviews. I thank all of the additional authors for their work on this review. We thank Drs. Haig Kazazian and Mark Batzer for their critical and insightful review of the manuscript. We thank Aurélien Doucet, Billy Giblin, Nancy Leff, John Moldovan, Sandra Richardson, and other members of the Moran lab for helpful comments and discussions. C.R.B. was supported in part by NIH training grants T32GM7544 and T32000040. J.V.M. is supported by NIH grants GM060518 and GM082970 and is also an investigator of the Howard Hughes Medical Institute. J.L.G.-P. is supported by an ISCIII-CSJA (FEDER/EMER07/056), a Marie Curie IRG action (FP7-PEOPLE-2007-4-3-IRG), CICE (P09-CTS-4980), Proyectos en Salud PI-002 from Junta de Andalucía (Spain), and the Spanish Ministry of Health (FEDER/FIS PI08171). R.M.B. was supported by a Wellcome Trust project grant (075163/Z/04/Z) to R.M.B. and Prof. Sir Alec Jeffreys, FRS. Finally, we thank our spouses for their support and patience while writing this review.

Figure 1.1: Mobile Elements in Human Genomes

The classes of mobile genetic elements in the human genome, showing the type of mobile element, the structure of representative elements, the percentage of each element in the human genome reference sequence (HGR), and whether each class of elements is currently active (Lander et al., 2001). Abbreviations for human endogenous retrovirus-K (HERV-K): LTR, long terminal repeat; Gag, group-specific antigen; Pol, polymerase; Env, envelope protein (dysfunctional). For LINE-1: UTR, untranslated region; CC, coiled coil; RRM, RNA recognition motif; CTD, carboxyl-terminal domain; EN, endonuclease; RT, reverse transcriptase; C, cysteine-rich domain. For Alu: A and B, component sequences of the RNA polymerase III promoter; AR, the adenosine-rich segment separating the 7SL monomers. For SINE-R/VNTR/Alu (SVA): VNTR, variable number of tandem repeats; SINE-R, domain derived from a HERV-K. A_n signifies a poly(A) tail.

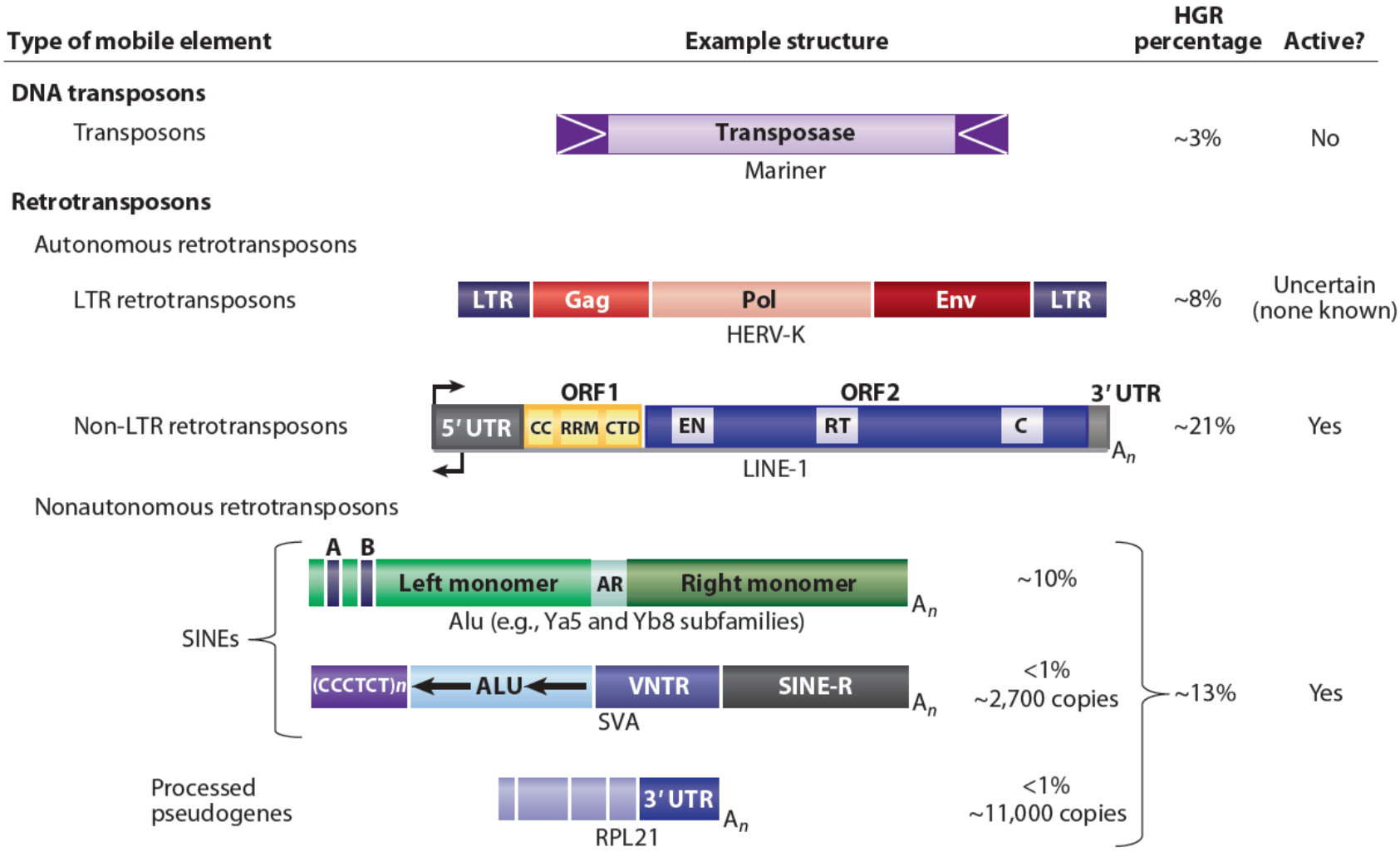


Figure 1.1: Mobile Elements in Human Genomes

Figure 1.2: A LINE-1 Retrotransposition Cycle

A LINE-1 retrotransposition cycle. A full-length L1 (*light blue bar* on *gray chromosome*) is transcribed, the L1 messenger RNA (mRNA) is exported to the cytoplasm, and translation of ORF1p (*yellow circles*) and ORF2p (*blue oval*) leads to ribonucleoprotein (RNP) formation. Components of the L1 RNP are transported to the nucleus, and retrotransposition occurs by target-site primed reverse transcription (TPRT). During TPRT, the L1 endonuclease (EN) nicks genomic DNA, exposing a free 3′-OH that can serve as a primer for reverse transcription of the L1 RNA. The processes of second-strand cleavage, second-strand complementary DNA (cDNA) synthesis, and completion of L1 integration require elucidation. TPRT results in the insertion of a new, often 5′-truncated L1 copy at a new genomic location (*gray bar* on *purple chromosome*) that generally is flanked by target-site duplications (*red arrows*). Alu, SINE-R/VNTR/Alu (SVA), and cellular mRNAs may hijack the L1-encoded protein(s) in the cytoplasm to mediate their *trans* mobilization. U6 small nuclear RNA (snRNA) may be integrated with L1 during TPRT. Question marks denote steps in the retrotransposition pathway of unknown mechanism.

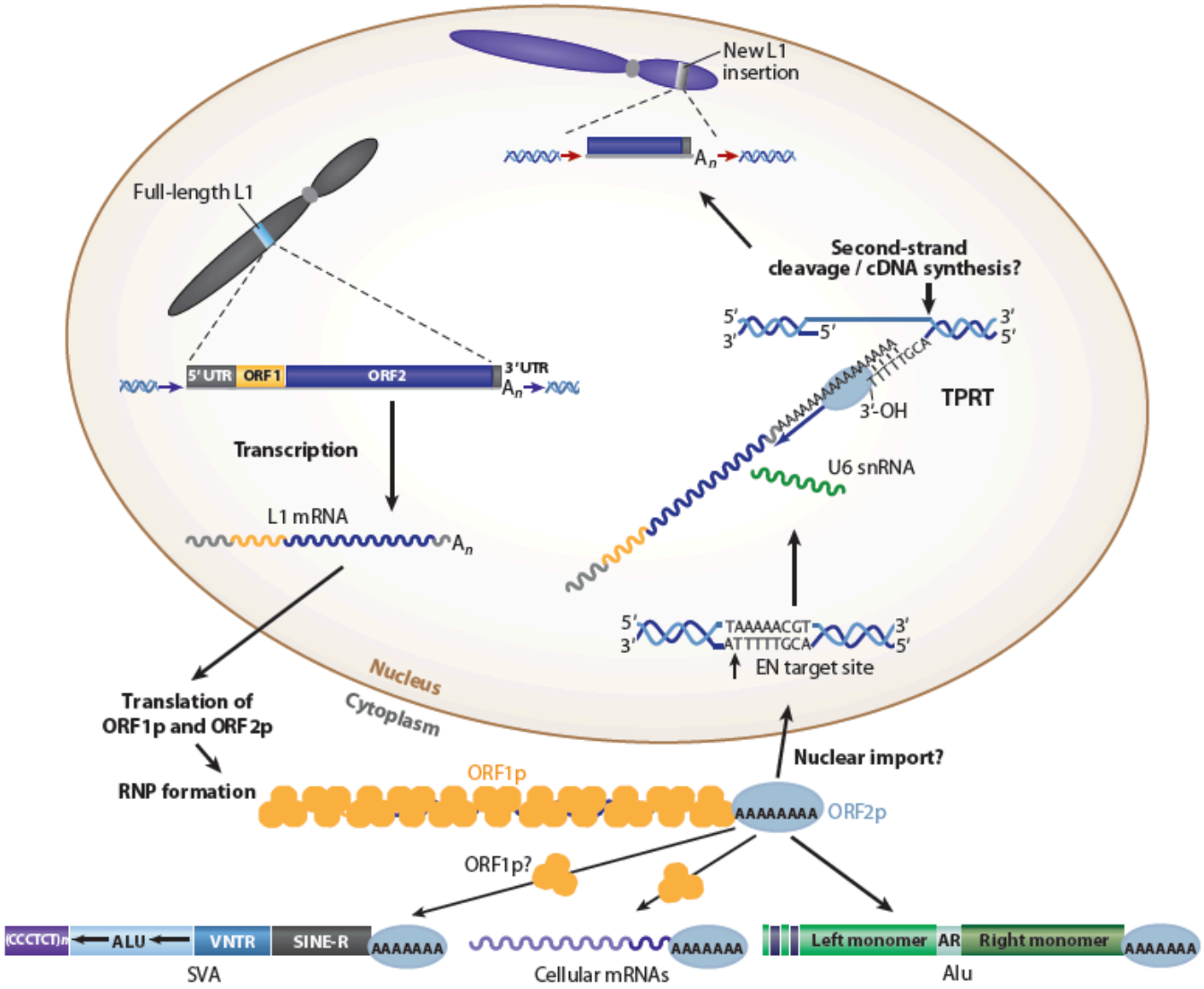


Figure 1.2: A LINE-1 Retrotransposition Cycle

Figure 1.3: Methods to Detect LINE-1-Mediated Polymorphic Human Retrotransposition Events in Individual Genomes

Modifications of these assays can also be used to identify other polymorphic retrotransposons in human DNA. **(a)** Polymerase chain reaction (PCR)-based methodologies. PCR using primers specific to diagnostic sequence variants in the L1 (*red triangles* and the corresponding *maroon primer*) and arbitrary oligonucleotides or primers complementary to ligated linkers (*yellow line*) can be used to amplify human-specific L1s and their associated flanking sequences (*maroon/yellow line* flanked by *triangles*). The amplicon libraries are then resolved using electrophoresis, and individual products are cloned and sequenced (*left*). Alternatively, the amplicons can be hybridized to genome tiling microarrays (*center*), or directly characterized using high-throughput sequencing methodologies (*right*). Abbreviations: ATLAS, amplification typing of L1 active subfamilies; TIP-chip, transposon insertion profiling by microarray **(b)** Mining of L1s in individual genome sequences. Whole-genome sequences, comparative genomics, or mining trace sequence databases can discover dimorphic L1s in individual genomes that are absent from reference genome assemblies. **(c)** Paired-end sequencing. Mate-pair reads containing one sequence from a uniquely mapping portion of genomic DNA and one sequence from an L1 can be used to identify novel retrotransposons (*left*) in individual genomes. Paired-end sequencing of fosmid inserts with restricted size distributions (~40 kb) allows the discovery of novel ~6-kb insertions (*right*) as well as deletions and inversions relative to a reference sequence. Fosmids containing insertions can then be screened for the presence of human-specific L1s. Abbreviation: HGR, human genome reference sequence**.** These methods are also described in another recent review (O'Donnell and Burns, 2010).

Figure 1.3: Methods to Detect LINE-1-Mediated Polymorphic Human Retrotransposition Events in Individual Genomes

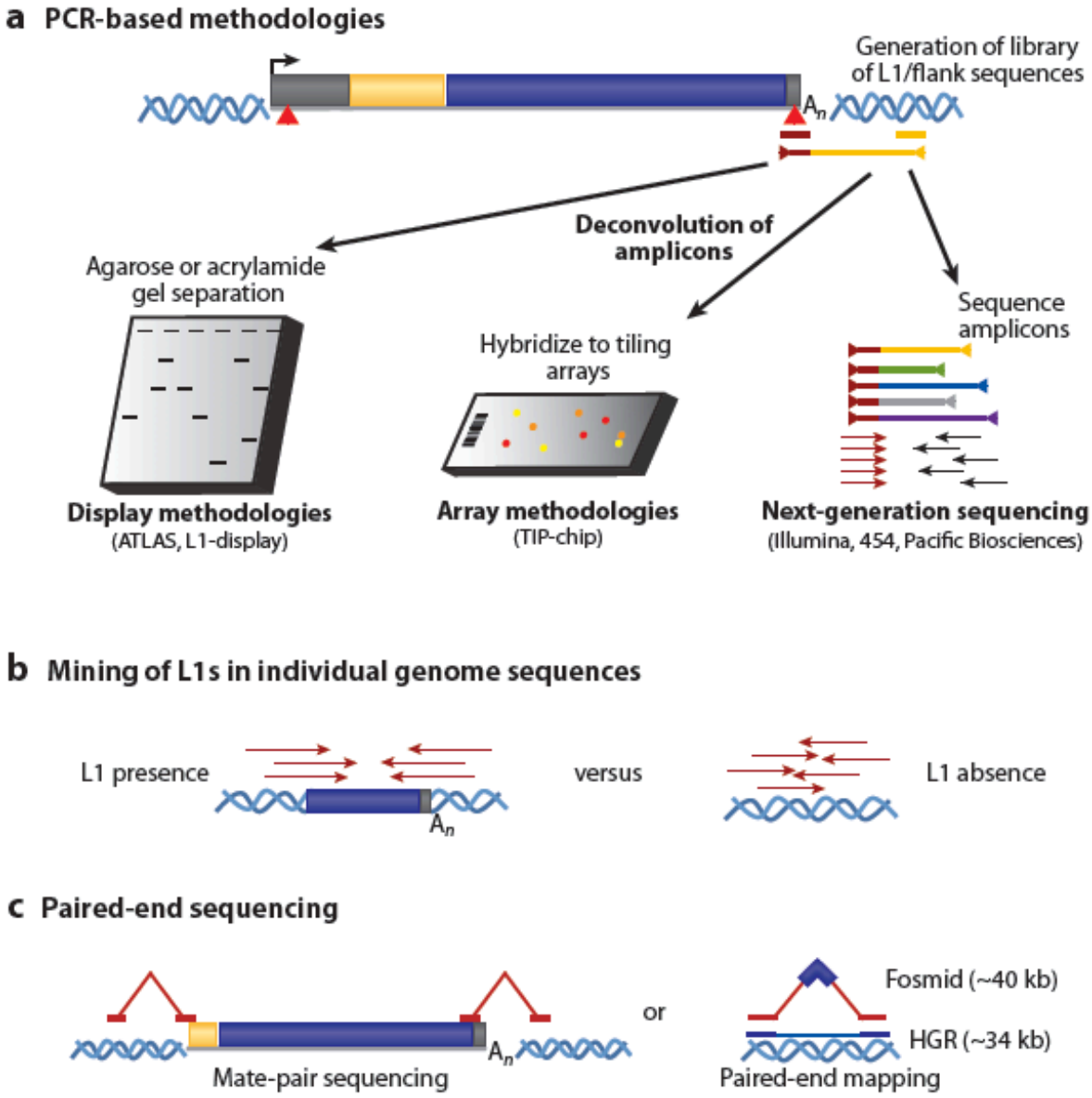


Figure 1.4: A Cultured Cell Assay to Detect LINE-1 Retrotransposition

A cultured cell assay to detect LINE-1 retrotransposition: **(a)** Candidate active human L1s (Beck et al., 2010; Brouha et al., 2003; Moran et al., 1996) are tagged in their 3′ untranslated region (UTR) with an indicator cassette designed to detect retrotransposition events in cultured cells. The selectable/screenable markers [*e.g.*, NEO (Freeman et al., 1994; Moran et al., 1996), GFP (Ostertag et al., 2000), blasticidin (Morrish et al., 2002), and luciferase (Xie et al., 2011)] are in the opposite transcriptional orientation compared with the L1 and contain their own promoters (*backward blue arrow*) and polyadenylation sequences (*upside-down filled red lollipop*); they also contain an intron in the same transcriptional orientation as the L1 [splice donor (SD) and splice acceptor (SA), respectively]. This arrangement ensures that the reporter gene (*e.g., NEO*) will become activated only upon a successful round of retrotransposition (*bottom*). Flags represent epitope tags that can be placed on ORF1p and/or ORF2p (Doucet et al., 2010; Kulpa and Moran, 2005), and open lollipops represent the polyadenylation sequences flanking the L1s. Details regarding the assay can be found in (Moran et al., 1996; Wei et al., 2000). **(b)** Representative results of an L1 retrotransposition assay in cultured HeLa cells. WT is an active L1 allele, L1.3 [accession number L19088 (Dombroski et al., 1993)]. RT− is a retrotransposition-defective control containing a missense mutation (D702A) in the reverse transcriptase domain of ORF2p (Wei et al., 2001). The number of HeLa cells transfected in each experiment is depicted below the wells.

Figure 1.4: A Cultured Cell Assay to Detect LINE-1 Retrotransposition

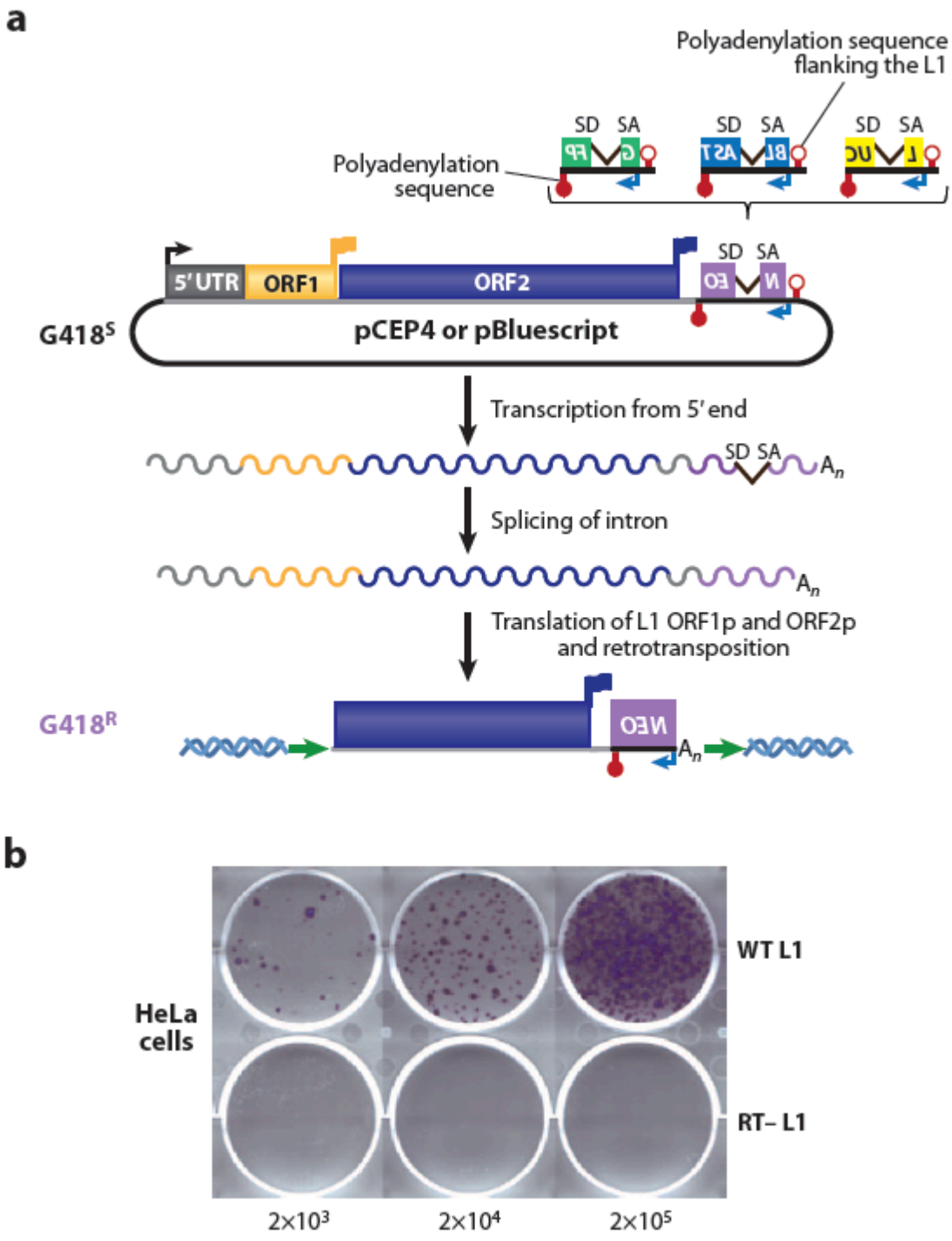


Figure 1.5: The Impact of Mobile Elements on the Human Genome

Schematics highlighting the various ways that LINE-1-mediated retrotransposition events can impact the human genome. **(a)** A hypothetical wild-type gene locus. Light-gray rectangles represent exons, black lines represent introns, and helical lines represent flanking genomic DNA sequence. A full-length (*left*), 5′-truncated (*center*), and inverted/deleted L1 formed by twin priming (*right*) (Ostertag and Kazazian, 2001b) are shown as intronic insertions. The arrows indicate target-site duplications (TSDs), and for simplicity are shown only in this panel. **(b)** Examples of L1-mediated processes that may result in disease. **(c)** Examples of structural variation caused by L1 insertions. The transduction figures show a 3' transduction in light purple with its own poly(A) tail, and a 5' transduction in orange. The nonallelic homologous recombination figure shows L1s at different loci (*light* and *dark gray* exons) acting as substrates for aberrant recombination (*red arrow*). **(d)** Potential effects on gene expression caused by L1 insertion. Note that L1s in all the depicted events would generally contain TSDs, with the exception of endonuclease (EN)-independent retrotransposition events and some genomic deletions. Abbreviations: C, cysteine-rich domain; CC, coiled coil; CTD, carboxyl-terminal domain; EN, endonuclease; RRM, RNA recognition motif; RT, reverse transcriptase; SVA, SINE-R/VNTR/Alu; UTR, untranslated region.

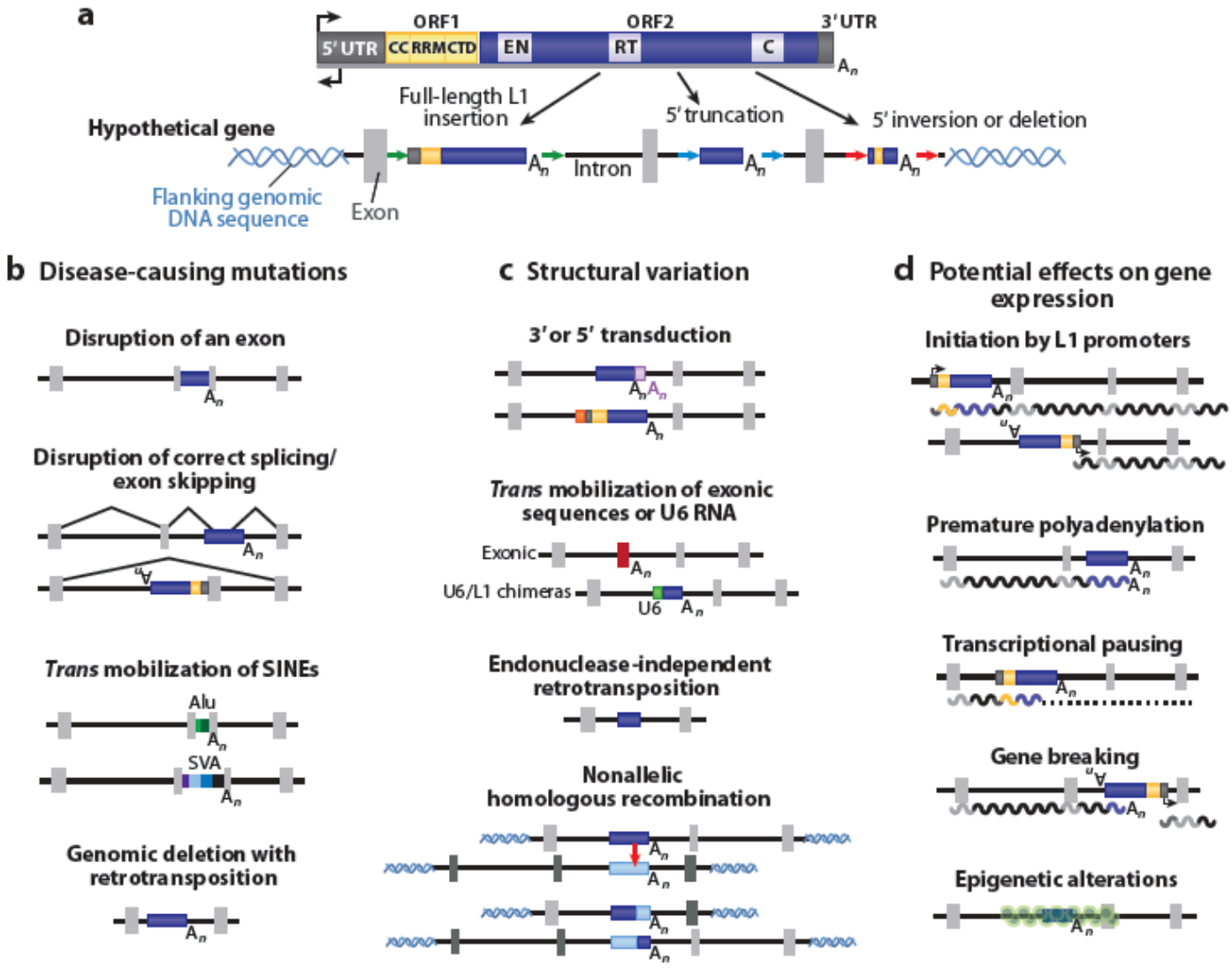


Figure 1.5: The Impact of Mobile Elements on the Human Genome

References

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 Dev *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.

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. Genomics *29*, 136-144.

Asch, H.L., Eliacin, E., Fanning, T.G., Connolly, J.L., Bratthauer, G., and Asch, B.B. (1996). Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues. Oncol Res *8*, 239-247.

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 Res *32*, 3846-3855.

Babushok, D.V., and Kazazian, H.H., Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat *28*, 527-539.

Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S., and Kazazian, H.H., Jr. (2007). A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Res *17*, 1129-1138.

Babushok, D.V., Ostertag, E.M., Courtney, C.E., Choi, J.M., and Kazazian, H.H., Jr. (2006). L1 integration in a transgenic mouse model. Genome Res *16*, 240-250.

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

Bailey, J.A., Carrel, L., Chakravarti, A., and Eichler, E.E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc Natl Acad Sci U S A *97*, 6634-6639.

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

Bannert, N., and Kurth, R. (2006). The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet *7*, 149-173.

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. J Mol Biol *357*, 351-7.

Batzer, M.A., and Deininger, P.L. (2002). Alu repeats and human genomic diversity. Nat Rev Genet *3*, 370-379.

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.

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. Hum Mol Genet *2*, 1697-1702.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature *441*, 87-90.

Belancio, V.P., Hedges, D.J., and Deininger, P. (2008a). Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res *18*, 343-358.

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

Belancio, V.P., Roy-Engel, A.M., and Deininger, P.L. (2010a). All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol *20*, 200-210.

Belancio, V.P., Roy-Engel, A.M., Pochampally, R.R., and Deininger, P. (2010b). Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res *38*, 3909-3922.

Belshaw, R., Dawson, A.L., Woolven-Allen, J., Redding, J., Burt, A., and Tristem, M. (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol *79*, 12507-12514.

Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S., and Devine, S.E. (2004). Natural genetic variation caused by transposable elements in humans. Genetics *168*, 933-951.

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 Res *18*, 1875-1883.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R*., et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

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

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

Boissinot, S., and Furano, A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol *18*, 2186-2194.

Branciforte, D., and Martin, S.L. (1994). Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. Mol Cell Biol *14*, 2584-2592.

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. Am J Hum Genet *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. Proc Natl Acad Sci U S A *100*, 5280-5285.

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

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003a). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res *31*, 4385-4390.

Buzdin, A., Ustyugova, S., Gogvadze, E., Lebedev, Y., Hunsmann, G., and Sverdlov, E. (2003b). Genome-wide targeted search for human specific and polymorphic L1 integrations. Hum Genet *112*, 527-533.

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.

Callinan, P.A., Wang, J., Herke, S.W., Garber, R.K., Liang, P., and Batzer, M.A. (2005). Alu retrotransposition-mediated deletion. J Mol Biol *348*, 791-800.

Carrel, L., Park, C., Tyekucheva, S., Dunn, J., Chiaromonte, F., and Makova, K.D. (2006). Genomic environment predicts expression patterns on the human inactive X chromosome. PLoS Genet *2*, e151.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M.*, et al.* (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol *311*, 17-40.

Chen, J., Rattner, A., and Nathans, J. (2006). 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. Hum Mol Genet *15*, 2146-2156.

Chen, J.M., Chuzhanova, N., Stenson, P.D., Ferec, C., and Cooper, D.N. (2005a). Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat *25*, 207-221.

Chen J.M., Stenson P.D., Cooper D.N., and Ferec C. (2005b). A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet *117*, 411-27.

Chen, L.L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol Cell *35*, 467-478.

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

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E.*, et al.* (2010). LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell *141*, 956-969.

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. Mol Cell Biol *25*, 6617-6628.

Chu, W.M., Liu, W.M., and Schmid, C.W. (1995). RNA polymerase III promoter and terminator elements affect Alu RNA expression. Nucleic Acids Res *23*, 1750-1757.

Cohen, C.J., Lock, W.M., and Mager, D.L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene *448*, 105-114.

Collier, L.S., and Largaespada, D.A. (2005). Hopping around the tumor genome: transposons for cancer gene discovery. Cancer Res *65*, 9607-9610.

Comeaux M.S., Roy-Engel A.M., Hedges D.J., and Deininger P.L. 2009. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? Genome Res *19*, 545-55.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat Genet *38*, 75-81.

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

Cordaux, R., Hedges, D.J., and Batzer, M.A. (2004). Retrotransposition of Alu elements: How many sources? Trends Genet *20*, 464-467.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. Embo J *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.

Craig, N., Craigie, R., Gellert, M., and Lambowitz, A. (2002). Mobile DNA II (Washington, DC, ASM).

Cruickshanks, H.A., and Tufarelli, C. (2009). Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics *94*, 397-406.

Curcio, M.J., and Belfort, M. (2007). The beginning of the end: links between ancient retroelements and modern telomerases. Proc Natl Acad Sci U S A *104*, 9107-9108.

Damert, A., Raiz, J., Horn, A.V., Lower, J., Wang, H., Xing, J., Batzer, M.A., Lower, R., and Schumann, G.G. (2009). 5'-transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res *19*, 1992-2008.

Davidson, E.H., and Britten, R.J. (1979). Regulation of gene expression: possible role of repetitive sequences. Science *204*, 1052-1059.

Deininger, P.L., Batzer, M.A., Hutchison, C.A. III, and Edgell, M.H. (1992). Master genes in mammalian repetitive DNA amplification. Trends Genet *8*, 307-311.

Denison, R.A., Van Arsdell, S.W., Bernstein, L.B., and Weiner, A.M. (1981). Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc Natl Acad Sci U S A *78*, 810-814.

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

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res *16*, 1548-1556.

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

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell *122*, 473-483.

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. Mol Cell Biol *27*, 4685-4697.

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.

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci U S A *90*, 6513-6517.

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 Genet *6,* e1001150.

Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Ejima, Y., and Yang, L. (2003). Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Hum Mol Genet *12*, 1321-1328.

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. J Biol Chem *279*, 27753-27763.

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

Ewing, A.D., and Kazazian, H.H. Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

Ewing, A.D., and Kazazian, H.H. Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res *21*, 985-990.

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 Res *15*, 2251-2260.

Farley, A.H., Luning Prak, E.T., and Kazazian, H.H., Jr. (2004). More active human L1 retrotransposons produce longer insertions. Nucleic Acids Res *32*, 502-510.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T.*, et al.* (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat Genet *41*, 563-571.

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. (2008). Transposable elements and the evolution of regulatory networks. Nat Rev Genet *9*, 397-405.

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques *17*, 46, 48-49, 52.

Gabriel, A., Dapprich, J., Kunkel, M., Gresham, D., Pratt, S.C., and Dunham, M.J. (2006). Global mapping of transposon location. PLoS Genet *2*, e212.

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 Res *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. Hum Mol Genet *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.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol *357*, 1383-1393.

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. Hum Mol Genet *18*, 1221-1228.

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

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

Gladyshev, E.A., and Arkhipova, I.R. (2007). Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci U S A *104*, 9352-9357.

Goodier, J.L., and Kazazian, H.H. Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23-35.

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. Hum Mol Genet *19*, 1712-1725.

Goodier, J.L., and Maraia, R.J. (1998). Terminator-specific recycling of a B1-Alu transcription complex by RNA polymerase III is mediated by the RNA terminus-binding protein La. J Biol Chem *273*, 26110-26116.

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

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. Mol Cell Biol *27*, 6469-6483.

Graur, D., Shuali, Y., and Li, W.H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol *28*, 279-285.

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

Hackett, P.B., Largaespada, D.A., and Cooper, L.J. (2010). A transposon and transposase system for human application. Mol Ther *18*, 674-683.

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. Proc Natl Acad Sci U S A *105*, 19366-19371.

Han, K., Lee, J., Meyer, T.J., Wang, J., Sen, S.K., Srikanta, D., Liang, P., and Batzer, M.A. (2007). Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet *3*, 1939-1949.

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 Res *33*, 4040-4052.

Hancks, D.C., Ewing, A.D., Chen, J.E., Tokunaga, K., and Kazazian, H.H. Jr. (2009). Exon-trapping mediated by the human retrotransposon SVA. Genome Res *19*, 1983-1991.

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. Hum Mol Genet *20*, 3386-3400.

Harrison, P.M., Zheng, D., Zhang, Z., Carriero, N., and Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res *33*, 2374-2383.

Hayakawa, T., Satta, Y., Gagneux, P., Varki, A., and Takahata, N. (2001). Alu-mediated inactivation of the human CMP-N-acetylneuraminic acid hydroxylase gene. Proc Natl Acad Sci U S A *98*, 11399-11404.

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

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. Nat Genet *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. J Biol Chem *267*, 19765-19768.

Hormozdiari, F., Alkan, C., Ventura, M., Hajirasouliha, I., Malig, M., *et al.* (2011). Alu repeat discovery and characterization within human genomes. Genome Res *21*, 840-849.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat Genet *36*, 949-951.

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.

Ivics, Z., Li, M.A., Mates, L., Boeke, J.D., Nagy, A., Bradley, A., and Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. Nat Methods *6*, 415-422.

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. J Biol Chem *282*, 24893-24904.

Jurka, J. (2008). Conserved eukaryotic transposable elements and the evolution of gene regulation. Cell Mol Life Sci *65*, 201-204.

Kano, H., Godoy, I., Courtney, C., Vetter, M.R., Gerton, G.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev *23*, 1303-1312.

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

Kawakami, K. (2005). Transposon tools and methods in zebrafish. Dev Dyn *234*, 244-254.

Kazazian, H.H. Jr., and Moran, J.V. (1998). The impact of L1 retrotransposons on the human genome. Nat Genet *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.

Keng, V.W., Ryan, B.J., Wangensteen, K.J., Balciunas, D., Schmedt, C., Ekker, S.C., and Largaespada, D.A. (2009). Efficient transposition of Tol2 in the mouse germline. Genetics *183*, 1565-1573.

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 Res *16*, 78-87.

Khazina, E., and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A *106*, 731-736.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F.*, et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

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.

Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R., and Eichler, E.E. (2007). Population stratification of a common APOBEC gene deletion polymorphism. PLoS Genet *3*, e63.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet *8*, 1557-1560.

Kolosha, V.O., and Martin, S.L. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A *94*, 10155-10160.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L.*, et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420-426.

Kubo, S., Seleme Mdel, 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. Proc Natl Acad Sci U S A *103*, 8036-8041.

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

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *13*, 655-660.

Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D.C., Moore, L., Nakashima, K., Asashima, M., and Gage, F.H. (2009). Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat Neurosci *12*, 1097-1105.

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.

Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res *14*, 2253-2260.

Lee, Y.N., and Bieniasz, P.D. (2007). Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog *3*, e10.

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.

Lemmers, R.J., van der Vliet, P.J., Klooster, R., Sacconi, S., Camano, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W.*, et al.* (2010). A unifying genetic model for facioscapulohumeral muscular dystrophy. Science *329*, 1650-1653.

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. Science *300*, 1288-1291.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G.*, et al.* (2007). The diploid genome sequence of an individual human. PLoS Biol *5*, e254.

Li, X., Scaringe, W.A., Hill, K.A., Roberts, S., Mengos, A., Careri, D., Pinto, M.T., Kasper, C.K., and Sommer, S.S. (2001). Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat *17*, 511-519.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.G., Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K.*, et al.* (2009). Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell *139*, 1069-1083.

Liu, W.M., and Schmid, C.W. (1993). Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. Nucleic Acids Res *21*, 1351-1359.

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.

Lupski, J.R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet *1*, e49.

Lutz, S.M., Vincent, B.J., Kazazian, H.H. Jr., Batzer, M.A., and Moran, J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am J Hum Genet *73*, 1431-1437.

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet *80*, 133-137.

Macfarlane, C., and Simmonds, P. (2004). Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. J Mol Evol *59*, 642-656.

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. Mol Cell Biol *31*, 300-316.

Maestre, J., Tchenio, T., Dhellin, O., and Heidmann, T. (1995). mRNA retroposition in human cells: processed pseudogene formation. Embo J *14*, 6333-6338.

Maksakova, I.A., Romanish, M.T., Gagnier, L., Dunn, C.A., van de Lagemaat, L.N., and Mager, D.L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet *2*, e2.

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. J Mol Biol *247*, 49-59.

Martin, S.L. (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol *11*, 4804-4807.

Martin, S.L. (2009). Developmental biology: jumping-gene roulette. Nature *460*, 1087-1088.

Martin, S.L., Branciforte, D., Keller, D., and Bain, D.L. (2003). Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci U S A *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. Mol Cell Biol *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. Cytogenet Genome Res *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. J Biomed Biotechnol *2006*, 71753.

Mayer, J., Meese, E., and Mueller-Lantzsch, N. (1997). Multiple human endogenous retrovirus (HERV-K) loci with gag open reading frames in the human genome. Cytogenet Cell Genet *78*, 1-5.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A *36*, 344-355.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C.*, et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res *19*, 1527-1541.

McMillan, J.P., and Singer, M.F. (1993). Translation of the human LINE-1 element, L1Hs. Proc Natl Acad Sci U S A *90*, 11533-11537.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S.*, et al.* (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature *403*, 785-789.

Michaud, E.J., van Vugt, M.J., Bultman, S.J., Sweet, H.O., Davisson, M.T., and Woychik, R.P. (1994). Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. Genes Dev *8*, 1463-1472.

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 Res *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. Am J Hum Genet *78*, 671-679.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K.*, et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M., and 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 Res *20*, 3139-3145.

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. Hum Mutat *28*, 137-142.

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.

Morisada, N., Rendtorff, N.D., Nozu, K., Morishita, T., Miyakawa, T., Matsumoto, T., Hisano, S., Iijima, K., Tranebjaerg, L., Shirahata, A.*, et al.* (2010). Branchio-oto-renal syndrome caused by partial EYA1 deletion due to LINE-1 insertion. Pediatr Nephrol *25*, 1343-1348.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208-212.

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. Nat Genet *31*, 159-165.

Moyes, D., Griffiths, D.J., and Venables, P.J. (2007). Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet *23*, 326-333.

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.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V.*, et al.* (2002). A

comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet *71*, 312-326.

Narita, N., Nishio, H., Kitoh, 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. J Clin Invest *91*, 1862-1867.

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.

O'Donnell, K.A., and Burns, K.H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob DNA *1*, 21.

OhAinle, M., Kerns, J.A., Li, M.M., Malik, H.S., and Emerman, M. (2008). Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. Cell Host Microbe *4*, 249-259.

Ohshima, K., and Igarashi, K. (2010). Inference for the initial stage of domain shuffling: tracing the evolutionary fate of the PIPSL retrogene in hominoids. Mol Biol Evol *27*, 2522-2533.

Ono, M., Kawakami, M., and Takezawa, T. (1987). A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res *15*, 8725-8737.

Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., Hino, T., Suzuki-Migishima, R., Ogonuki, N., Miki, H.*, et al.* (2006). Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet *38*, 101-106.

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

Ostertag, E.M., and Kazazian, H.H. Jr. (2001a). Biology of mammalian L1 retrotransposons. Annu Rev Genet *35*, 501-538.

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

Ostertag, E.M., Prak, E.T., DeBerardinis, R.J., Moran, J.V., and Kazazian, H.H. Jr. (2000). Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res *28*, 1418-1423.

Ovchinnikov, I., Rubin, A., and Swergold, G.D. (2002). Tracing the LINEs of human evolution. Proc Natl Acad Sci U S A *99*, 10522-10527.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res *11*, 2050-2058.

Pace, J.K. II, and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res *17*, 422-432.

Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G.*, et al.* (2009). An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science *325*, 995-998.

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

Peterson, S.E., Westra, J.W., Paczkowski, C.M., and Chun, J. (2008). Chromosomal mosaicism in neural stem cells. Methods Mol Biol *438*, 197-204.

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

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W.*, et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444-454.

Robins, D.M., and Samuelson, L.C. (1992). Retrotransposons and the evolution of mammalian gene expression. Genetica *86*, 191-201.

Rouchka, E., Montoya-Durango, D.E., Stribinskis, V., Ramos, K., and Kalbfleisch, T. (2010). Assessment of genetic variation for the LINE-1 retrotransposon from next generation sequence data. BMC Bioinformatics *11 Suppl 9*, S12.

Roy, A.M., Carroll, M.L., Kass, D.H., Nguyen, S.V., Salem, A.H., Batzer, M.A., and Deininger, P.L. (1999). Recently integrated human Alu repeats: finding needles in the haystack. Genetica *107*, 149-161.

Rozmahel, R., Heng, H.H., Duncan, A.M., Shi, X.M., Rommens, J.M., and Tsui, L.C. (1997). Amplification of CFTR exon 9 sequences to multiple locations in the human genome. Genomics *45*, 554-561.

Salem, A.H., Kilroy, G.E., Watkins, W.S., Jorde, L.B., and Batzer, M.A. (2003). Recently integrated Alu elements and human genomic diversity. Mol Biol Evol *20*, 1349-1361.

Sarrowa, J., Chang, D.Y., and Maraia, R.J. (1997). The decline in human Alu retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric Alu RNA and increased expression of small cytoplasmic Alu RNA. Mol Cell Biol *17*, 1144-1151.

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., Kimura-Yoshida, C., Matsuo, I., Sumiyama, K., Saitou, N.*, et al.* (2008). Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A *105*, 4220-4225.

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. Nat Genet *16*, 37-43.

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.

Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G., Kirschner, R., Hemberger, M., Bergen, A.A., Rosenberg, T.*, et al.* (1998). Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nat Genet *19*, 327-332.

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.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M.*, et al.* (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525-528.

Segal, Y., Peissel, B., Renieri, A., de Marchi, M., Ballabio, A., Pei, Y., and Zhou, J. (1999). LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. Am J Hum Genet *64*, 62-69.

Sela, N., Mersch, B., Hotz-Wagenblatt, A., and Ast, G. (2010). Characteristics of transposable element exonization within human and mouse. PLoS One *5*, e10907.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H. Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A *103*, 6611-6616.

Sen, S.K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P.A., Dyer, M., Cordaux, R., Liang, P., and Batzer, M.A. (2006). Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet *79*, 41-53.

Sen, S.K., Huang, C.T., Han, K., and Batzer, M.A. (2007). Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res *35*, 3741-3751.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., and Swergold, G.D. (2000). Reading between the LINEs:

human genomic variation induced by LINE-1 retrotransposition. Genome Res *10*, 1496-1508.

Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R., Dangel, A.W., Carroll, M.C., Zipf, W.B., and Yu, C.Y. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region: molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem *269*, 8466-8476.

Shen, S., Lin, L., Cai, J.J., Jiang, P., Kenkel, E.J., Stroik, M.R., Sato, S., Davidson, B.L., and Xing, Y. (2011). Widespread establishment and regulatory impact of Alu exons in human genes. Proc Natl Acad Sci U S A *108*, 2837-2842.

Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G., and Gage, F.H. (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci *33*, 345-354.

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

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. Curr Opin Genet Dev *6*, 743-748.

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

Snider, L., Geng, L.N., Lemmers, R.J., Kyba, M., Ware, C.B., Nelson, A.M., Tawil, R., Filippova, G.N., van der Maarel, S.M., Tapscott, S.J*., et al.* (2010). Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. PLoS Genet *6*, e1001181.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

Srikanta, D., Sen, S.K., Huang, C.T., Conlin, E.M., Rhodes, R.M., and Batzer, M.A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. Genomics *93*, 205-212.

Stetson, D.B., Ko, J.S., Heidmann, T., and Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. Cell *134*, 587-598.

Sun, L.V., Jin, K., Liu, Y., Yang, W., Xie, X., Ye, L., Wang, L., Zhu, L., Ding, S., Su, Y*., et al.* (2008). PBmice: an integrated database system of piggyBac (PB) insertional mutations and their characterizations in mice. Nucleic Acids Res *36*, D729-734.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol *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.

Takasu, M., Hayashi, R., Maruya, E., Ota, M., Imura, K., Kougo, K., Kobayashi, C., Saji, H., Ishikawa, Y., Asai, T.*, et al.* (2007). Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon. Tissue Antigens *70*, 144-150.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M.*, et al.* (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534-538.

Tang, Y.A., Huntley, D., Montana, G., Cerase, A., Nesterova, T.B., and Brockdorff, N. (2010). Efficiency of Xist-mediated silencing on autosomes is linked to chromosomal domain organisation. Epigenetics Chromatin *3*, 10.

Tchenio, T., Casella, J.F., and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res *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. Hum Mutat *29*, 931-938.

Tomlins, S.A., Laxman, B., Dhanasekaran, S.M., Helgeson, B.E., Cao, X., Morris, D.S., Menon, A., Jing, X., Cao, Q., Han, B.*, et al.* (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature *448*, 595-599.

Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. (2003). A genome-wide survey of human pseudogenes. Genome Res *13*, 2559-2567.

Trelogan, S.A., and Martin, S.L. (1995). Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc Natl Acad Sci U S A *92*, 1520-1524.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.*, et al.* (2005). Fine-scale structural variation of the human genome. Nat Genet *37*, 727-732.

Ullu, E., Esposito, V., and Melli, M. (1982). Evolutionary conservation of the human 7 S RNA sequences. J Mol Biol *161*, 195-201.

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

Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T., and Gesteland, R.F. (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell *26*, 11-17.

Van den Hurk, J.A., Meij, I.C., Seleme, M.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. Hum Mol Genet *16*, 1587-1592.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. Annu Rev Genet *19*, 253-272.

Voliva, C.F., Martin, S.L., Hutchison, C.A. III, and Edgell, M.H. (1984). Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J Mol Biol *178*, 795-813.

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. J Mol Biol *354*, 994-1007.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., and Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat *27*, 323-329.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y.*, et al.* (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60-65.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A *104*, 18613-18618.

Weber, M.J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet *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. Mol Cell Biol *21*, 1429-1439.

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem *284*, 435-438.

Weiner, A.M., Deininger, P.L., and Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem *55*, 631-661.

Westra, J.W., Rivera, R.R., Bushman, D.M., Yung, Y.C., Peterson, S.E., Barral, S., and Chun, J. (2010). Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. J Comp Neurol *518*, 3981-4000.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res *15*, 1073-1078.

Wheelan, S.J., Scheifele, L.Z., Martinez-Murillo, F., Irizarry, R.A., and Boeke, J.D. (2006). Transposon insertion site profiling chip (TIP-chip). Proc Natl Acad Sci U S A *103*, 17632-17637.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T.*, et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872-876.

Witherspoon, D.J., Xing, J., Zhang, Y., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2010). Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics *11*, 410.

Xie, Y., Rosser, J.M., Thompson, T.L., Boeke, J.D., and An, W. (2011). Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. Nucleic Acids Res *39*, e16.

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. Proc Natl Acad Sci U S A *103*, 17608-17613.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res *19*, 1516-1526.

Yang, N., and Kazazian, H.H. Jr. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol *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 Res *31*, 4929-4940.

Zhang, Z., Harrison, P., and Gerstein, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res *12*, 1466-1482.

Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res *13*, 2541-2558.

Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F., and Craig, N.L. (2004). Transposition of hAT elements links transposable elements and V(D)J recombination. Nature *432*, 995-1001.

Zingler, N., Willhoeft, U., Brose, H.P., Schoder, V., Jahns, T., Hanschmann, K.M., Morrish, T.A., Lower, J., and Schumann, G.G. (2005). Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. Genome Res *15*, 780-789.

Chapter 2

LINE-1 Retrotransposition Activity in Human Genomes

Abstract

Highly active (*i.e.*, “hot”) long interspersed element-1 (LINE-1 or L1) sequences comprise the bulk of retrotransposition activity in the human genome; however, the abundance of hot L1s in the human population remains largely unexplored. Here, we used a fosmid-based, paired-end DNA sequencing strategy to identify 68 full-length L1s that are differentially present among individuals but are absent from the human genome reference sequence. The majority of these L1s were highly active in a cultured cell retrotransposition assay. Genotyping 26 elements revealed that two L1s are only found in Africa and that two more are absent from the H952 subset of the Human Genome Diversity Panel. Therefore, these results suggest that hot L1s are more abundant in the human population than previously appreciated, and that ongoing L1 retrotransposition continues to be a major source of inter-individual genetic variation.

Introduction

L1s comprise ~17% of human DNA and have been an instrumental force in shaping genome architecture (Lander et al., 2001). Most L1s are molecular fossils that cannot move (retrotranspose) to new genomic locations (Grimaldi and Singer, 1983; Lander et al., 2001). However, a small number of human-specific L1 (L1Hs) elements remain retrotransposition-competent (Badge et al., 2003; Brouha et al., 2003; Sassaman et al., 1997). On occasion, their retrotransposition has resulted in sporadic cases of human disease (reviewed in Babushok and Kazazian, 2007; Kazazian et al., 1988).

During the past 15 years, computational, molecular biological, and genomic approaches have been used to identify and characterize L1Hs elements (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2000; Boissinot et al., 2004; Brouha et al., 2003; Lander et al., 2001; Moran et al., 1996; Myers et al., 2002; Ovchinnikov et al., 2001; Sheen et al., 2000; Xing et al., 2009). Several themes have emerged from these studies. First, L1Hs elements can be stratified into several subfamilies (pre-Ta, Ta-0, Ta-1, Ta1-d, Ta1-nd) based upon the presence of diagnostic sequence variants contained within their 5′ and/or 3′ untranslated regions (UTRs) (Boissinot et al., 2000; Skowronski et al., 1988; Smit et al., 1995). Second, many L1Hs elements are dimorphic in that they are differentially present in individual genomes and/or are present in an individual but absent from the haploid Human Genome Reference sequence (HGR) (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2004; Brouha et al., 2003; Lander et al., 2001; Myers et al., 2002; Xing et al., 2009). Third, it has been estimated

that the average human genome contains ~80-100 active (retrotransposition-competent) L1Hs elements, and that only a small number of highly active L1Hs elements ("hot" L1s) account for the bulk of retrotransposition activity in the HGR (Brouha et al., 2003). Those studies, as well as recent efforts to identify insertion, deletion, and inversion polymorphisms (structural variants) in humans (Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005; Xing et al., 2009), indicate that ongoing L1 retrotransposition contributes to inter-individual genetic variation.

Here, we employed a fosmid-based, paired-end DNA resource to identify full-length L1Hs elements in the genomes of six individuals of diverse geographic origin. Over half (37/68) of the newly identified L1s were hot for retrotransposition when examined in a cultured cell assay (Moran et al., 1996). Genotyping a subset of these L1s further revealed that some are likely restricted to Africans, whereas others are absent from the Human Genome Diversity Panel (HGDP) (Cann et al., 2002), suggesting that they are present at very low allele frequencies.

Results

An Experimental Strategy to Identify Full-Length Human Specific L1s

To identify novel, full-length L1s in the genomes of geographically diverse individuals, we exploited a fosmid-based, paired-end DNA sequencing strategy that previously was used to identify structural variants in human DNA (Kidd et al., 2008; Tuzun et al., 2005). Fragments of genomic DNA approximately 40 kb in size were individually cloned using fosmid vectors (see Experimental Procedures). Sequence reads were obtained from both ends of each insert

(paired-end sequences) and compared to the HGR. End sequences from genomic fragments that do not differ significantly in size from the HGR will map ~40 kb away from each other. In contrast, paired-end sequences derived from genomic fragments containing a full-length, dimorphic ~6 kb L1Hs element will be separated by ~34 kb when mapped to the HGR (Figure 2.1) (Tuzun et al., 2005). In general, the predicted variants were required to be supported by two fosmid clones containing putative insertions from the same individual. The size cutoffs used in our screening protocols are biased to allow the identification of full-length or near full-length L1 insertion polymorphisms, but not severely 5′ truncated L1 sequences, which are replication deficient (Table 2.1). Through this scheme, we should be able to identify the bulk of full-length L1s in an individual genome that are dimorphic when compared to the HGR.

Fosmids fulfilling the above mapping criterion were subjected to a series of screens (Figure 2.1). First, allele-specific oligonucleotide hybridization using probes directed against diagnostic sequences in the L1Hs 5′ UTR identified insertion fosmids that contained putative dimorphic L1Hs elements (Boissinot et al., 2000; Tuzun et al., 2005). Second, Southern blotting with a probe directed against the 5′ UTR of L1.3 (Accession# L19088) enabled the identification of fosmids that contained putative full-length L1Hs elements (Dombroski et al., 1993; Sassaman et al., 1997). Third, a suppression PCR-based method (ATLAS) (Badge et al., 2003) and/or direct sequencing was used to verify the presence of a full-length (or near full-length) L1Hs element in the fosmid. Finally, genomic sequences flanking the 5′ and 3′ ends of the newly identified L1Hs elements were

used as probes in BLAT searches (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) (Kent, 2002) to confirm that the L1 was absent from the HGR (NCBI build 36.1/hg18). Flanking sequences also were used to determine whether any of the L1Hs elements were present in a database of known polymorphic retrotransposon insertions (dbRIP; http://dbrip.brocku.ca/) (Wang et al., 2006). Two additional L1Hs elements were identified through direct sequencing of the fosmids (#1-(2-1) and 3-(2-1)).

Identification of Full-Length L1Hs Elements from Geographically Diverse Individuals

We first conducted a pilot study to examine a fosmid library from a female individual (G248; NA15510) for full-length L1Hs insertions (Table 2.1) (Tuzun et al., 2005). Despite the fact that this library was optimized for identifying ~8 kb insertion polymorphisms as part of the Human Genome Structural Variation project (HGSV) (Kidd et al., 2008; Tuzun et al., 2005), we were able to identify five novel L1Hs elements using our screening protocol (Table 2.1).

The above data provided “proof of principle” that our strategy was effective for identifying full-length, dimorphic L1Hs elements. Thus, we next screened fosmid libraries from five females representing four distinct geographic populations that were studied as part of the HapMap project (one Japanese (NA18956), one Chinese (NA18555), one Western European CEPH (NA12878), and two Yoruban individuals (NA19240, NA19129)) (International HapMap Consortium, 2005; Kidd et al., 2008). Size cutoffs allowed detection of insertion polymorphisms as small as ~4.2-5.5 kb and enabled the identification of an

additional 64 L1Hs elements (Table 2.1) (Kidd et al., 2008). As our strategy is biased toward finding novel, full-length L1s, we generally observed a decrease in the number of L1Hs elements identified in each successive library screen (*e.g.*, ABC13 was the last library analyzed and contained relatively few novel L1Hs elements). In total, we identified 69 L1Hs elements that were absent from the HGR, one of which was identified in two different individuals (#4-1 and 5-77). This element also was completely annotated in dbRIP, unlike 65 of the distinct 68 L1s identified in this study (Table 2.1). The number of elements discovered at each stage of the analysis is detailed in the Extended Experimental Procedures.

Many of the Newly Identified L1Hs Elements are Hot for Retrotransposition

We next tested if the L1Hs elements identified in our screens were active for retrotransposition in cultured cells. Sixty-seven elements were cloned into either a pBluescript and/or pCEP4 L1 expression vector that contained an *mneoI* retrotransposition indicator cassette in its 3′ UTR (#2-42 was refractory to cloning; details in Experimental Procedures) (Freeman et al., 1994; Moran et al., 1996). The pBluescript-based L1 constructs lack an exogenous promoter; thus, L1 expression is driven from its native 5′ UTR. Elements isolated from libraries ABC11-13 were assayed in this context. L1s isolated from the G248, ABC9, and ABC10 libraries were assayed in pCEP4 (CMV+/5′ UTR+) and/or pBluescript (5′ UTR+) based contexts. The resultant plasmids were transfected into HeLa cells and successful retrotransposition events were detected as G418-resistant foci (Figure 2.2A) (Moran et al., 1996). Retrotransposition activities are reported relative to L1.3, and hot refers to an L1 that jumps at >10% of L1.3 (see Table

2.2). Notably, 22 elements yielded similar retrotransposition efficiencies relative to L1.3 when tested in either a CMV+/5′ UTR+ or a 5′ UTR+ context (data not shown). Since the subcloning procedure does not involve PCR, we truly are testing the retrotransposition capability of each of the identified L1Hs elements in our screen.

Each individual contained between three and nine highly active L1s in their genome and 55% (37/67) of the L1Hs elements tested were hot for retrotransposition (Figures 2.2A and 2.2B; Table 2.1). These 37 highly active L1Hs elements represent an approximately 4-fold increase in the number of hot L1s identified in previous studies (Badge et al., 2003; Brouha et al., 2002, 2003; Kimberland et al., 1999; Lander et al., 2001; Sassaman et al., 1997). Examination of the 3′ UTR sequences of the 68 L1s uncovered six elements that contain an ACG in place of the Ta subfamily diagnostic ACA characters. These elements are termed “pre-Ta” and represent an older L1 subfamily (Boissinot et al., 2000; Brouha et al., 2003; Kazazian et al., 1988; Lander et al., 2001; Myers et al., 2002; Skowronski et al., 1988). Two pre-Ta L1s (#3-5 and 5-55) were hot for retrotransposition (Figure 2.2B; Table 2.2). These data agree with previous studies, which showed that a *de novo* insertion of a pre-Ta L1 into the *Factor VIII* gene resulted in a sporadic case of hemophilia A (Kazazian et al., 1988).

Hallmarks and Insertion Locations of L1s Identified in This Study

We next sequenced each L1Hs element in its entirety and compared these data to fosmid sequences previously deposited in GenBank (Kidd et al., 2008). We annotated each L1 for hallmarks of retrotransposition as well as their

chromosomal environment (Table 2.3). In general, the L1Hs elements were flanked by target-site duplications that ranged from 6 to 20 bp, inserted into an L1 endonuclease consensus cleavage sequence (Cost and Boeke, 1998; Feng et al., 1996; Morrish et al., 2002), and their 3′ ends had either homopolymeric poly(A) tails that ranged from ~8-41 bp in size or interrupted poly(A) tails/3′ transductions ranging from ~18 bp to 1,105 bp in length (Table 2.3) (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000).

A subset of the elements (~32/68) contained an additional 1-14bp of untemplated nucleotides at their 5′ ends, termed 5′ end heterogeneity (Athanikar et al., 2004; Lavie et al., 2004). Five of these L1s have an extra G at their 5′ ends, and one has three extra Gs when compared to a hot L1Hs consensus sequence (Brouha et al., 2003). These extra nucleotides potentially could result either from a terminal transferase activity associated with the L1 reverse transcriptase or from reverse transcription of the 7-methyl-guanosine cap at the 5′ end of L1 RNA (Boeke, 2003; Gilbert et al., 2005; Symer et al., 2002). The majority of elements identified were full-length; however, we also found seven elements (*e.g.,* #1-5 and 2-30) that were truncated within their 5′ UTR. These data, along with the fact that the fosmid libraries provided ~4-5 fold coverage of each haplotype from the 6 individuals (Kidd et al., 2008), indicate that our screening procedure identified the majority of the dimorphic full-length L1s in these genomes.

The 68 L1Hs elements were dispersed throughout the genome. We did not identify L1Hs elements on chromosomes 16 or 19 (Figure 2.2C); however, this

result probably reflects our small sample size rather than a systematic bias against their ability insert on these chromosomes (Lander et al., 2001). Consistently, we previously were able to detect the insertion of engineered L1s into chromosomes 16 and 19 of HeLa cells (Gilbert et al., 2005).

Approximately 32% (22/68) of L1Hs elements were present in the introns of known RefSeq genes (http://www.ncbi.nlm.nih.gov/RefSeq/), and mutations in several of these genes are implicated in human genetic disorders (Table 2.4). Thirteen L1 insertions were in the antisense orientation (*i.e.*, were transcribed in the opposite orientation to the gene), whereas nine L1 insertions were in the same transcriptional orientation as the gene. Since ~26-38% of the genome is spanned by genes (Venter et al., 2001), the data suggest that the L1s have inserted randomly with respect to gene content, which is in agreement with previous studies (Gilbert et al., 2002, 2005; Ovchinnikov et al., 2001; Symer et al., 2002).

Our sequencing studies uncovered several expected trends and some unexpected results. All 37 hot L1 elements and the 6 low-level activity elements had two intact open reading frames (ORFs). A consensus sequence derived from these 37 L1s was identical at the amino acid level to a previously derived consensus (Brouha et al., 2003) (data not shown).

Inactive elements generally had frameshift (5/24) or chain-terminating nonsense mutations (9/24) in at least one of the L1 ORFs. However, ten of these low-level activity or inactive elements contained two intact open reading frames. One L1 (#3-24) contained an S228P missense mutation within the endonuclease

(EN) domain of ORF2p (Feng et al., 1996; Weichenrieder et al., 2004). Though L1s containing EN mutations are unable to retrotranspose in HeLa cells, they can retrotranspose in Chinese Hamster Ovary (CHO) cells deficient in the nonhomologous end-joining (NHEJ) pathway of DNA repair, presumably by parasitizing a free 3′ OH group to initiate target-primed reverse transcription (TPRT) (Morrish et al., 2002, 2007). Interestingly, although #3-24 is inactive in NHEJ-proficient cell lines, the L1 retrotransposed at roughly 60% the efficiency of the wild-type control, L1.3, in NHEJ-deficient CHO cells (Morrish et al., 2002). Introducing the S228P change into L1.3 (Sassaman et al., 1997) also allowed efficient EN-independent retrotransposition, indicating that this mutation is largely responsible for the inactivity of #3-24 in HeLa cells (Figure 2.7).

Analysis of genomic sequences flanking the 68 L1Hs elements revealed a number of interesting findings. The poly(A) tails of 25 L1s were interrupted or contained 3′ transductions (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000), 17 of which clustered into “subfamilies” of L1Hs elements. In one case, we identified an L1 (#2-1) as the likely source element for one of these subfamilies. For #1-3, 3-31, and 1-5, these transductions/interrupted poly(A) tails were identical to those in L1Hs elements that have caused disease-producing mutations (*e.g.,* $L1_{RP}$, LRE3) (Brouha et al., 2002; Kimberland et al., 1999). In other cases, the transductions denote examples of recently amplified subfamilies (Goodier et al., 2000; Lander et al., 2001; Pickeral et al., 2000).

Examining the 5′ genomic flanks showed that the retrotransposition of a full-length L1 from the ABC9 genomic library (#2-24) that integrated on chromosome

10 was accompanied by ~250 bp of an Alu element that maps to chromosome 16. The Alu sequence is in the opposite transcriptional orientation to the L1, 13 bp of unmapped sequence separates the elements, and the whole insertion is flanked by target-site duplications (TSDs) (Figure 2.8A). Thus, though most of the full-length L1Hs elements identified here have been amplified by canonical retrotransposition, recombination- and/or replication-mediated repair processes may facilitate the integration of some elements (Gilbert et al., 2002, 2005; Symer et al., 2002). Additionally, our screen allowed us to resolve possible sequence anomalies in the HGR. For example, one fosmid that lacks a dimorphic L1Hs element (#6-105) actually contains two L1s (a PA2 and pre-Ta element) that likely were collapsed into a harlequin element during the HGR assembly (Figure 2.8B).

Finally, the data also enabled us to examine allelic heterogeneity associated with L1Hs elements. For example, one L1 (#5-70) was present in the HGR, but contained a stop codon in ORF2 and was not previously tested for activity (Brouha et al., 2003). Interestingly, #5-70 retrotransposed at ~8% of the level of L1.3, further illustrating how allelic heterogeneity can impact retrotransposon activity (Lutz et al., 2003; Seleme et al., 2006).

Allele Frequencies of Genotyped Elements

The 68 L1Hs elements identified here are dimorphic with respect to presence; thus, we tested if a subset of these L1s represented population-restricted or potentially private alleles. To address this question, we first compiled existing genotyping data (Badge et al., 2003; Myers et al., 2002; Xing et al.,

2009). Additional genotyping then was conducted on a subset of the L1s discovered here (26 in total; see Extended Experimental Procedures for selection criteria). The 26 L1s first were genotyped in a CEPH panel of 129 unrelated individuals. Nine L1s absent from the CEPH panel then were genotyped in a Zimbabwean panel of 72 unrelated individuals. Finally, if the element was absent from both panels, it was genotyped on the H952 subset of the HGDP consisting of ~1050 individuals from ~51 worldwide populations (Figure 2.3A and Table 2.5) (Cann et al., 2002; Rosenberg, 2006).

Two elements (#3-5 and 3-31) genotyped on the HGDP exist at very low allele frequencies and were only found in Africans. Two other L1Hs elements (#1-5 and 3-24) were absent from the HGDP (Table 2.5). Element #3-24 (the S228P mutant described above) was found in the ABC10 Yoruban library. Further genotyping revealed that the L1Hs element containing the mutation was present in her mother (but not her father), excluding a *de novo* origin (Figure 2.3B). The other putatively "private" L1Hs element was from G248 (#1-5), so we could not examine its segregation in a trio. Interestingly, this hot L1 insertion occurred into an intron of the *ABCA1* gene (Figure 2.3C); mutations in *ABCA1* have been associated with Tangier disease and low serum HDL levels (Frikke-Schmidt, 2010).

The Total Number of Active L1Hs Elements Present in ABC13

To estimate the total number of active L1s in one individual, we carried out *in silico* genotyping of the 68 L1Hs elements in ABC13, the last library examined in our subtractive scheme. We identified 20 regions containing distinct L1

insertions identified in the first 5 individuals that corresponded to insertion fosmids in the ABC13 HGSV track (http://hgsv.washington.edu/) of the UCSC genome browser (Figure 2.4A, Table 2.5) (Kent et al., 2002; Kidd et al., 2008). PCR genotyping confirmed that ABC13 contained 18 of these 20 elements (Figure 2.4B), and was homozygous with respect to presence for 3 of the elements. This result suggests that *in silico* genotyping could be used as a screening tool to identify L1Hs elements present at low allele frequencies in the population (Table 2.5).

Adding the 18 L1Hs elements identified by *in silico* genotyping to the seven novel L1Hs elements identified in the ABC13 genome through our fosmid screens revealed that this individual contains 25/68 L1Hs elements identified in this study. Additional genotyping revealed that this individual contains 2 of the hot L1s characterized in a previous study (Table 2.1) (Brouha et al., 2003). Combining these numbers with our retrotransposition data indicates that the ABC13 genome contains 14 potentially hot L1Hs elements, and that at least 3 of these elements are present in a homozygous state.

Estimates of L1 Age

Our data suggest that, on average, the 68 L1Hs elements identified here are present at lower allele frequencies, are more active, and may be evolutionarily younger than those in previous studies (Brouha et al., 2003). To test this hypothesis, we derived maximum likelihood estimates for the ages of Ta-1 L1Hs elements in our dataset and that of Brouha *et al.* (Brouha et al., 2003; Marchani et al., 2009). This analysis revealed that the Ta-1 L1Hs elements

identified here are significantly younger (1.0 million years [MY] 95% confidence interval [C.I.] 0.98 – 1.01 MY) than those reported previously (2.01 MY 95% C.I. 2.00 – 2.02 MY) (Marchani et al., 2009) (1.73 MY 95% C.I. 1.69 – 1.77 MY) (Brouha et al., 2003).

The maximum likelihood estimated age (Marchani et al., 2009) (1.0 MY) of the L1s reported here differs significantly from that calculated using the *ad hoc* method, which uses sequence divergence within subfamilies of elements to determine age (Carroll et al., 2001) (1.18 MY old). These two methods are known to be respectively robust (the maximum likelihood method) and sensitive (the *ad hoc* method) to the presence of multiple active lineages in the dataset (*i.e.,* departures from the master gene model of L1 evolution) (Cordaux et al., 2004). The difference in these two estimates may indicate that members of multiple active L1Hs subfamilies are present in our dataset and suggests that the true age of the L1s may be younger than either calculation suggests. Indeed, the above data are consistent with the hypothesis that the HGR is strongly biased in favor of older, fixed L1Hs elements.

We next used a neighbor-joining approach, rooted with an intact chimpanzee L1 element, to generate a phylogenetic tree of the 68 full-length L1Hs elements (Figure 2.5, see Experimental Procedures). As predicted, pre-Ta elements were located near the root of the tree. Interestingly, two known ($L1_{RP}$ and LRE3) and five other currently amplifying subfamilies clustered together on the tree (Figure 2.5; see groups of colored elements), even though the

interrupted poly(A) tail/transduction sequences themselves were excluded from the sequence alignments.

Discussion

We have developed a systematic process to identify novel, dimorphic, active L1Hs elements in genomes of individuals from diverse geographic populations. Many of the newly identified L1Hs elements exist at low allele frequencies in the population and four L1Hs elements represent “rare” alleles, three of which appear to be restricted to Africans. Sequence-based age estimates further reveal that these L1Hs elements appear to be, on average, evolutionarily younger than those identified in previous studies (Brouha et al., 2003; Marchani et al., 2009). These data are consistent with the notion that full-length active L1s are systematically underrepresented in available genome reference sequences (Badge et al., 2003; Boissinot et al., 2004; Brouha et al., 2003; Sassaman et al., 1997; Sheen et al., 2000; Xing et al., 2009).

Our study has underscored the effectiveness of fosmid paired-end libraries in the discovery of novel, active L1Hs elements. Though a number of technologies have been developed to identify polymorphic L1s (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2004; Brouha et al., 2003; Moran et al., 1996; Myers et al., 2002; Sheen et al., 2000; Xing et al., 2009), the approach described here is not reliant upon PCR fidelity, readily allows the identification of active L1Hs elements, and makes sequencing of genomic flanking sequences, poly(A) tails, and L1-mediated transductions relatively straightforward. Thus, we predict that the fosmid-based approach likely will be superior to second-

generation, low-coverage genome sequencing methodologies (*e.g.*, many individual genomes characterized in the 1000 genomes project; http://www.1000genomes.org/page.php) for comprehensively identifying and characterizing rare L1 alleles in individual genomes. Indeed, recently published genome sequences highlight the difficulties in detecting and unambiguously mapping highly repetitive insertions (relative to a reference genome), including L1Hs elements (Bentley et al., 2008; McKernan et al., 2009; Wang et al., 2008; Wheeler et al., 2008).

Our analysis revealed that many active L1s cluster in small subfamilies. In the strictest sense, these data argue against a master gene model (Deininger et al., 1992) and instead support a model in which multiple active source L1Hs elements (including members of both the pre-Ta and Ta subfamilies) are currently retrotransposing in modern human genomes (Cordaux et al., 2004). We cannot formally exclude a "stealth" model, where L1s in unfavorable expression contexts sometimes give rise to new retrotransposition-competent source elements that can be expressed from a more favorable genomic context (Han et al., 2005). However, the most parsimonious explanation of our data is that multiple source L1Hs elements and subfamilies with limited "life-spans" exist in the genome. We posit that hot L1Hs elements must give rise to new, active progeny at a faster rate than they are inactivated by cellular mutational processes (see Figure 2.6 for model); this can lead to a scenario where small numbers of currently active L1Hs lineages may out-compete older L1s for limiting reagents, such as host factors (Boissinot and Furano, 2001). This competition

scenario both supports and extends current lineage succession models and could potentially explain the monophyletic history of L1s and the appearance of a replication-dominant L1Hs subfamily (Boissinot et al., 2000; Cordaux et al., 2004; Seleme et al., 2006).

Our dataset is still relatively small, and it remains difficult to estimate the actual number of highly active L1s in the extant population. However, our ability to readily identify rare hot L1s in the genomes of geographically diverse individuals strongly suggests that these highly active L1Hs elements are more abundant in the population than previously appreciated. Indeed, these results are in general agreement with recently published studies (Iskow et al., 2010; Huang et al., 2010).

The active L1Hs elements identified here also have the potential to impact modern human genomes by retrotransposing flanking genomic sequences to new chromosomal locations and by serving as substrates for nonallelic homologous recombination (reviewed in Cordaux and Batzer, 2009; Moran et al., 1999). The proteins encoded by these L1s also may promote the retrotransposition of Alu elements and noncoding RNAs (Bennett et al., 2008; Dewannieux et al., 2003; Garcia-Perez et al., 2007). Indeed, our data support the hypothesis that hot L1s are actively retrotransposing in modern-day human genomes and suggest that some of the L1 alleles identified here could serve as source elements for disease-producing L1 insertions.

Experimental Procedures

Creation of Fosmid Libraries and Identification of Insertion-Containing Fosmids

Genomic DNA from the six individuals was obtained from transformed lymphoblastoid cell lines (available from the Coriell Cell Repository). The DNA was hydrodynamically sheared, end-repaired, size selected for 40 kb fragments by pulsed field gel electrophoresis, and ligated into fosmid vectors (Donahue and Ebling, 2007). Agencourt Biosciences Corporation constructed all libraries, with the exception of the G248 library, which was constructed as part of the human genome project finishing effort. From each library, approximately 1 million individual cloned fragments were arrayed into 384-well plates. End-sequence pairs were obtained from both ends of each DNA fragment using standard capillary sequencing and were mapped back to the HGR. Insertion-containing fosmids were identified as the subset of fosmids containing an apparent insert that was ~3 standard deviations smaller than the library mean (Kidd et al., 2008; Tuzun et al., 2005).

Screening of Fosmid Clones for LINE-1 Insertions

Insertion-containing fosmids identified *in silico* were screened for L1Hs elements in the following manner. First, all insertion fosmids were subjected to allele-specific oligonucleotide hybridization to identify characters in the 5′ UTRs of newer L1 subfamilies (Badge et al., 2003; Boissinot et al., 2000). This protocol was adapted from “hybridization of bacterial DNA on filters” (Sambrook, 1989). Fosmid DNAs were prepared according to the Very Low-Copy Plasmid/Cosmid Purification protocol for the Qiagen-tip 100 Midi prep kit (Qiagen). Those DNAs

were subjected to Southern blotting followed by ATLAS (Badge et al., 2003) and/or direct sequencing to identify L1Hs elements that were absent from the HGR. Sequences flanking the L1Hs elements then were used as probes in BLAT searches at the UCSC genome browser (http://genome.ucsc.edu/) to determine the insertion site in the HGR (Kent, 2002; Kent et al., 2002). Detailed protocols for each step of the screening process, as well as the number of fosmids positive at each stage of the analysis, can be found in the Extended Experimental Procedures.

Cloning of L1s

In general, L1Hs elements were cloned directly from insertion-containing fosmids by digestion with *Acc*I (Sassaman et al., 1997). The restricted DNA was separated on a 0.8% agarose gel, and the ~6 kb L1-containing restriction fragment was cloned into an L1 expression vector. This method captures the vast majority of the L1Hs sequence, leaving only the first ~35 bp and last ~50 bp of the original L1 5′ and 3′ UTRs present in the cloning vector, respectively. One element, #2-42, was refractory to this cloning procedure, as it contains a polymorphism near the 3′ end of ORF2 that creates an additional *Acc*I site. The PDH L1.3 mutant was generated by site-directed mutagenesis. Each L1Hs element was sequenced in its entirety. Detailed protocols for the creation of each construct are included in the Extended Experimental Procedures.

L1 Retrotransposition Assays

We used a modification of a transient transfection protocol to conduct retrotransposition assays in HeLa and CHO cells (Moran et al., 1996; Morrish et

al., 2002; Wei et al., 2000). Briefly, cells in 6-well dishes were transfected using the Fugene 6 agent (Roche) with 1 µg of plasmid (containing the indicator cassette) per each well. Cells were fed with media ~24 hours post-plating, and daily from 72 hr or 5 days with media containing either 400 µg/mL G418 or 10 µg/mL blasticidin, respectively. Fourteen days post-transfection, cells were fixed and stained with 0.1% crystal violet. Colonies were counted in the appropriate wells, and these counts were normalized to green fluorescent protein (GFP) transfection efficiency. Detailed protocols for culture and assay conditions are found in the Extended Experimental Procedures.

Genotyping and Panels

The genomic locations of L1Hs insertions were compared to a database of human retrotransposon insertion polymorphisms (dbRIP; http://dbrip.brocku.ca/) (Wang et al., 2006). PCR genotyping assays were designed for a subset of L1Hs elements that were not completely annotated in dbRIP. Genotyping initially was conducted on a CEPH panel of 129 unrelated individuals of Northern European ancestry. If an L1Hs element was absent from the CEPH panel, it was genotyped on a panel containing genomic DNAs from 72 unrelated Zimbabwean individuals. Finally, if an L1Hs element was absent from both genotyping panels, it was genotyped on the H952 subset (Rosenberg, 2006) of the HGDP (Cann et al., 2002) (see Figure 2.3A). *In silico* genotyping was conducted using the HGSV track of the UCSC genome browser (Kent et al., 2002; Kidd et al., 2008). Details about these analyses are in the Extended Experimental Procedures.

Estimation of L1 Element Age

Sequences of the 69 full-length L1 elements were classified into subfamilies using the L1Xplorer analysis website (Penzkofer et al., 2005). Ta-1, Ta-0 and Non-Canonical (NC) (Brouha et al., 2003) elements were separately aligned using Muscle 3.52 (Edgar, 2004) on the Phylemon web server (http://phylemon.bioinfo.cipf.es/cgi-bin/home.cgi) (Tarraga et al., 2007). Raw alignments were manually refined using Jalview to remove all indels, all variable CpG sites, and the L1 polypurine tract (Waterhouse et al., 2009). Maximum likelihood estimates of the age (T) of each group, the sampling variance of T, and its 95% C.I. were calculated using the mleT script (Marchani et al., 2009) running under Matlab 7.2 -2007a (The Mathworks Inc., Natick, MA, USA). The subroutine CountMutations (Marchani et al., 2009) was also utilized to calculate the number of substitutions in the datasets to enable the "ad hoc" subfamily age estimation method (Marchani et al., 2009).

Phylogenetic Tree

The sequences of the 69 elements were aligned as described above. An intact chimpanzee element (BS000022_PTROG) was used to root the tree. The alignment also includes an intact Ta-1 L1 (L19088_L1.3), a non-Ta L1 (AL022171_NTA), a pre-Ta L1 (AL357559), and the "Hot Consensus" L1 element from Brouha et al. (2003). Raw alignments were manually refined using Jalview (Waterhouse et al., 2009) to remove large indels and truncated elements; this led to the exclusion of #6-113 due to a large 5′ UTR deletion.

A single neighbor-joining tree of the 68 remaining full-length elements was constructed using the PHYLIP package (Felsenstein, 1989). Branch lengths were corrected using the Kimura 2 parameter model (Kimura, 1980). To assess the reliability of the phylogeny, 1000 bootstrapped resamples of the multiple alignment were made using the seqboot program of the PHYLIP package (Felsenstein, 1989). The neighbor joining tree derived from the full dataset was manually annotated with bootstrap values using Dendroscope (Huson et al., 2007) (Figure 2.5). Only bifurcations that occurred in more than 70% of bootstrap resamples are labeled.

Accession Numbers

Accession numbers for all elements are tabulated in Table 2.6. Two L1Hs elements (accession numbers (#1-5) GU477636 and (#6-102) GU477637) were recently posted in GenBank.

Extended Experimental Procedures

Identification of Insertion-Containing Fosmids

Paired-end sequence analysis was carried out as previously described (Kidd et al., 2008; Tuzun et al., 2005). Briefly, end-sequence pairs were mapped against the HGR sequence (NCBI build 35/hg17) using megaBLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul et al., 1990). Any clone having an apparent insert size (based on the distance between the mapped positions of the reads) approximately 3 standard deviations smaller than the library mean was considered to represent a potential insertion event (Kidd et al., 2008; Tuzun

et al., 2005). We required that each identified variant be supported by at least two clones from the same individual. Since the individual clones from each library are retained, the identified cloned segments could be directly retrieved for further analysis. For this study, a reduced size threshold (approximately 2 standard deviations) was employed to screen the ABC9 library (Table 2.1). This method provides us with a minimal estimate of the fosmids that may contain L1 insertions.

Allele-Specific Oligonucleotide Hybridization

Insertion fosmid containing bacteria were spotted onto Amersham Hybond-XL nylon membrane (GE healthcare) and grown on 12.5μg/mL chloramphenicol plates over two nights (~40 hours) at room temperature. Bacteria then were lysed and the DNA denatured and fixed to the membrane using standard methods. Fosmids were screened with two γ-^{32}P end-labeled oligonucleotides specific for the 5′ UTR of full-length, recently inserted L1s (oligonucleotide sequences available upon request) (Badge et al., 2003). The protocol was adapted from 'hybridization of bacterial DNA on filters' (Sambrook, 1989).

Fosmid DNA Preparation

Insertion-containing fosmids that scored positive by oligonucleotide hybridization were shipped to the Badge and Moran labs as LB media stabs, and subsequently were grown and archived as glycerol stocks. Fosmid-containing bacterial cultures were grown in 5-10mL (mini-prep) or 300-500 mL (midi-prep) of LB liquid media containing 12.5 μg/mL chloramphenicol overnight (~16 hours) at

37°C. Minipreps were carried out using the standard protocol for the Wizard Plus SV DNA Purification System (Promega) and DNA was eluted in 60 µl of DNAse/RNAse free water (Gibco). Midi-preps were prepared according to the Very Low-Copy Plasmid/Cosmid Purification protocol for the Qiagen-tip 100 Midi prep kit (Qiagen). Midi-prep DNAs were reconstituted in 150 µl of DNAse/RNAse free water (Gibco).

Southern Blots

Fosmid DNA mini-preps (50 µl, ~1 µg) were digested overnight at 37°C with *Acc*I (New England Biolabs-NEB), which is predicted to cut L1.3 at bp positions 41 and 5965 ((Accession number: L19088 (Sassaman et al., 1997)). Digests were fractionated on 0.7% agarose gels containing ethidium bromide (0.5 µg/ml), and digital images were recorded as a fingerprint of each fosmid. Southern blotting was carried out using an adaptation of a standard protocol (Sambrook, 1989). Briefly, gels were treated to depurinate (0.25M HCl) and denature (0.5M NaOH/1.5M NaCl) the DNA. They then were treated with neutralization solution (1.5M NaCl/0.5M TrisHCl) and the DNA was transferred to an Amersham Hybond-XL Nylon membrane by capillary action. Membranes were baked at 80°C for 1.5 - 2 hours to fix the DNA. Probe DNA was created from the *Not*I to *Bgl*II (NEB) fragment of JM101/L1.3 (bp 1-661 of the 5′UTR of L1.3 (Sassaman et al., 1997)), and radio-labeled using [α-^{32}P] dCTP (GE healthcare) with the Rediprime II labeling kit (Amersham / GE healthcare). Excess [α-^{32}P] dCTP was removed using a G-50 MicroSpin column (Amersham / GE healthcare).

ATLAS

Putative full-length L1-containing fosmids were screened for newer elements using a modification of the previously described technique ATLAS (amplification typing of L1 active subfamilies) (Badge et al., 2003). Briefly, 5 µl (~100 ng) of fosmid DNA minipreps were digested overnight at 37^{o}C with *Mse*I, *Taq*I, or *Nla*III (New England Biolabs-NEB), followed by restriction enzyme inactivation at 65^{o}C for 20 minutes (*Mse*I and *Nla*III only). Digested fosmid DNA was ligated to annealed ATLAS linkers with appropriate terminal sequences, as described previously (Badge et al., 2003). After ligase inactivation (20 minutes at 65^{o}C), DNA was added directly to L1-to-linker amplification reactions using either 5′ UTR specific (RB5PA2 and RBX4) or 3′ UTR specific (RB3PA2 and RBX4) PCR primers (sequences below). Control reactions performed in the absence of annealed linker or ligase, reactions lacking digested fosmid DNA, or reactions lacking fosmid DNA were used to ensure amplification was specific for reactions containing all components and that amplicons were derived from linkered fosmid DNA. Amplification reactions were fractionated on 2% agarose gels containing 0.5µg/ml ethidium bromide. Fosmid-specific amplicons were excised, the DNA purified using the Qiagen Gel Extraction Kit (Qiagen UK, Crawley UK)), cloned into pGEMT-Easy (PromegaUK, Southampton, UK), and sequenced using primers flanking the vector multiple cloning site. Sequencing was carried out using an Applied Biosystems 3730 sequencer by the University of Leicester PNACL core facility. Sequences flanking the L1 elements then were used to map the insertion point in the human genome using BLAT (Kent, 2002) at UCSC (Kent

et al., 2002) (http://genome.ucsc.edu/) as discussed below. Insertion fosmids from G248, ABC9, and ABC10 were screened using both 5′ UTR and 3′ UTR specific ATLAS.

Sequencing and Analysis

Fosmid midi-prep DNA was sequenced from the L1 3′ and 5′ UTRs into flanking regions using the HS or ORF2L oligos, and the RB5PA2 or 5′ UTR AS oligos, respectively. The University of Michigan Sequencing Core Facilities performed DNA sequencing using an Applied Biosystems ABI Model 3730XL sequencer.

L1 insertion locations were determined by comparison of 5′ and 3′ flanking sequence of the L1s to the HGR (NCBI build 36.1/hg18) using BLAT (BLAST-like alignment tool- http://genome.ucsc.edu/cgi-bin/hgBlat) (Kent, 2002), and the presence or absence of the element in the HGR was determined using the UCSC genome browser (http://genome.ucsc.edu/) (Kent et al., 2002). Where the flanking sequence was too short to enable precise mapping, the region of genomic sequence corresponding to the region between the mapped end sequences of the fosmid was downloaded from UCSC and the best sequence match within this region was identified as the most likely insertion point using BLAT (Kent, 2002). Sequences of the distal 5′ and 3′ UTRs from the new L1s were ascertained from direct sequencing of the fosmid or were obtained from fosmid sequences deposited in online databases (NCBI) by the Washington University Genome Sequencing Center (St. Louis). Accession numbers for fosmid clones sequenced in their entirety (67/69), as well as those created for the

two L1 elements that lack sequenced fosmids (#1-5 and 6-102) are reported in Table 2.6.

Screening of Fosmid Clones

Insertions in two or more clones for a particular region were scored as an insertion-containing region in the fosmid genomic library of an individual. For the G248 library, these clones were subjected to hybridization and further analysis. For all other libraries (ABC9-13), only one insertion-containing clone per region was hybridized and subjected to the following steps in our analysis scheme. Additionally, regions found in previous libraries were excluded from analysis in subsequent individuals, therefore generating a subtractive set of L1 insertions where there is minimal overlap of regions between the 6 examined genomes.

G248

In total, 108 predicted insertions for the G248 library were identified by paired-end sequence analysis, and 32/108 insertion fosmids were positive by oligonucleotide hybridization. Thirteen of the 32 fosmids were positive by Southern blot, and 8 of these 13 were positive by both 5′ and 3′ ATLAS. Five fosmids that were positive by Southern blot and negative by ATLAS were examined and found, in the case of 3 fosmids, to contain older, full-length elements from the L1PA3-4 families already in the HGR. One more of these 'false positive' fosmids yielded equivocal results using primers at both the 5′ and 3′ ends. Sequencing verified that 7/8 fosmids that were positive for ATLAS were dimorphic. Of these 7, 3 were duplicates of a given L1. One allele of each element (4 in total) was cloned and tested. Three of these 4 elements were

active. One more element from the G248 library was identified in fosmid sequencing data from the Eichler lab. This element was not included in the 32 original clones. The fosmid was acquired, and the element (G248 #1-(2-1)) was cloned and found to be active.

ABC9

The ABC9 library contained 186 fosmids with predicted insertions, 64 were positive by oligonucleotide hybridization and 38 of the 64 were positive by Southern blot. Fourteen of these 38 were positive by 5′ and 3′ ATLAS, 7 were positive by 5′ ATLAS only, and 4 were positive by 3′ ATLAS only. The 14 plus the four 3′ ATLAS only fosmids were sequenced and cloned. Sixteen were absent from the HGR, and 9 were active.

ABC10

In the ABC10 library, 297 fosmids contained predicted insertions and 46 of the 297 fosmids were positive by oligonucleotide hybridization. Of the 46, 37 were positive by Southern blot, 14/37 were both 5′ and 3′ ATLAS positive, 7/37 were only 5′ ATLAS positive, and 3/37 were only 3′ ATLAS positive. The 14 dual positives, the 3 that only were 3′ ATLAS positive, and 3 fosmids that were only 5′ ATLAS positive were sequenced and cloned. Nineteen of these elements were dimorphic, and of them, 10 were active. One more L1 (#3-(2-1)) was identified via sequencing by the Eichler laboratory. The fosmid was obtained, and the L1 was shown to be active.

ABC11

Of 246 fosmids containing predicted insertions, 35 were positive by oligonucleotide hybridization. Of the 35, 22 were full-length by Southern blot, 13 of these 22 were dimorphic by sequencing, and, of these, 9 were active in the retrotransposition assay.

ABC12

Of 258 fosmids containing predicted insertions, 52 were positive by oligonucleotide hybridization. Of the 52, 18 were full-length by Southern blot, 8 of these 18 were dimorphic, and, of these, 4 were active in the retrotransposition assay.

ABC13

Of the 265 fosmids containing predicted insertions, 29 were positive by oligonucleotide hybridization. Of these 29, 17 were full-length by Southern blot, 7 of the 17 were dimorphic, and, of these, 6 were active in the retrotransposition assay.

Cloning of L1s

L1s were cloned directly from fosmids into the context of the *mneoI* retrotransposition indicator cassette without the use of PCR. All pBluescript (Stratagene) vectors used in this study contain the 5′ UTR of an L1 and lack an exogenous promoter; the pCEP4 (Invitrogen) vectors have the cytomegalovirus (CMV) immediate early promoter, the 5′ UTR, and contain the *mneoI* retrotransposition indicator cassette.

G248

Three G248 fosmids were digested with *Acc*I (NEB) and ligated (T4 DNA ligase NEB, overnight at 16^{o}C) into the context of a pBluescript vector (pD100), containing the entire L1.3 element with a T7 *gene 10* epitope tag on ORF1p (#1-2, 1-4, and 1-5) (Kulpa and Moran, 2005). The ~6kb *Acc*I restriction fragment from element #1-2-1 was cloned into JCC9/L1.3 RT-, a pBluescript based plasmid containing both L1.3 with a D702A mutation in the reverse transcriptase active site and *mneoI* cassette in the 3′ UTR (Moran et al., 1996; Morrish et al., 2002). The first ~35bp and the last ~50bp of the new L1s are thereby replaced with those regions of the known active element L1.3 due to the location of the *Acc*I sites. Inserts were then verified by diagnostic restriction digest and sequencing to identify polymorphisms with respect to L1.3. Due to cloning difficulties, #1-3 was cloned into this vector by digestion with *Kas*I and *Nco*I, replacing the first 439bp and the last 346bp of the element with L1.3 sequences. These regions contained differences in the 5′ UTR as well as the 3′ end of ORF2 and the 3′ UTR with respect to L1.3. Both fosmid and cloned #1-3 elements have two intact open reading frames; however, there are two non-synonymous amino acid changes between the fosmid sequence and the clone. A pBluescript, JCC9-based clone of #1-3, generated using an *Acc*I digest, lacks these sequence changes, and retrotranspose at a level similar to L1.3 (data not shown). All five G248 elements then were cloned into JM105/L1.3 (pCEP4 backbone with an L1.3 RT- element), tagging their 3′ UTRs with the *mneoI* cassette. This was done through the use of *Not*I and *BstZ17*I (NEB), which cut at either the 5′ end (*Not*I)

or within the 3′ UTR (*BstZ17*I) of the L1 in pBluescript. Once again, DNA sequencing was performed to validate distinguishing polymorphisms in the respective 5′ UTR sequences of each clone.

ABC9 and ABC10

Both ABC9 and ABC10 elements were first cloned via *Acc*I into JCC9/L1.2A (similar to JCC9 L1.3 RT- described above, but with L1.2A in the pBluescript based vector; the first ~35 and last ~50 bp are identical to L1.3) (Moran et al., 1996) or JCC9/L1.3/RT-. The one exception to this (#2-42) was unable to be cloned in this manner due to a nucleotide change producing a third *Acc*I site near the 3′ end of ORF2p. This change was verified by sequencing the fosmid insert. All elements, except #3-5, were then subcloned into JM105/L1.3 with the use of *Not*I and *BstZ17*I as described above. All steps were verified through diagnostic digests and sequencing to identify distinguishing polymorphisms in the elements.

ABC11, ABC12 and ABC13

ABC11, 12, and 13 elements were cloned via *Acc*I into JCC9/L1.3/RT-. These clones were subjected to restriction digests and DNA sequencing to verify diagnostic polymorphisms in the respective L1Hs element.

L1.3 PDH mutant

The #3-24 fosmid L1 element containing the PDH mutation was cloned from JM #3-24 construct into JJ105/L1.3 (identical to JM105/L1.3, except with an *mblastI* cassette in the 3′ UTR (Morrish et al., 2002)) directly from the JM backbone with the use of *Not*I and *BstZ17*I. Subsequently, L1.3 S228P was

created by a standard site-directed mutagenesis protocol. Briefly, linear PCR was conducted using 250ng of pBluescript backbone (JCC5) containing L1.3 (Sassaman et al., 1997) and both forward and reverse-complementary 49-mer primers (IDT, www.idtdna.com) containing the mutation in the center (bp 25) of each primer using 5U of Pfu Turbo polymerase (Stratagene), mixed dNTPs (10mM, Invitrogen) and 17 cycles with extension for 18 minutes (~2 minutes per kb). Amplified plasmids then were incubated with *Dpn*I to digest methylated DNA and then were transformed into *E. coli*. Ampicillin resistant clones were verified by diagnostic digest and sequencing, and the mutation was then cloned back into JCC5/L1.3 by use of an *Age*I to *Eco*RI digest that allowed a swap of the mutation region for bp 1896 to 3431 of L1.3. The resultant construct was sequenced across the mutation and restriction enzyme sites used to generate the clone, and the L1.3 PDH element was then transferred to JM and JJ pCEP4 based vectors using *Not*I and *BstZ17*I.

Cell Culture Conditions

The cell culture conditions used have been described previously (Moran et al., 1996; Morrish et al., 2002). Briefly, HeLa cells (ATCC) were grown at 37^{o}C in DMEM-high glucose media with 10% FBS, 20U/mL penicillin/streptomycin and 0.4 mM glutamine (HeLa complete media) (Gibco) in the presence of 7% CO_2 and 100% humidity (Moran et al., 1996). The 4364a cell line was derived from Chinese Hamster Ovary (CHO-K1) cells, and is auxotrophic with respect to proline and glycine (Morrish et al., 2002). XR-1 cells were derived from the 4364a cell line, lack the *XRCC4* gene product, and are deficient in the NHEJ pathway of

DNA repair (Morrish et al., 2002). These cells were maintained in DMEM-low glucose media with 10% FBS, 20U/mL penicillin/streptomycin and 0.4 mM glutamine plus non-essential amino acids (CHO complete media) (Gibco) in the presence of 7% CO_2 and 100% humidity (Morrish et al., 2002).

Transfection and Retrotransposition Assays

We used a modification of a transient transfection protocol (Wei et al., 2000). Approximately $2x10^3$, $2x10^4$, or $2x10^5$ cells per well of a 6-well plate were used in the transfection of HeLa cells. Three wells of each cell concentration were transfected with 1 µg of plasmid (containing the *mneoI* cassette) using Fugene 6 transfection reagent (Roche) and Opti-mem media (Gibco) ~24 hours post plating. Cells were fed ~18-24 hours post transfection with HeLa complete media (described above), and daily from 72 hours with complete media plus 400µg/mL G418 (Gibco). Fourteen days after transfection, cells were washed with phosphate buffered saline (PBS) (Gibco), fixed in a solution of 2% formaldehyde/0.2% glutaraldehyde in PBS, and stained with 0.1% crystal violet. Colonies of G418 resistant cells were counted on the $2x10^3$ or $2x10^4$ plates as necessary. Transfection efficiency was obtained for each plasmid by transfecting $2x10^4$ and $2x10^5$ cells with 0.5µg of both the construct of interest and pCEP/GFP (Alisch et al., 2006). FACS analysis for %GFP positive cells at 72 hours post transfection yielded the transfection efficiency for the construct of interest. L1 retrotransposition in CHO cells was conducted as previously described (blasticidin was used at a concentration of 10µg/mL) (Morrish et al., 2002). Percent retrotransposition is reported relative to the rate of L1.3, adjusting for

transfection efficiency. L1Hs elements were classified as 'hot' if they retrotransposed with a frequency greater than ~10% of L1.3. This activity reflects the retrotransposition efficiency of precursors of mutagenic insertions or full-length mutagenic insertions themselves (Brouha et al., 2003).

Genotyping

After determining L1 insertion locations, we then compared these elements to a database of human retrotransposon insertion polymorphisms (dbRIP; http://dbrip.brocku.ca/) (Wang et al., 2006). An element was determined to be absent from dbRIP if, by January of 2009, there was no completely annotated full-length element at a given LINE-1 insertion site (*i.e.*, with respect to sequence, size and known TSDs). Seven elements (#2-38, 2-6, 2-7, 3-17, 5-58, 5-66, 5-86) in our study had dbRIP entries nearby (within ~1kb), but were not definitively mapped/annotated. For a subset of 'novel' elements absent from dbRIP, PCR genotyping assays were designed. These novel elements were selected from the first three individuals on the basis of a unique genomic insertion location, presence of non-repetitive sequence on both the 5′ and 3′ flanking regions for primer design, and for the ability to detect a single amplicon of the correct size upon PCR.

Dimorphism was determined with primers 5′ and 3′ of the insertion to detect the empty site as well as the HS or RB3PA2 primers paired with the 3′ flanking DNA primer for the filled site (common primer sequences below) (Badge et al., 2003). When the 3′ filled site was refractory to genotyping, either due to a transduction or variant sequences in the 3′ forward primer of the L1 (as in pre-Ta

elements), these elements were typed for the 5′ filled site using RB5PA2 and a 5′ flanking sequence oligonucleotide. Many of the filled site genotyping reactions were carried out at both the 5′ and 3′ ends to control for potentially unidentified 3′ transductions (Badge et al., 2003).

Genotyping Panels

Genotyping was carried out on a CEPH panel of 129 unrelated individuals of Northern European ancestry, a panel of 72 unrelated Zimbabwean individuals, or the H952 subset (Rosenberg, 2006) of the HGDP (Cann et al., 2002) as shown in Figure 2.3B. The 129 unrelated individuals are a subset of the grandparents and parents from the CEPH Pedigree DNA resource, within which there are no individuals with previously established lineal relationships. The Zimbabwean panel is composed of genomic DNA from 72 unrelated anonymous native African male semen donors collected in Harare, Zimbabwe. The H952 subset of the HGDP comprises all individuals within the 1051 samples that likely share no closer relationships than first cousins, as determined by RELPAIR analysis of SNP genotyping data (Rosenberg, 2006). Two elements positive for 5′ but not 3′ ATLAS, and that were not tested for activity were also genotyped on these panels. Allele frequencies are reported in Table 2.5 across all panels genotyped (*e.g.*, the four elements typed on the HGDP were calculated as the number of L1-containing alleles over the number of alleles genotyped on CEPH, Zimbabwean, and HGDP panels). Additionally, if an element was previously described in dbRIP or an L1 polymorphism study not included in dbRIP, efforts were made to obtain genotyping data from these studies (see Table 2.5).

In Silico Genotyping

As the set of elements that we examined is subtractive, we used the human genome structural variation (HGSV: http://hgsv.washington.edu/) track (Kidd et al., 2008) of the UCSC browser (http://genome.ucsc.edu/) (Kent et al., 2002) to examine whether elements found in the first 5 individuals were present in the last person examined in our study (ABC13). This analysis was conducted by ascertaining whether the ABC13 genomic library contained 'insertion' clones where other genomes (G248, ABC9-12) contained a full-length L1 insertion with respect to the HGR (Figure 2.4A). Putative insertions in the ABC13 genome were tested by genotyping as described above. *In silico* genotyping was conducted for each of the 68 distinct L1Hs insertions in this study, and is detailed in Table 2.5.

Estimation of L1 Element Age

The determination of L1 element age using maximum likelihood and *ad hoc* methods are described in the Experimental Procedures section of the main text. While the *ad hoc* method (Marchani et al., 2009) for estimating L1 subfamily age through divergence from consensus is simple, the newer maximum likelihood measure enables the calculation of 95% confidence limits about the estimate. Analysis of 886bp from 191 Ta-1 L1 elements gave an estimated age for this subfamily of 2.01 MY (95% C.I. 2.00 – 2.02 MY) using an L1-specific substitution rate of 0.25% per million years, based on the human–orangutan divergence (Marchani et al., 2009). The 48 Ta-1 elements in the current dataset are significantly younger than most Ta-1 L1 elements, with an estimated age of 1.0 MY (95% C.I. 0.98 – 1.01 MY). Applying the same analysis to 37 previously

reported Ta-1 elements (Brouha et al., 2003) gives an estimated age of 1.73 MY (95% C.I. 1.69 – 1.77 MY).

Where multiple lineages of L1 elements within a subfamily are simultaneously active, the *ad hoc* estimate of element age will significantly deviate from the maximum likelihood estimate (although both overestimate element age under a transposon model). We compared the maximum likelihood age of the 48 Ta-1 elements reported here (1.0 MY (95% C.I. 0.98 – 1.01 MY)) with the *ad hoc* estimate of their age (1.18 MY), and found them to be significantly different.

Oligonucleotides

HS- 5′- ATACCTAATGCTAGATGACACA- 3′

ORF2L- 5′- ATAGCAAAGACTTGGAACCAACCC- 3′

RB5PA2- 5′- TGGAAATGCAGAAATCACCG- 3′

5′UTR AS- 5′- CAGGCAGGCCTCCTTGAGCTG- 3′

RB3PA2- 5′- ACCTAATGCTAGATGACACA- 3′

RBX4- 5′- GTGGCGGCCAGTATTC- 3′

The sequences of additional oligonucleotides (for genotyping assays, *etc*.) used in this study are available upon request.

Acknowledgments

This Chapter was previously published: 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, and is reproduced with the permission of Cell and Elsevier. I thank all of the additional authors for experimental contributions as well as their input on the manuscript. We thank Prof. Sir Alec Jeffreys FRS for access to CEPH and Zimbabwean DNA samples, and Prof. Mark Jobling for access to HGDP DNA samples. We thank Dr. Elizabeth Marchani for advice on maximum likelihood age estimates and Dr. José Luis Garcia-Perez for plasmid JJ105/L1.3. We thank Dr. Garcia-Perez and members of the Moran lab for helpful comments. C.R.B. was supported in part by NIH training grants T32GM7544 & T32000040. J.M.K. was supported by a National Science Foundation Graduate Research Fellowship. Work in the laboratory of E.E.E. was supported by grant HG004120. P.C. and C.M. were supported by a Wellcome Trust Project Grant (075163/Z/04/Z) to R.M.B. and Prof. Sir Alec Jeffreys FRS. J.V.M. is supported by NIH grants GM066695 and GM060518. The University of Michigan Cancer Center Support Grant (5P30CA46592) helped defray sequencing costs incurred in this study. J.V.M. and E.E.E. are Investigators of the Howard Hughes Medical Institute.

Figure 2.1: A Strategy for Identifying Dimorphic L1Hs Elements in Individual Human Genomes

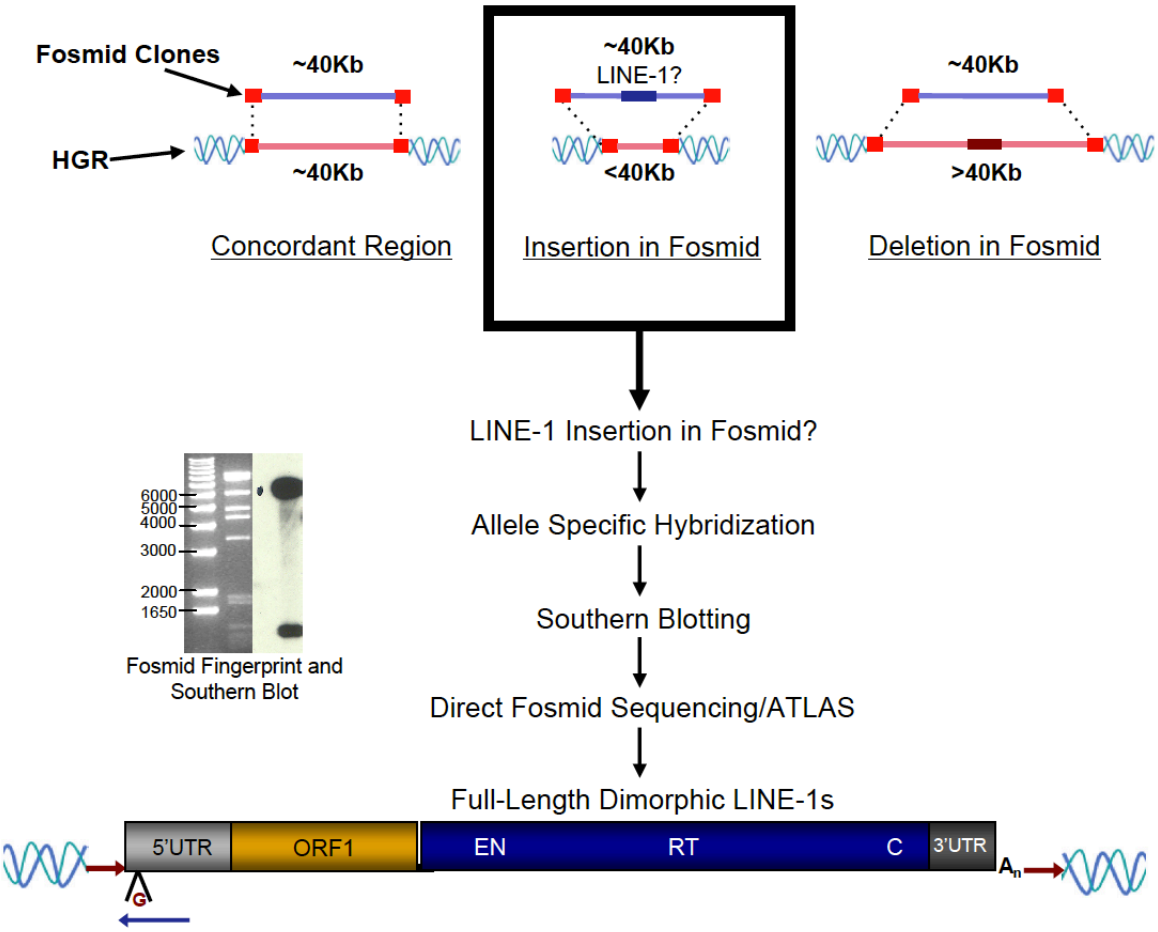


Figure 2.1: A Strategy for Identifying Dimorphic L1Hs Elements in Individual Human Genomes

In silico comparison of the fosmid end sequences (red squares) from individual genomic libraries (blue horizontal line) and the HGR (pink horizontal line) enables the detection of fosmids that may contain insertions or deletions with respect to the HGR (see dashed lines). Insertion fosmids were screened by allele specific oligonucleotide hybridization to detect characters that are present in the 5′ UTR of newer L1 elements (one discriminating character utilized, a deletion of the G residue at bp 74 in recent L1s, is indicated in maroon). Putative L1Hs-containing fosmids were analyzed by Southern blotting with a 5′ UTR probe (blue arrow). A representative digest and Southern blot is shown. The ~6 kb band is diagnostic for the full-length L1. The additional hybridizing band (~1.3 kb band liberated from the L1 5′ flank in this Southern blot example) serves to distinguish individual fosmids. ATLAS and/or DNA sequencing confirmed the presence of a dimorphic, full-length L1Hs insertion. The endonuclease (EN), reverse transcriptase (RT), and cysteine-rich (C) domains of ORF2 (blue rectangle) are indicated.

Figure 2.2: L1Hs Activity in Six Human Genomes

(A) Cloning strategy: All but one L1Hs element were cloned directly from fosmids using *Acc*I sites in their 5′ UTR and 3′ UTRs, respectively (red vertical lines; see Experimental Procedures). The L1s then were ligated into vectors that either contain or lack a CMV promoter (black rectangle). Both vectors contain the *mneoI* retrotransposition indicator cassette (light blue) in the L1 3′ UTR. This cassette allows for detection of retrotransposition events in a cell culture retrotransposition assay. SD = splice donor. SA = splice acceptor. Active elements confer G418 resistance to HeLa cells, whereas defective elements, as illustrated by the RT mutant control (RT- L1), do not. **(B)** Representative G418-resistant foci for the 20 elements from the Yoruban library, ABC10: Nine of these elements were highly active (large suns to the left of assay image), and two more retained a low level of activity (small suns). One element (#3-5, red box) is a hot pre-Ta L1 (#3-5 was tested in a pBluescript backbone (5′ UTR+); all others were tested in a pCEP4 (CMV+/5′ UTR+)) backbone (Extended Experimental Procedures). Table 2.2 displays retrotransposition efficiencies for each L1 identified in this study. Figure 2.7 provides details on the EN-deficient element #3-24. **(C)** The 68 distinct L1Hs elements identified in this study and their positions in the genome: Red vertical lines and text represent hot or highly active elements. Orange vertical lines with black text represent low-level activity elements. Blue vertical lines with black text represent 'dead' or inactive elements. The black line indicates the one untested element (#2-42). Ideograms were adapted from UCSC genome browser: http://genome.ucsc.edu (Kent et al., 2002).

Figure 2.2: L1Hs Activity in Six Human Genomes

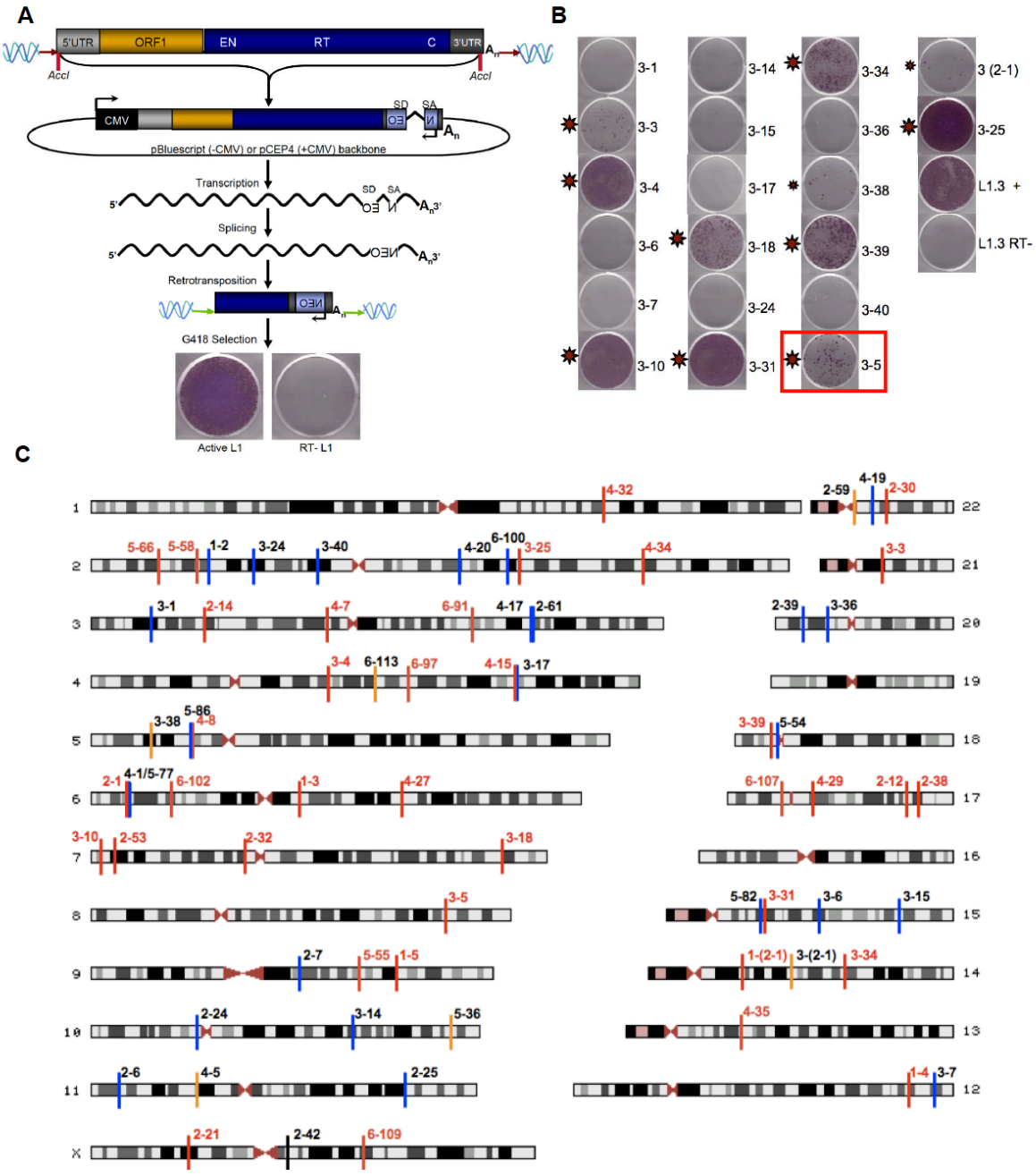


Figure 2.3: Allele Frequencies of L1Hs Elements in the Population

(A) Genotyping assays: L1s were queried in panels of individuals for their absence (solid grey lines), or presence (red line). Genotyping of 26 elements in the three panels allowed the discovery of population restricted or potentially "private" L1Hs elements. The expected amplicon sizes are diagrammed for element #3-24. **(B)** Pedigrees showing the inheritance of two elements typed in the ABC10 trio: Genotyping gels show the heritability of #3-31 (African specific) and #3-24 (absent from the HGDP). E and F at the top of the gel image indicate PCR results for empty and filled sites. M, F, and C at the bottom of the image indicate lanes for the mother, father, and child of the trio. **(C)** Example data sheet for the G248 element #1-5: Empty site: insertion site in the HGR. EN cleavage site: the endonucleolytic cleavage site used by L1 EN to initiate retrotransposition. pA length: the approximate L1 poly(A) tail length; 3′ transductions and interrupted poly(A) tails also are annotated. TSD length: the length of the target site duplication flanking the L1Hs element (underlined lettering). Table 2.3 contains data sheets for each L1 in this study. Table 2.4 contains L1Hs insertion locations with respect to genes. Figure 2.8 displays a non-canonical L1Hs insertion and documents a possible sequence anomaly in the HGR.

Figure 2.3: Allele Frequencies of L1Hs Elements in the Population

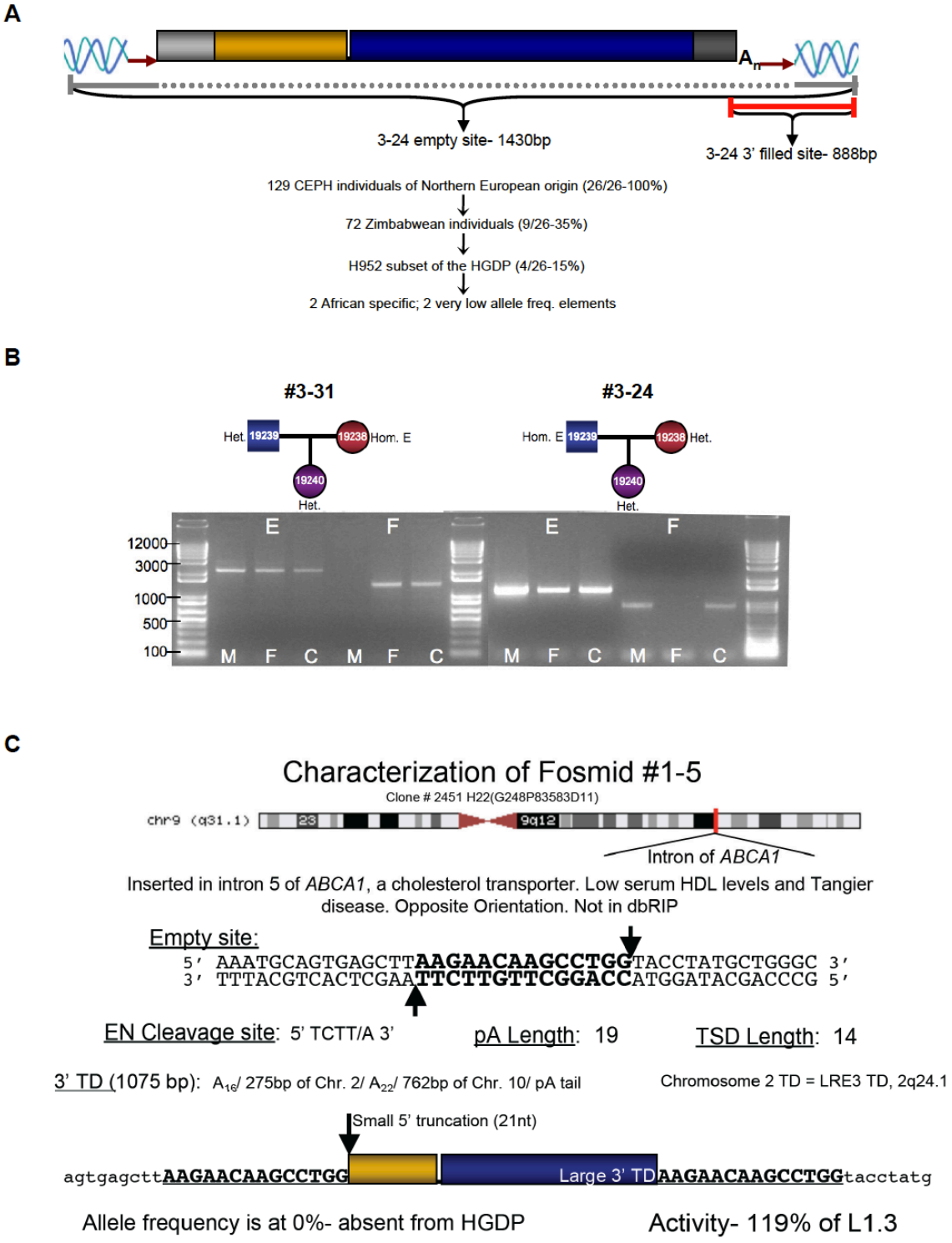


Figure 2.4: An Estimate of the Number of Active L1Hs Elements in an Individual (ABC13) Genome

(A) *In silico* genotyping: The last library in our study, ABC13, was examined *in silico* (see text) for the presence of insertion fosmids mapping to the location of L1Hs elements found in other individuals. Element 3-17 is used as an example. All blue lines represent insertion fosmids in the genomes of the 8 individuals on the HGSV track (http://hgsv.washington.edu/) of the UCSC genome browser (http://genome.ucsc.edu) (Kent et al., 2002). The ABC7, 8, and 14 libraries were not investigated in this study. **(B)** PCR validation: The elements identified *in silico* were genotyped using similar schemes to that shown in Figure 2.3A to validate the predictions from the HGSV track of the UCSC browser. Element 3-17 is used to illustrate the genotyping. ABC10 and ABC13 are heterozygous with respect to the L1Hs insertion. ABC11 lacks the L1Hs insertion. Table 2.5 displays genotyping results for all elements in this study.

Figure 2.4: An Estimate of the Number of Active L1Hs Elements in an Individual (ABC13) Genome

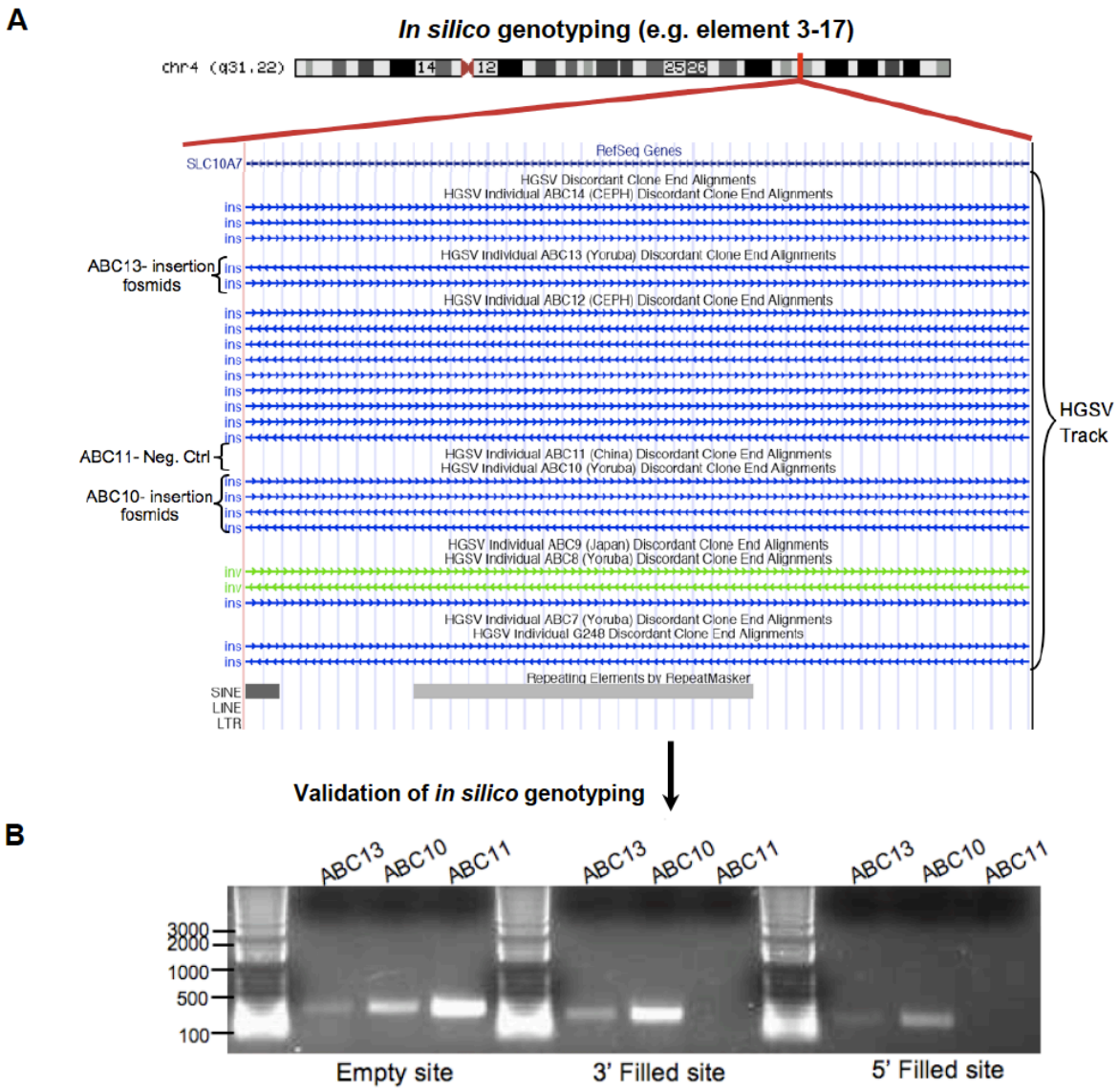


Figure 2.5: Phylogenetic Tree of the L1Hs Elements Identified in This Study

The tree is a single neighbor-joining tree (with branch lengths corrected using the Kimura 2 parameter model of nucleotide substitution) with 68 full-length elements from our study. The numbers at particular nodes indicate the number of times that node was observed in 1000 bootstrap replicates of the dataset. Only bootstrap values exceeding 70% are shown. The brackets at the right side indicate previously described 'transduction subfamilies' ($L1_{RP}$ (labeled RP in the Figure) and LRE3) and distinct L1Hs subfamilies currently capable of amplifying in human genomes (I-V) (Goodier et al., 2000; Pickeral et al., 2000). Those subfamilies are highlighted in the same color to show their clustering on the tree. Retrotransposition activity (% relative to L1.3) as well as allele frequency (*e.g.*, AF= 0.012), if determined, are appended to the sequence identifiers. Element #4-17 contains ACG characters in its 3′ UTR, which are diagnostic for pre-Ta L1s; however, the element clusters with the Ta0 subfamily. Activities for elements AL357559 and AL022171 were previously determined (Brouha et al., 2003). n/a = an L1 element not assayed for retrotransposition. The tree and age estimates use sequences indicated Table 2.6.

Figure 2.5: Phylogenetic Tree of the L1Hs Elements Identified in This Study

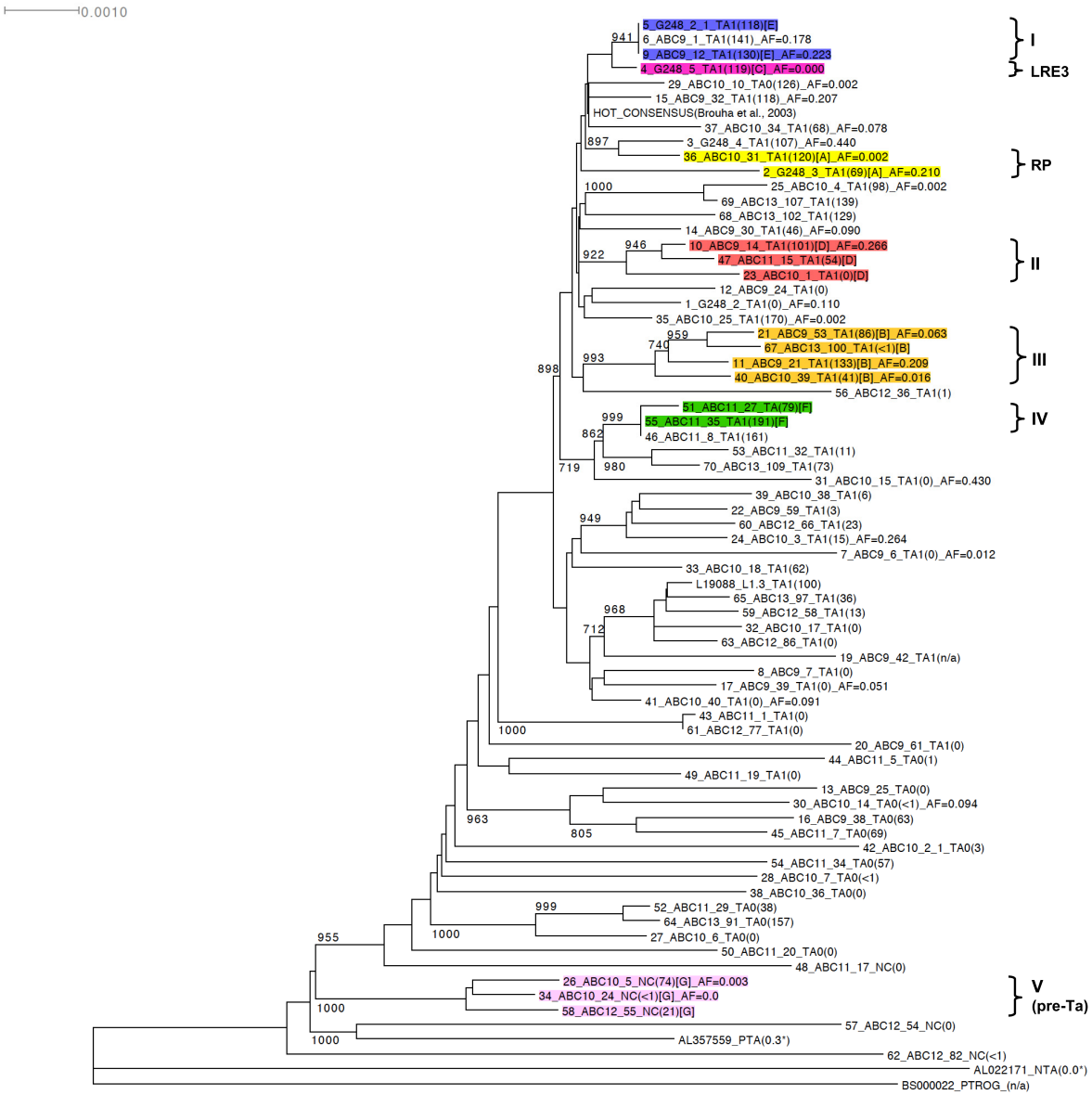


Figure 2.6: Multiple Source Loci Model for Continued L1Hs Activity

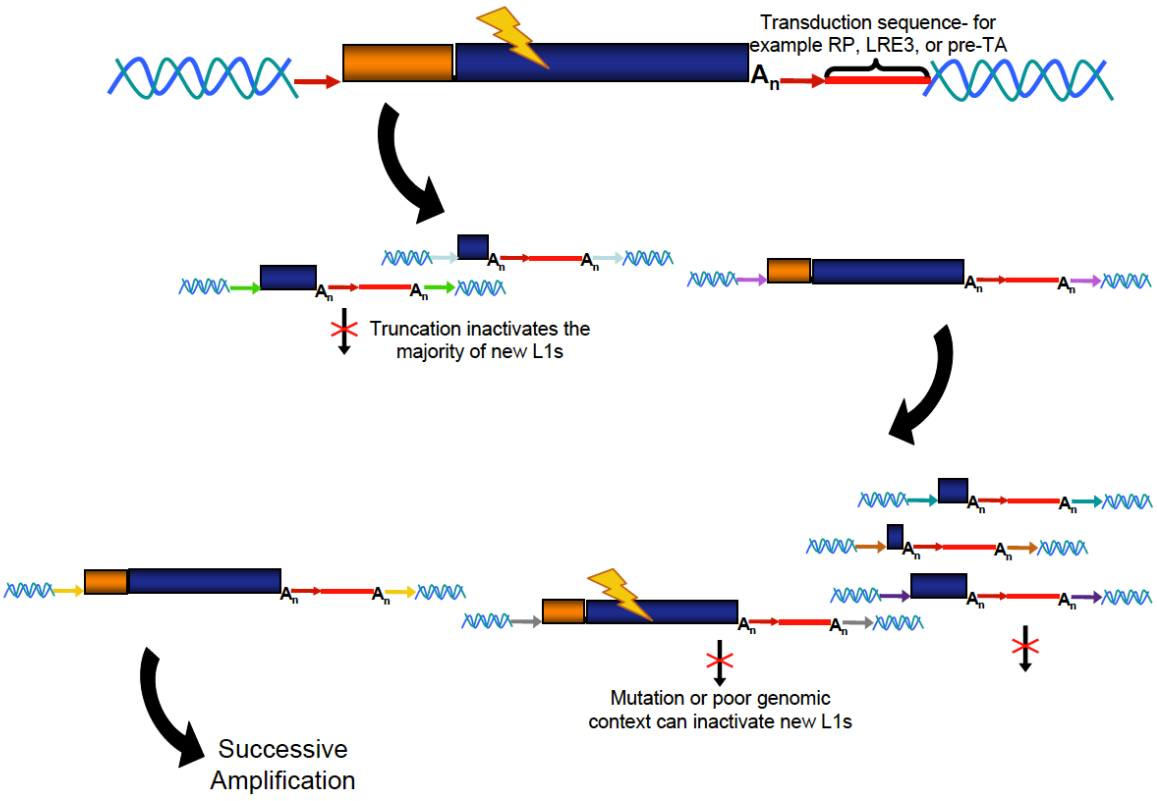


Figure 2.6: Multiple Source Loci Model for Continued L1Hs Activity

An element (source locus) that is both active and in a conducive genomic environment can retrotranspose. Shown here is an example of a progenitor element that can be associated with subsequent members of a family through the use of interrupted poly(A) tails and/or 3′ transduced sequence (3′ red arrow and line). Distinct elements are marked by distinguishing TSDs specific for their new integration site (different colored horizontal arrows). There are many of these 'families' active in human genomes, such as $L1_{RP}$, LRE3, and the 5 'families' noted in Figure 2.5. Although host processes (lightning bolt) may inactivate some older elements, some of their descendents may retain the ability to retrotranspose and could harbor the 3′ transduction/interrupted poly(A) tail.

Table 2.1: Summary of Data for the Six Libraries

Table 1. Summary of Data for the Six Libraries

Individual/Library Data						LINE-1 Data				
Library ID	Coriell ID	Population	Library Mean In Silico Insert Size	SD (kb)	Detection Limit (kb)	Dimorphic Elements	Novel (Not in dbRIP)	Active	Hot	HGR "Hot" Elements
G248	NA15510	N/A	39.89	2.75	8.25	5	5	4	4	2
ABC9	NA18956	Japan	39.51	2.26[b]	4.52[b]	16	16	9	8	2
ABC10	NA19240	Yoruba[a]	41	1.84	5.52	20	18	11	9	2
ABC11	NA18555	China	40.03	1.77	5.31	13	12	9	8	2
ABC12	NA12878	CEPH[a]	39.75	1.4	4.2	8	7	4	3	2
ABC13	NA19129	Yoruba[a]	39.29	1.77	5.31	7	7	6	5	2
Total						69/68[c]	65	43	37	

[a] Daughters of HapMap trios.
[b] Differs from Kidd et al (2008).
[c] One element was observed twice, in ABC11 &12- #4-1 and #5-77. Neither allele is active, and the element is in dbRIP.

Table 2.1: Summary of Data for the Six Libraries

Column 1: library identifiers. Column 2: Coriell identifier of individuals analyzed. Column 3: population of origin for individuals in the HapMap study. Column 4: the average insert size of each individual library (in kb). Column 5: the standard deviation in insert size of each individual library. Column 6: the detection limit for the size of insertions in each library. For ABC9 a more reduced threshold was applied than that used previously (Kidd et al., 2008). Column 7: the number of elements found in each library that are absent from the HGR. Column 8: the number of elements from column 7 that are not completely annotated in dbRIP (Wang et al., 2006). Column 9: the number of elements from column 7 that were active in retrotransposition assays. Column 10: elements from column 9 that retrotransposed at levels >10% of L1.3, a known active element. Column 11: The number of the HGR hot elements that were present in each individual (Brouha et al., 2003).

Figure 2.7: Endonuclease-Deficient Element #3-24, Related to Figure 2.2

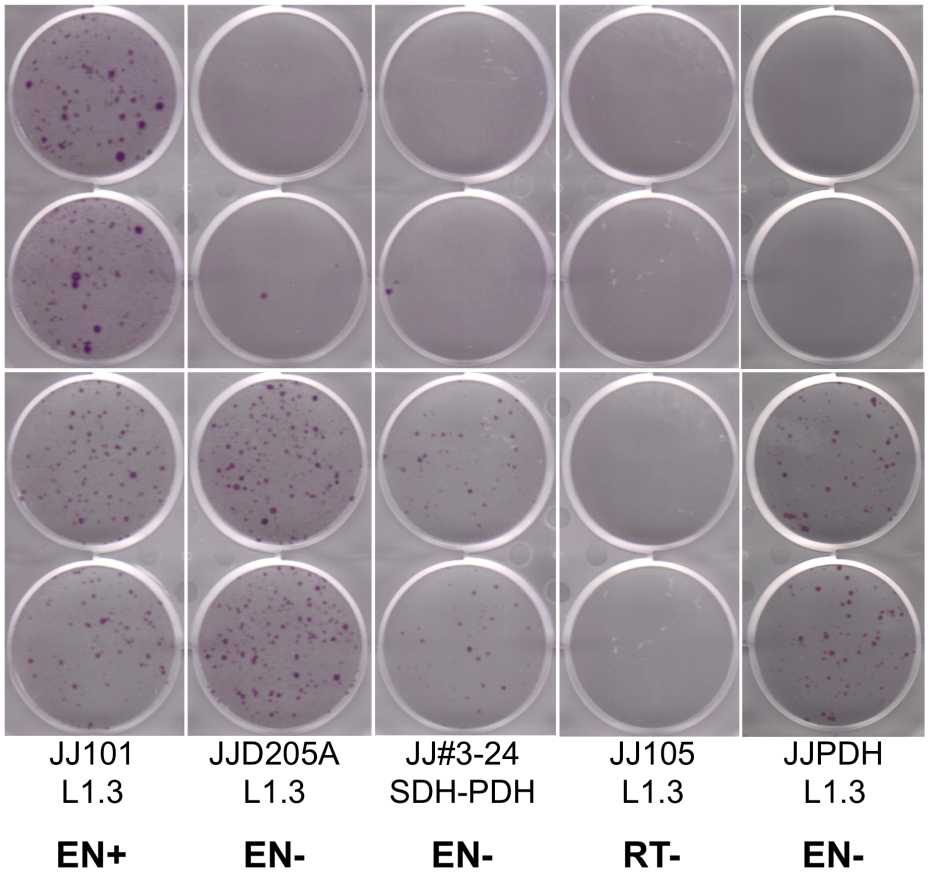


Figure 2.7: Endonuclease-Deficient Element #3-24, Related to Figure 2.2

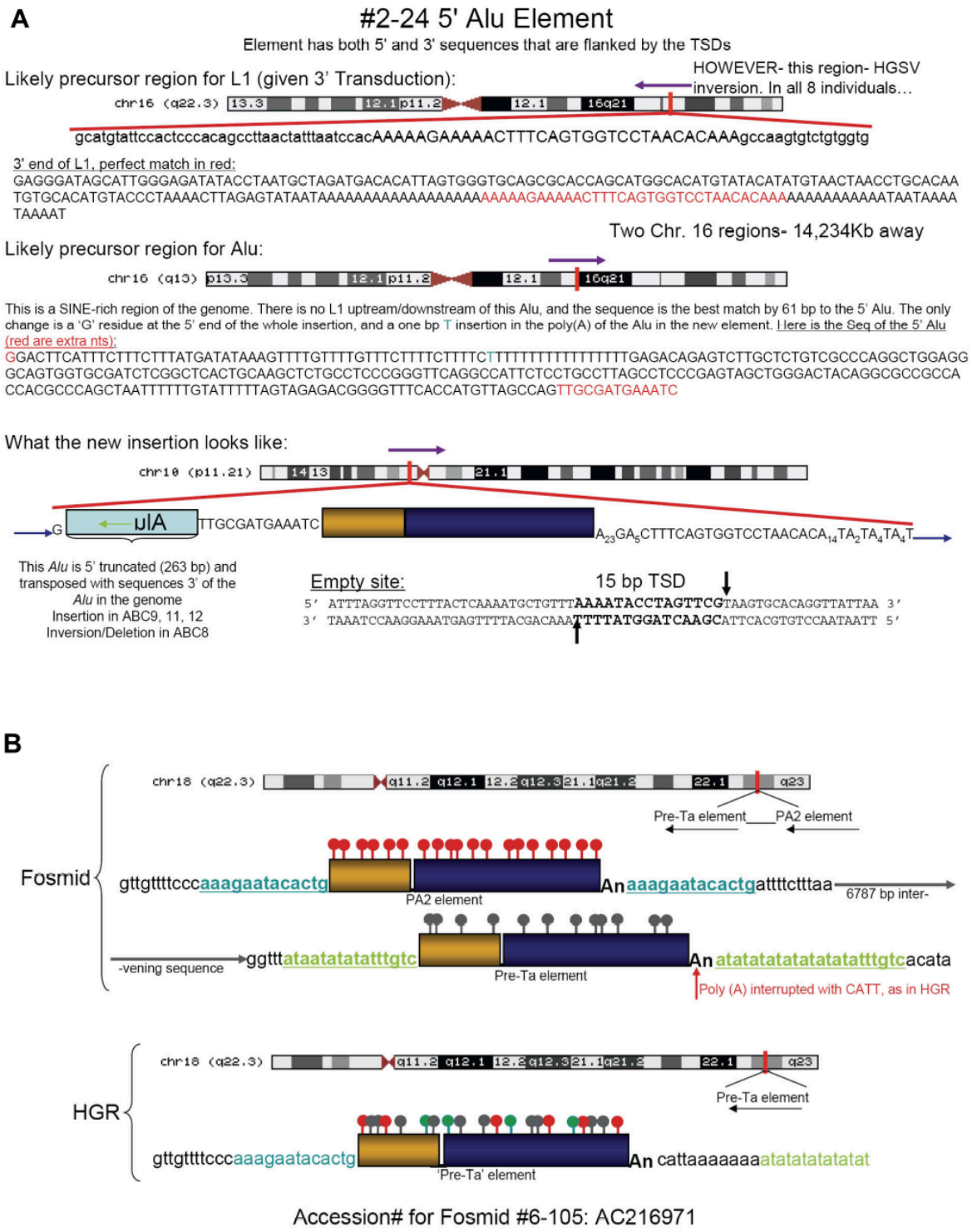
The endonuclease-dependent and endonuclease-independent retrotransposition assays are shown for element #3-24, which contains an S228P mutation in the EN domain of ORF2p. 4364a cells are NHEJ proficient, and are parental to the *XRCC4*-deficient XR-1 cell line. Retrotransposition events are detected as blasticidin resistant colonies in an assay analogous to that shown in Figure 2.2A. L1.3 is able to retrotranspose in both parental and XR-1 cells, whereas a known EN mutant (D205A), #3-24, and the S228P L1.3 mutant retrotranspose in XR-1 cells. An RT mutant (JJ105, D702A) cannot retrotranspose in either cell line (Morrish et al., 2002).

Figure 2.8: A Noncanonical L1 Retrotransposition Event and a Possible Sequence Anomaly in the HGR, Related to Figure 2.3

(A) *A Noncanonical Retrotransposition Event*: the retrotransposition of a L1Hs element (#2-24) was accompanied by the insertion of a segment of an Alu element at its 5′ end. Possible precursors of both the L1 and Alu are shown (14,234kb apart on chromosome 16), as is a diagram of the L1Hs insertion (*i.e.*, empty) site on chromosome 10. The purple arrows above the ideograms indicate the orientation of each element. The empty site on chromosome 10 was PCR amplified from the genomic DNA of ABC9, and sequencing showed that it lacks the Alu/L1Hs insertion.

(B) *A Possible Sequence Anomaly in the HGR:* The fosmid sequence for the region of chromosome 18 is shown at the top of the Figure. The L1PA2 and L1Hs pre-Ta element are in the same orientation, separated by ~6.8kb of sequence that is absent from the HGR, yet is present in a recently completed human genome diploid sequence (Venter et al., 2001). Red lollipops signify sequences specific for the L1PA2, while grey lollipops indicate sequences peculiar to the pre-Ta. The corresponding sequence in the HGR is indicated at the bottom of the Figure. Aqua lollipops indicate sequence changes specific to the harlequin element. Portions of the TSDs for the two elements present in the fosmid flank the harlequin sequence (aqua and light green lettering, respectively). The interspersion of red and grey lollipops makes it unlikely that the L1 was formed by non-allelic homologous recombination. Instead, it is most likely that a sequencing assembly error is responsible for misrepresentation of this region in the HGR. However, we cannot formally rule out that the sequence collapse is due to an unconventional inter-L1 recombination event.

Figure 2.8: A Noncanonical L1 Retrotransposition Event and a Possible Sequence Anomaly in the HGR, Related to Figure 2.3



Accession# for Fosmid #6-105: AC216971

Table 2.2: Activity of the L1 Elements, Related to Figure 2.2

L1 ID	Chromosome	Activity	L1 ID	Chromosome	Activity
1-2	2	0	3-14	10	<1
1-3	6	69	3-15	15	0
1-4	12	107	3-36	20	0
1-5	9	119	3-6	15	0
1-(2-1)	14	118	3-5	8	74
2-1	6	141	3-(2-1)	14	3
2-6	11	0	4-1	6	0
2-7	9	0	4-5	11	1
2-12	17	130	4-7	3	69
2-14	3	101	4-8	5	161
2-21	X	133	4-15	4	54
2-24	10	0	4-17	3	0
2-30	22	46	4-19	22	0
2-32	7	118	4-20	2	0
2-39	20	0	4-27	6	79
2-42	X	n/a	4-29	17	38
2-53	7	86	4-32	1	11
2-59	22	3	4-34	2	57
2-25	11	0	4-35	13	191
2-38	17	63	5-36	10	1
2-61	3	0	5-54	18	0
3-1	3	0	5-55	9	21
3-3	21	15	5-58	2	13
3-4	4	98	5-66	2	23
3-7	12	<1	5-77	6	0
3-10	7	126	5-82	15	<1
3-17	4	0	5-86	5	0
3-18	7	62	6-91	3	157
3-24	2	<1	6-97	4	36
3-25	2	170	6-100	2	<1
3-31	15	120	6-102	6	129
3-34	14	68	6-107	17	139
3-38	5	6	6-109	X	73
3-39	18	41	6-113	4	4
3-40	2	0			

Table 2.2: Activity of the L1 Elements, Related to Figure 2.2

Chart showing the activity of the elements investigated in this study relative to L1.3 (Sassaman et al., 1997). First column: the identifier of each element, where the number preceding the hyphen indicates the individual library (1=G248, 2=ABC9, 3=ABC10, 4=ABC11, 5=ABC12, 6=ABC13). Second column: chromosomal location of the insertion. Third column: the activity of the element in the retrotransposition assay (relative to L1.3 (L19088) (Sassaman et al., 1997)).

Table 2.3: Datasheets for the Elements in This Study, Related to Figure 2.3

Each L1 identified in this study is illustrated in a separate datasheet. Included for each element is an ideogram with the location of the L1 insertion, the empty site sequence from the HGR, and characteristics of the insertion site (including genes in the area). Hallmarks of the L1 retrotransposition event are annotated for the following: the endonuclease (EN) cleavage site, the target site duplication (TSD) size, the approximate size of the poly(A) tail and 3′ transduction/interrupted poly(A) tail details, the activity of the element in the cell culture retrotransposition assay, and the allele frequency of the element (if genotyped). Graphs of representative retrotransposition assays also are shown (related to Table 2.2). Error bars represent the standard deviation of the % retrotransposition, when normalized to the activity of L1.3. In some graphs, we report the retrotransposition efficiencies of additional positive controls (L1.2A, L1RP, LRE3, and #2-12). A L1.3 RT mutant (JM105 or J9105) served as a negative control. L1 elements from individuals ABC11-13 (J9#1-113) are numbered sequentially. Some L1 elements were cloned and assayed for retrotransposition before we determined their presence in the HGR. These elements (*e.g.,* #3-8) are present in the graph, but not the datasheets. The datasheet for element #1-5 was also shown in Figure 2.3.

Table 2.3: Datasheets for the Elements in This Study, Related to Figure 2.3

G248

Characterization of G248 Fosmid #1-2

Clones #2859 B17(G248P88405A9) and #1171 M13(G248P84644G7)

No genes within >50 Kb of insertion site, occurred in a spliced EST/mRNA. Inserted in an L1 MD. Not in dbRIP.

Empty site:

5′ AATGTGACAATT**AAGAAACATCAGA**TAAGAAAATGAG 3′
3′ TTACACTGTTAA**TTCTTTGTAGTCT**ATTCTTTTACTC 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 17 TSD Length: 13

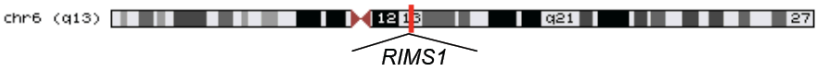
acaatt**AAGAAACATCAGA**GA **AAGAAACATCAGA**taagaa

Non-templated

Allele frequency is at 11% (Badge Lab) Activity- 0% of L1.3

Characterization of Fosmid #1-3

Clones #3369 E07(G248P89400C4) and #3492 J18(G248P801415E9)



Inserted in Intron 2 of *RIMS1* (regulating synaptic membrane exocytosis 1) in the Same Orientation. Not in dbRIP.

Empty site:

5′ CTGGATT**AAGAAAATGTGGCACA**TATACACCATGGAATACTAT 3′
3′ GACCTAA**TTCTTTTACACCGTGT**ATATGTGGTACCTTATGATA 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 30 TSD Length: 16

3′ TD: A_{23} gttttaaattt pA tail (Same as RP, but different polyA lengths)

Non-templated

gattg**AAGAAAATGTGGCACA** **AATTG** **AAGAAAATGTGGCACA**tata

Allele frequency is at 21% (Badge Lab) Activity- 69% of L1.3

Characterization of Fosmids #1-4

Clones #2178 G16(G248P87518D8) and # 2458 K22(G248P80878F11)

Insertion within an LTR retrotransposon about 15Kb upstream of the *NOS* gene (same orientation). Not in dbRIP.

Empty site:

5′ CATGGGGCACCTC**AAAGAATGAGAGGG**AGAGAGCAGCTGGC 3′
3′ GTACCCCGTGGAG**TTTCTTACTCTCCC**TCTCTCGTCGACCG 5′

EN Cleavage site: 5′ CTTT/G 3′ pA Length: 17 TSD Length: 14

3′ Transduction (67bp): A_9TAAATAAATA$_{12}$TAAATAAATA$_{10}$TA$_{11}$GA$_4$G pA tail

cctc**AAAGAATGAGAGGG**G — 3′ TD ~66nt — **AAAGAATGAGAGGG**agag

Non-templated

Allele frequency is at 44% (Badge Lab)

Activity- 107% of L1.3

Characterization of Fosmid #1-5

Clone # 2451 H22(G248P83583D11)

Inserted in intron 5 of *ABCA1*, a cholesterol transporter. Low serum HDL levels and Tangier disease. Opposite Orientation. Not in dbRIP

Empty site:

5′ AAATGCAGTGAGCTT**AAGAACAAGCCTGG**TACCTATGCTGGGC 3′
3′ TTTACGTCACTCGAA**TTCTTGTTCGGACC**ATGGATACGACCCG 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 19 TSD Length: 14

3′ TD (1075 bp): A_{16}/ 275bp of Chr. 2/ A_{22}/ 762bp of Chr. 10/ pA tail

Chromosome 2 TD = LRE3 TD, 2q24.1

Small 5′ truncation (21nt)

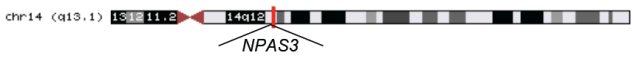
agtgagctt**AAGAACAAGCCTGG** — Large 3′ TD — **AAGAACAAGCCTGG**tacctatg

Allele frequency is at 0%- absent from HGDP (Badge Lab)

Activity- 119% of L1.3

Characterization of Fosmid #1-(2-1)

Clone #G248_P801757_D9



Inserted in Intron 4 of *NPAS3* (Neuronal PAS domain protein 3 isoform 1). Transcription factor- involved in neuronal development. Associated with psychiatric illness- especially schizophrenia

Opposite Orientation. Not in dbRIP

Empty site:

5′ AAAGCTTTTGTAAGGCTGATATT**AAAAAATACTTTG**AGGGCATTGAGGTTACCACAAG 3′
3′ TTTCGAAAACATTCCGACTATAA**TTTTTTATGAAAC**TCCCGTAACTCCAATGGTGTTC 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 102* TSD Length: 13

*Interrupted with another pA site in tail #1- 2nd pA is 30bp

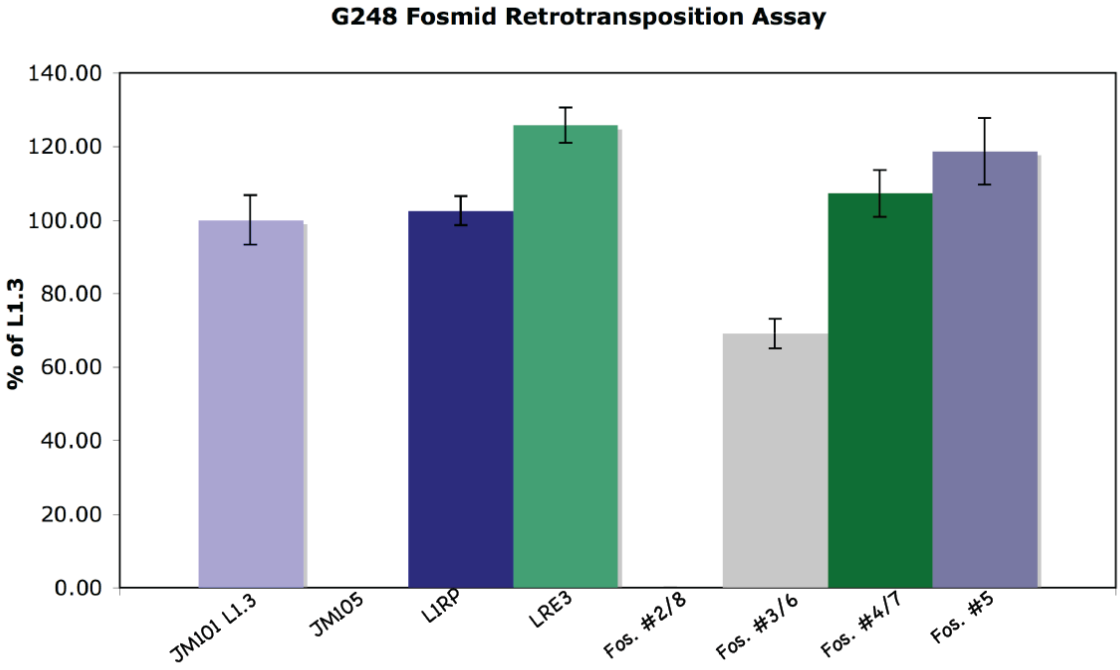
Small 5′ truncation (12nt)

gatatt**AAAAAATACTTTGG** — **AAAAAATACTTTG**agggca

Non-templated

Found from Eichler lab sequencing

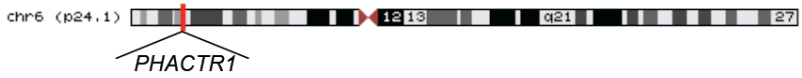
Activity- 118% of L1.3



ABC9

Characterization of Fosmid #2-1

Clone #ABC9_3_2_000043835400_B17



Inserted in Intron 5 of *PHACTR1* (phosphate and actin regulator 1)

Same Orientation. This element is the likely progenitor of two transductions in this study (#2-12 and 1-2-1). Transduction not seen in HGR. One SNP in 3'UTR. Not in dbRIP

Empty site:

5' GACTTATTTGCACAGTGGTT**AGAATTCAAAATACAA**GGAAACTTAATATATTCTCAGA 3'
3' CTGAATAAACGTGTCACCAA**TCTTAAGTTTTATGTT**CCTTTGAATTATATAAGAGTCT 5'

EN Cleavage site: 5' TTCT/A 3' pA Length: 32 TSD Length: 16

tggtt**AGAATTCAAAATACAA** **AGAATTCAAAATACAA**ggaaa

Allele frequency is at **18**% (Badge Lab)

Activity- 141% of L1.3

Characterization of Fosmid #2-6

Clone #ABC9_3_5_000044009800_K17

Inserted in Intron 6 of *SBF2* (SET binding factor 2) aka *MTMR13* (myotubularin -related protein- Pseudophosphatase) CMT type 4B2

Same Orientation. 2 non-syn cSNPs. 6 UTR SNPs. Not in dbRIP, however an unresolved L1 is close.

Empty site:

5′ AAATATAGGGTTTTTTTTTAAAGTTCTT**AAAAAAGTTTTTGTT**CAGTTTGTCTTACAAATATTA 3′
3′ TTTATATCCCAAAAAAAAATTTCAAGAA**TTTTTTCAAAAACAA**GTCAAACAGAATGTTTATAAT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 11 TSD Length: 15

Prior to pure pA tail of 11, A8-T-A4-T-A4-T-A4-T-A4-T

gttctt**AAAAAAGTTTTTGTT** **AAAAAAGTTTTTGTT**cagtttg

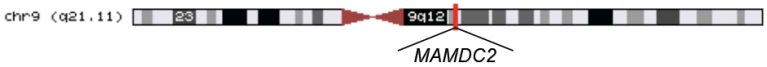
Allele frequency is at **1.2**% (Badge Lab)

7% (ATLAS paper)

Activity- 0% of L1.3

Characterization of Fosmid #2-7

Clone #ABC9_3_5_000045471900_O4



Inserted in Intron 8 of *MAMDC2* (MAM domain containing 2) in the Opposite Orientation.

1 non-syn SNP/5 syn./ 2 UTR SNPs. Not in dbRIP, however an unresolved L1 is close.

Empty site:

5′ CTGGGTGACAGAGTGAGATCCTGTCTT**AAAAAAACAAAACA**AAACAAAAGTACCCAGGAAA 3′
3′ GACCCACTGTCTCACACTAGGACAGAA**TTTTTTTGTTTTGT**TTTGTTTTCATGGGTCCTTT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 26 TSD Length: 14

A4-T-A21-GAGCCATGAC-A3

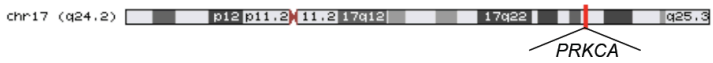
gtctt**AAAAAAACAAAACA**G **AAAAAAACAAAACA**aaacaa

Non-templated

Activity- 0% of L1.3

Characterization of Fosmid #2-12

Clone #ABC9_3_2_000046009500_E6



Inserted in Intron 3 of *PRKCA* (Protein kinase C, alpha)

Deletions and mutations- predisposition to cancer, modulation of contractile heart motion, prefrontal cortical regulation of working memory

Same Orientation close to exon. 1 non-syn SNP/1 syn. SNP/ 8 UTR SNPs. Not in dbRIP

Empty site:

5′ AGCAAGACTCTGTCTC**AAAAAACAACAGTAAATATC**TAAAATATTTAGTGGCGAAGTCACCACC 3′
3′ TCGTTCTGAGACAGAG**TTTTTTGTTGTCATTTATAG**ATTTTATAAATCACCGCTTCAGTGGTGG 5′

EN Cleavage site: 5′ TTTT/G 3′ pA Length: 52 TSD Length: 20

ORF2L out is the same as 2-1 (2bp pA change) Re-sequenced; Southerns different, too

ctc**AAAAAACAACAGTAAATATC**GAA **TD** **AAAAAACAACAGTAAATATC**taa

Non-templated

Allele frequency is at **22**% (Badge Lab)

Activity- 130% of L1.3

Characterization of Fosmid #2-14

Clone #ABC9_3_5_000043957900_J21

Inserted in Intron 13 of *SCN5A* (Voltage gated sodium channel type V alpha)

Opposite Orientation. 8 non-syn SNPs/ 2 synonymous. Not in dbRIP

Missense mutations- myotonia, Long QT or Brugada syndrome, cardiomyopathy, variant (S1103Y)-present in 13.2% of A. Americans. Arrhythmia…

Empty site:

5′ GAACAGATAAGTAAATGAATGAAAAC**AGAATGAGTAAATAATG**AATAAATAGGTAAGTAAGTGGG 3′
3′ CTTGTCTATTCATTTACTTACTTTTG**TCTTACTCATTTATTAC**TTATTTATCCATTCATTCACCC 5′

EN Cleavage site: 5′ TTCT/G 3′ pA Length: 22 TSD Length: 17

3′ Transduction: 17 pA tail-GAATTGTAAAAAAAAAAATTATAAATAAAACAAAGAAGAATATG

This Transduction is 100% match to 5 other L1HS 3′ Transductions from BLAT

aac**AGAATGAGTAAATAATG**GA **AGAATGAGTAAATAATG**aataa

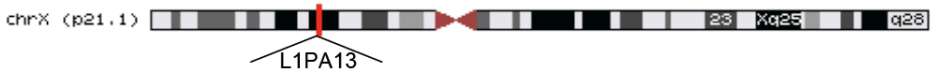
Non-templated

Allele frequency is at **27**% (Badge Lab)

Activity- 101% of L1.3

Characterization of Fosmid #2-21

Clone #ABC9_3_2_000043852700_D21



Inserted within an old LINE-1 element.

~ 55000 bp upstream of DMD in opposite orientation to the gene. Not in dbRIP

Empty site:

5′ TATCTTTGTTCACATTAGTTTC**AAAGAATTTCTTGA**TTTCTGCCTTAATTTCACTATTT 3′
3′ ATAGAAACAAGTGTAATCAAAG**TTTCTTAAAGAACT**AAAGACGGAATTAAAGTGATAAA 5′

EN Cleavage site: 5′ CTTT/G 3′ pA Length: 27-cacat-15-caacaaaacaaagc-20As TSD Length: 14

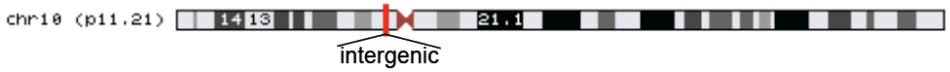
agtttc**AAAGAATTTCTTGA** **AAAGAATTTCTTGA**tttctgccttaa

Allele frequency is at **21**% (Badge Lab)

Activity- 133% of L1.3

Characterization of Fosmid #2-24

Clone #ABC9_3_2_000043873800_F17



Inserted in an intergenic region, with a couple spliced ESTs >25,000 bp away. Not in dbRIP

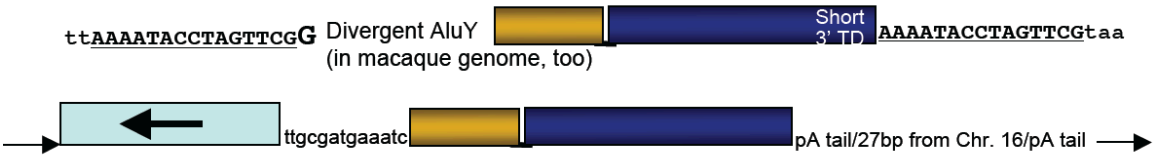
Insertion appears to coordinate with the mobilization of an older AluY. Recombination mediated?

Empty site:

5′ ATTTAGGTTCCTTTACTCAAAATGCTGTTT**AAAATACCTAGTTCG**TAAGTGCACAGGTTATTAA 3′
3′ TAAATCCAAGGAAATGAGTTTTACGACAAA**TTTTATGGATCAAGC**ATTCACGTGTCCAATAATT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 76 bp (interrupted) TSD Length: 15

3'Transduction/interrupted pA tail matches chromosome 16

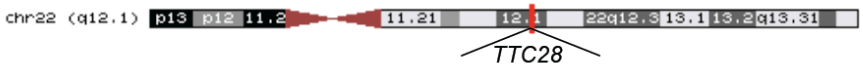


Emplty and filled sites confirmed by genomic DNA PCRs- similar to ATLAS paper (empty=empty, no Alu)

Activity- 0% of L1.3

Characterization of Fosmid #2-30

Clone #ABC9_3_2_000041232900_J18



Inserted within an L1ME2 in the Same Orientation as *TTC28* (tetracopeptide repeat protein 28) in intron 11.

Not in dbRIP

Empty site:

5′ ATCTCTGATATTTACCTATAGATT**GAAAATAAATGGGTA**GAAAACGATATACCATGTAAACGGT 3′
3′ TAGAGACTATAAATGGATATCTAA**CTTTTATTTACCCAT**CTTTTGCTATATGGTACATTTGCCA 5′

EN Cleavage site: 5′ TTTC/A 3′ pA Length: 52 TSD Length: 15

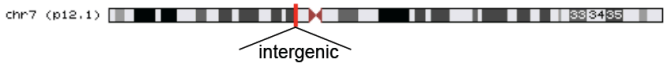
Ts throughout, many pA signals

Allele frequency is at **9**% (Badge Lab)

Activity- 46% of L1.3

Characterization of Fosmid #2-32

Clone #ABC9_3_2_000043849300_F23



Inserted within an intergenic region (spliced EST upstream), not within an element.

Not in dbRIP

Empty site:

5′ AGAAAAGGGCAGGTAACAAATATCATTTTTT**AAAAAAACAAATAA**CACAGAAGGAAAACTCAGGA 3′
3′ TCTTTTCCCGTCCATTGTTTATAGTAAAAAA**TTTTTTTGTTTATT**GTGTCTTCCTTTTGAGTCCT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 31 TSD Length: 14

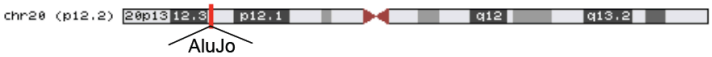
TTTTT**AAAAAAACAAATAA** **AAAAAAACAAATAA**CACAG

Allele frequency is at **21**% (Badge Lab)

Activity- 118% of L1.3

Characterization of Fosmid #2-39

Clone #ABC9_3_5_000043966000_G16



Inserted in an intergenic region (spliced EST upstream), within an AluJo.

Not in dbRIP

Empty site:

5′ CACAACCAGTTAAATGCAACCTT**AAAAATAATTTATA**GACAGGGTCTCACTGTGTTACCCAGACT 3′
3′ GTGTTGGTCAATTTACGTTGGAA**TTTTTATTAAATAT**CTGTCCCAGAGTGACACAATGGGTCTGA 5′

EN Cleavage site: 5′ AATT/A 3′ pA Length: 20 or 14 TSD Length: 8-14

Or, if pA tail changed, TTTT/A

SNP (in HGR is a C)

Was an A in pA tail (by sequencing)

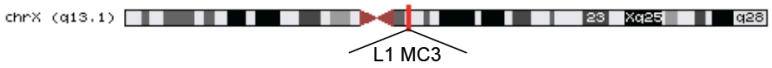
CAACCTT**AAAAATAATTTATA** **AAAAA**T**AATTTA**T**A**GACAGGG

Allele frequency is at **5**% (Badge Lab)

Activity- 0% of L1.3

Characterization of Fosmid #2-42

Clone #ABC9_3_2_000043888700_B3



Inserted within an older LINE-1

Not in dbRIP

Empty site:

5′ AACTTATATGAAATGGAACAATTTCTT**GAAAGACACAAACTA**TTAAACTGACACAAGGAGAAA 3′
3′ TTGAATATACTTTACCTTGTTAAAGAA**CTTTCTGTGTTTGAT**AATTTGACTGTGTTCCTCTTT 5′

EN Cleavage site: 5′ TTTC/A 3′ pA Length: 22 TSD Length: 15

TTTCTT**GAAAGACACAAACTA** **GAAAGACACAAACTA**TTAA

Activity- ND- *Acc*I site in ORF2, therefore refractory to cloning

Characterization of Fosmid #2-53

Clone #ABC9_3_5_000045468900_I15

Inserted within an older LINE-1 (L1 M4) in the opposite orientation.

Downstream of *NXPH1* in the same orientation (neurexophilin-1). Not in dbRIP

Empty site:

5′ ATGTTTATTCTCTAGATGATTTGTTCCTTTT**AACAATATGTAA**ATGTTTCACTTAATATTTTG 3′
3′ TACAAATAAGAGATCTACTAAACAAGGAAAA**TTGTTATACAT**TACAAAGTGAATTATAAAAC 5′

EN Cleavage site: 5′ TGTT/A 3′ pA Length: 41 TSD Length: 11

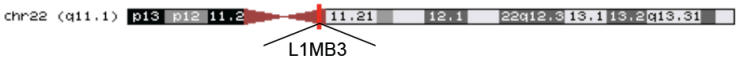
ORF2L out matches perfectly with 3 other L1HS elements in the genome (and #3-39), and contains a common transduction followed by a pA tail

Allele frequency is at **6**% (Badge Lab)

Activity- 86% of L1.3

Characterization of Fosmid #2-59

Clone #ABC9_3_5_000046213200_J3



Inserted within an intergenic region (spliced EST upstream and *XKR3* downstream), within an older L1.

Not in dbRIP

Empty site:

5′ TTTGATAAGGGATCAATATCCAGAATATAT**AAAGAACTCCT**ACAACTCAGCAATAAAGAA 3′
3′ AAACTATTCCCTAGTTATAGGTCTTATATA**TTTCTTGAGGA**TGTTGAGTCGTTATTTCTT 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 33 TSD Length: 11

2 elements are in this fosmid, one old, one new. (L1PA4 1300-6 kb here)

Within Segmental Duplication

Activity- 3% of L1.3

Characterization of Fosmid #2-25

Clone #ABC9_3_2_000043872200_O3

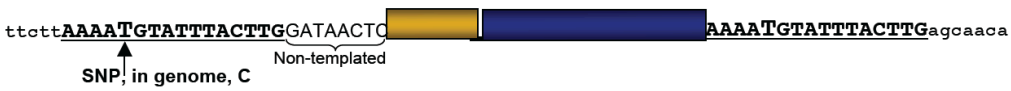
Inserted in an intergenic region.

Not in dbRIP

Empty site:

5′ GAAGAGGAGAGCGAGGCTGAGGGTTCTT**AAAATGTATTTACTTG**AGCAACACTGAAT 3′
3′ CTTCTCCTCTCGCTCCGACTCCCAAGAA**TTTTACATAAATGAAC**TCGTTGTGACTTA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 22 TSD Length: 16

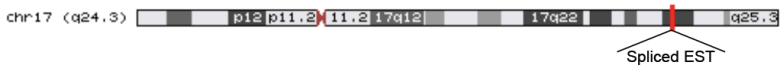


3' ATLAS only

Activity- 0% of L1.3

Characterization of Fosmid #2-38

Clone #ABC9_3_5_000043972200_E24



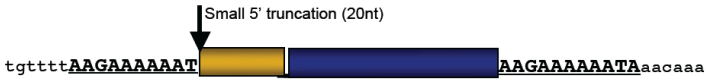
Not within a gene, occurred within a human spliced EST (opposite orientation).

Not in dbRIP, however an unresolved L1 is close

Empty site:

5′ CAGTTGCTGAATGATTTTGTTTTAAGAAAAAATAAACAAAAAGTACTTAGCCTCAAGGA 3′
3′ GTCAACGACTTACTAAAACAAAATTCTTTTTTTATTTGTTTTTCATGAATCGGAGTTCCT 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 41 TSD Length: 11



Might be due to sequencing problems, but ~500bp after the pA tail of the element, there is a large amount of sequence that doesn't BLAT to either this region of the genome, or anywhere else in the HGR

3′ ATLAS only

Activity- 63% of L1.3

Characterization of Fosmid #2-61

Clone #ABC9_3_5_000043931800_M16

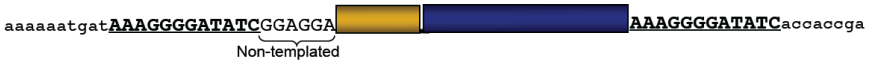
Inserted in an intergenic region. Occurred within a nearly full length L1PA5, towards the 3′end.

Not in dbRIP

Empty site:

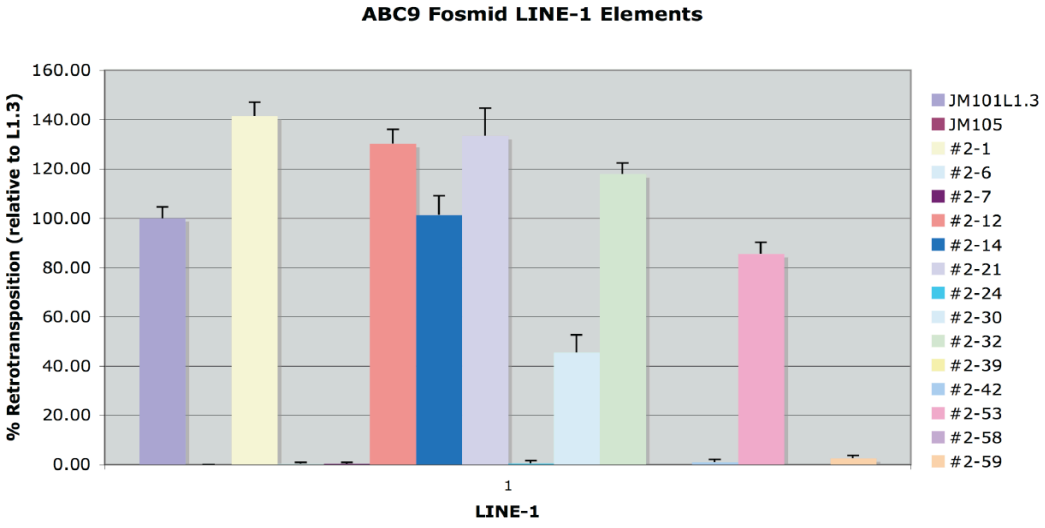
5′ GAAAAGAGAGAATAATCAAATAGACACAATAAAAAATGATAAAGGGGATATCACCACCGA 3′
3′ CTTTTCTCTCTTATTAGTTTATCTGTGTTATTTTTTACTATTTCCCCTATAGTGGTGGCT 5′

EN Cleavage site: 5′CTTT/A 3′ pA Length: 8 TSD Length: 12



3′ ATLAS only

Activity- 0% of L1.3



ABC10

Characterization of Fosmid #3-1

Clone #ABC10_2_1_000043667000_H21

Inserted within an older LINE-1 element in an intergenic region

Not in dbRIP

Empty site:

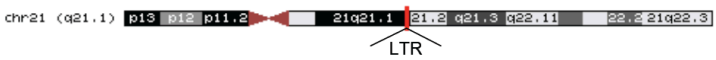
5′ TTAGGAAACCAACAATTTCTTTTAAAGTT**AAAGACAC**ATTTACCTTATTACCCTATTACCCA 3′
3′ AATCCTTTGGTTGTTAAAGAAAATTTCAA**TTTCTGTG**TAAATGGAATAATGGGATAATGGGT 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 75 TSD Length: 8

Activity- 0% of L1.3

Characterization of Fosmid #3-3

Clone #ABC10_2_1_000043666900_A3



Inserted within an LTR retrotransposon- MLT1H-int (MaLR)

In an intergenic region and within two spliced ESTs

Not in dbRIP

Empty site:

5′ TTCTCTGATTCCACATCTTTAT**GAAAAGAAAATAGGT**TAAGAAAATTTCTGTAAGAAAAT 3′
3′ AAGAGACTAAGGTGTAGAAATA**CTTTTCTTTTATCCA**ATTCTTTTAAAGACATTCTTTTA 5′

EN Cleavage site: 5′ TTTC/A 3′ pA Length: 28 TSD Length: 15

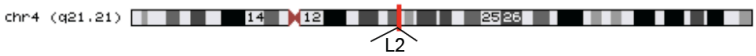
3′ Transduction: Chromosome 21:28214836-28215426 (no element here in draft)
Includes the middle to end of an Alu Ya5 and part of an LTR retrotransposon THE1D (family MaLR)

Allele frequency is at **26**% (Badge Lab)

Activity- 15% of L1.3

Characterization of Fosmid #3-4

Clone #ABC10_2_1_000044492700_P8



Inserted within an L2 distal to any known genes. There is either a sequence change in the 5′ poly G track or untemplated nucleotides for this element. Not in dbRIP

Empty site:

5′ TATGAACAATACCATT**AACAAAACAG**ACACAAATCTTTGTGATTGTAGAA 3′
3′ ATACTTGTTATGGTAA**TTGTTTTGTC**TGTGTTTAGAAACACTAACATCTT 5′

EN Cleavage site: 5′TGTT/A 3′ pA Length: 51 (one a to g change) TSD Length: 10

Allele frequency is at **0.25**% (Badge Lab)

Typed on Zimbabwean panel

Activity- 98% of L1.3

Characterization of Fosmid #3-7

Clone #ABC10_2_1_000045521700_B16

This location already contains a full-length LINE-1 HS.

However, location contains 2 full-length elements (12-13 kb apart). Not in dbRIP (non-HGR element)

HGR element was 0.3% diverged and was in the Brouha et al paper- AC005885- Ta1d with an allele freq. of 0.67 and an activity of 2.3% of RP.

Empty site:

5′ TTGTTTACTCTGGGAAATTC**AAAAGAGTACAAGTC**AATAATTTTGTTAATTG 3′
3′ AACAAATGAGACCCTTTAAG**TTTTCTCATGTTCAG**TTATTAAAACAATTAAC 5′

EN Cleavage site: 5′ TTTT/G 3′ pA Length: 15 TSD Length: 15

Ta-1nd

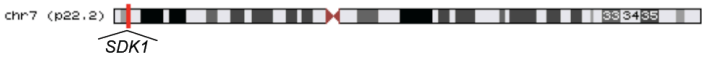
AAATTC**AAAAGAGTACAAGTC** **AAAAGAGTACAAGTC**AATAATT

Activity- <1% of L1.3

AC005885 (HGR element) also in ATLAS paper
Present in HGR/ 99.7% identity to L1.3
Insertion at 63%- Ubiquitous pop. distribution

Characterization of Fosmid #3-10

Clone #ABC10_2_1_000043644900_H4



Within intron **5** of the gene *SDK1* (sidekick 1)

Synaptic protein that directs homophilic adhesion and laminar targeting of neurites in vivo. Opp Orientation

4 non-syn/6 synonymous SNPs. Not in dbRIP

Empty site:

5′ CTTACTTAGTAAGTCTCCTCTT**AAAAATACACATCCTC**ATAACACTGGCATGATTGGCAAGC 3′
3′ GAATGAATCATTCAGAGGAGAA**TTTTTATGTGTAGGAG**TATTGTGACCGTACTAACCGTTCG 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 19 TSD Length: 16

CTCCTCTT**AAAAATACACATCCTC**GGA **AAAAATACACATCCTC**ATAA

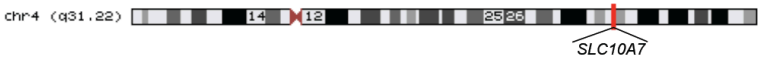
Non-templated

Allele frequency is at **0.25**% (Badge Lab)
Typed on Zimbabwean panel

Activity- 126% of L1.3

Characterization of Fosmid #3-17

Clone #ABC10_2_1_000044457200_I12



Within intron **7** of the gene *SLC10A7*

Solute carrier family 10 (sodium/bile acid co transporter member 7) is a membrane protein involved in ion transport. Opp. Orientation

No SNPs listed within transcript. Not in dbRIP, however an unresolved L1 is close

Empty site:

5′ TGCTTACTACTAACTTGTTAT**AAGAAGTTAAAATAG**AATTCCTATCAGTCCAATT 3′
3′ ACGAATGATGATTGAACAATA**TTCTTCAATTTTATC**TTAAGGATAGTCAGGTTAA 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 13 TSD Length: 15

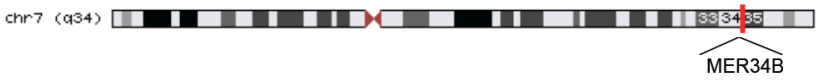
TAT**AAGAAGTTAAAATAG** **AAGAAGTTAAAATAG**AATT

Allele frequency is high (Xing *et al.*, 2009)

Activity- 0% of L1.3

Characterization of Fosmid #3-18

Clone #ABC10_2_1_000044548600_H9



Within a spliced EST in an ERV1 family LTR retrotransposon

Not in dbRIP

Empty site:

5′ ACTTGCCTAACATATAGTAATAT**AAAAATAGTTAATA**GAAAATTTAACTTGAGA 3′
3′ TGAACGGATTGTATATCATTATA**TTTTTATCAATTAT**CTTTTAAATTGAACTCT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 24 TSD Length: 14

A-G changes

ATAT**AAAAATAGTTAATA**T **AAAAATAGTTAATA**GAAAA

Untemplated T

Activity- 62% of L1.3

Characterization of Fosmid #3-24

Clone #ABC10_2_1_000044510700_F10

Intergenic region- in an older LINE-1

Not in dbRIP

Empty site:

5′ AGGCAAATCATGAAGTCTTCTTAATCTGAT**AAAAATGATACTTAG**AAAATCAACTTTGAGTAAGCATTAT 3′
3′ TCCGTTTAGTACTTCAGAAGAATTAGACTA**TTTTTTACTATGAATC**TTTTAGTTGAAACTCATTCGTAATA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 49 TSD Length: 15

TTAATCTGAT**AAAAATGATACTTAG** **AAAAATGATACTTAG**AAAATCAACTTT

ACG

Allele frequency is at **0**% (Badge Lab)
Absent from the HGDP

Activity- <1% of L1.3

Characterization of Fosmid #3-25

Clone #ABC10_2_1_000048936800_O12

Inserted in the *KIF5C* gene, in opposite orientation to the gene. Within intron 9. *KIF5C* is the kinesin heavy chain isoform 5C. May be a neuronal specific Kinesin heavy chain isoform.

Not in dbRIP

Empty site:

5′ TGAAGTTCACAGATAAAACGCTGTATG**AAAGAAGCCAGATAC**AAAACAGTACTGGGTGATTTCATTTTCAAA 3′
3′ ACTTCAAGTGTCTATTTTGCGACATAC**TTTCTTCGGTCTATG**TTTTGTCATGACCCACTAAAGTAAAAGTTT 5′

EN Cleavage site: 5′ CTTT/C 3′ pA Length: 55bp TSD Length: 15bp

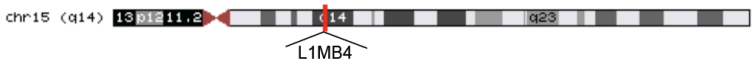
ACGCTGTATG**AAAGAAGCCAGATAC** **AAAGAAGCCAGATAC**AAAACAGTAC

Allele frequency is at **0.25**% (Badge Lab)
Typed on Zimbabwean panel

Activity- 170% of L1.3

Characterization of Fosmid #3-31

Clone #ABC10_2_1_000000840721_H19



Occurred within an old L1 (L1MB4) in a spliced EST/mRNA (same orientation).
Not in dbRIP

Empty site:

5′ CAAAATAGATCTTCAGAAACTGGCTCTG**AAAAAATGAAAATC**TATTAATATGAATTACTTGACAAAGAATTCA 3′
3′ GTTTTATCTAGAAGTCTTTGACCGAGAC**TTTTTTACTTTTAG**ATAATTATACTTAATGAACTGTTTCTTAAGT 5′

EN Cleavage site: 5′ TTTT/C 3′ pA Length: 66 bp TSD Length: 14bp

A_{33}GTTTTAAATTTA$_{22}$ (RP)

CTGGCTCTG**AAAAAATGAAAATC** **AAAAAATGAAAATC**TATTAATA

Allele frequency is at **0.17**% (Badge Lab)
Typed on HGDP

Activity- 120% of L1.3

Characterization of Fosmid #3-34

Clone #ABC10_2_1_000044573100_J8

Within intron **5** of the gene *RAD51L1* (*RAD51* Like 1) in the same orientation

RAD51L1 is a tumor-suppressing gene implicated in many tumors (e.g., uterine leiomyomas and pulmonary chondroid hamartomas) and involved in recombinatorial repair of DNA double-strand breaks. 6 non-syn/ 2 syn/ 1 in a UTR

Not in dbRIP

Empty site:

5′ ACTTTCTAGAGTTGATGTTACT**AAGAAATCAATGACATGG**TTAATAAGCTCACTAGTCCT 3′
3′ TGAAAGATCTCAACTACAATGA**TTCTTTAGTTACTGTACC**AATTATTCGAGTGATCAGGA 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 18 TSD Length: 18

CT**AAGAAATCAATGACATGG**GC **AAGAAATCAATGACATGG**TTA

Non-templated

Allele frequency is at **7.8**% (Badge Lab)
Typed on Zimbabwean panel

Activity- 68% of L1.3

Characterization of Fosmid #3-38

Clone #ABC10_2_1_000044679300_J6

Within a spliced EST, inserted in an intergenic region

Progenitor is likely on chromosome 4 in Anthrax toxin receptor 2.

(Poss. Progenitor-AC093886 4q21.21 Ta-0 0.00 allele freq 0%of RP activity in Brouha). 3-38 is Not in dbRIP

Empty site:

5′ GTCCACCTTTCTGAAGCATAT**AAAAATAACTTAAAATTCC**TTCAATATTAAATGATGTAG 3′
3′ CAGGTGGAAAGACTTCGTATA**TTTTTATTGAATTTTAAGG**AAGTTATAATTTACTACATC 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 18 TSD Length: 19

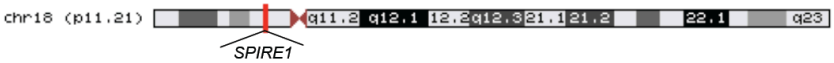
ATAT**AAAAATAACTTAAAATTCC** TD **AAAAATAACTTAAAATTCC**TTCA

Transduction: pA tail, 380 bp of chromosome 4

Activity- 6% of L1.3

Characterization of Fosmid #3-39

Clone #ABC10_2_1_000044679100_A17



Within intron **8** of the gene *SPIRE1* Opp. Orientation

SPIRE proteins are involved in actin nucleation and axis formation in oocytes and embryos.

6 SNPs in UTRs. This Element is in dbRIP, as it was found in the 2004 Boissinot paper

Empty site:

5′ CCAAAGCTATATACAGATCCAATAGAATCCCC**AAGAAAATCCT**AATGGTATTTTTTACAGAA 3′
3′ GGTTTCGATATATGTCTAGGTTATCTTAGGGG**TTCTTTTAGGA**TTACCATAAAAAATGTCTT 5′

EN Cleavage site: 5′ TCTT/G 3′ pA Length: 27 TSD Length: 11

AATCCCC**AAGAAAATCCT** TD **AAGAAAATCCT**AATGGT

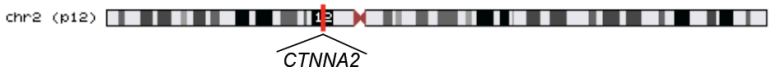
Common transduction to L1HS elements on chromosomes X, 5, and 6, as well as #2-53, but this insertion did not occur on the same chromosome.

Allele frequency is at **2**% (Badge Lab)

Activity- 41% of L1.3

Characterization of Fosmid #3-40

Clone #ABC10_2_1_000044091600_J2



Within intron **2** of the gene *CTNNA2* (*Catenin alpha-2*) Same orientation

Catenin Alpha-2 is a gene that has been found to play a role in the motility of dendritic spine heads of the hippocampus. 2 syn SNPs/1 in a UTR. Not in dbRIP

Empty site:

5′ TTACATTTTTACATGATTTTTTAAAAAT**AAGAAATTGA**ACTGTATACCATGATGTGTGAA 3′
3′ AATGTAAAAATGTACTAAAAAATTTTTA**TTCTTTAACT**TGACATATGGTACTACACACTT 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 18 TSD Length: 10

AAAAAT**AAGAAATTGA** **AAGAAATTGA**ACTGTATA

Allele frequency is at **9**% (Badge Lab)
Typed on Zimbabwean panel

Activity- 0% of L1.3

Characterization of Fosmid #3-14

Clone #ABC10_2_1_000044088700_G10

In an intergenic region. Upstream (1kb away) and in same orientation as predicted gene *AK093219*

Not in dbRIP

Empty site:

5′ GAAAATTATCCCACCT**AAAGAAAACAGAG**ACCTCAAAAAGACTAATCGCAGAGACTAA 3′
3′ CTTTTAATAGGGTGGA**TTTCTTTTGTCTC**TGGAGTTTTTCTGATTAGCGTCTCTGATT 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 20 TSD Length: 13

CACCT**AAAGAAAACAGAG**GAATTATAGGA **AAAGAAAACAGAG**ACCTCAAA

Non-templated

3′ ATLAS only

Allele frequency is at **9**% (Badge Lab)

Activity- <1% of L1.3

Characterization of Fosmid #3-15

Clone #ABC10_2_1_000045501700_B20

Inserted in Intron 2 of *HOMER2*
Same Orientation
Knock out mutations in mice cause cocaine sensitization. Neuronal gene. IN dbRIP- found by Boissinot

Empty site:

```
5' GTGGATGTGTTTCAGTAAGAAGTGGCTGGTGTAGAATGGACTCAGCTCATTTTGAGC  3'
3' CACCTACACAAAGTCATTCTTCACCGACCACATCTTACCTGAGTCGAGTAAAACTCG  5'
```

EN Cleavage site: 5' TCTT/A 3' pA Length: 24 TSD Length: 15

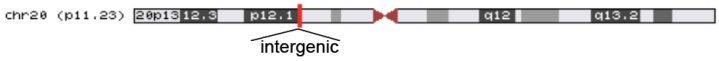
5' ATLAS only- 3' UTR is ACA- site downstream too close?

Allele frequency is at **43**% (Badge *et al.*, 2003)

Activity- 0% of L1.3

Characterization of Fosmid #3-36

Clone #ABC10_2_1_000044678300_B16



Inserted in an intergenic region. Not in dbRIP

Empty site:

```
5' ATTCAACACGTGCCTTGGAGAATGAAAATACAGTCAGGTTTGGGAACAACCACTT  3'
3' TAAGTTGTGCACGGAACCTCTTACTTTTATGTCAGTCCAAACCCTTGTTGGTGAA  5'
```

EN Cleavage site: 5' TTTC/A 3' pA Length: 27 TSD Length: 12

3' ATLAS only

Activity- 0% of L1.3

Characterization of Fosmid #3-6

Clone #ABC10_2_1_000045545200_J11

Inserted in an intergenic region.
Not in dbRIP

Empty site:

```
5' AGGTATTAATAGGAATTCTACAGCTTTGAAAATGAGTTGAAATATCCATATTAGTTT  3'
3' TCCATAATTATCCTTAAGATGTCGAAACTTTTACTCAACTTTATAGGTATAATCAAA  5'
```

EN Cleavage site: 5' TTTC/A 3' pA Length: 41 (a couple of gs in tail) TSD Length: 14

3' ATLAS only

Activity- 0% of L1.3

Characterization of Fosmid #3-5

Clone #ABC10_2_1_000045542900_P9

Within an intergenic region. Occurred in an old LTR retrotransposon, in a spliced EST.

Not in dbRIP.

Empty site:

5′ TAAGAATAGTCTAGGTTGTGCTGCAGTAACAAACAATCCCAAAATCTCAGTGGCTTTTAACAAC 3′
3′ ATTCTTATCAGATCCAACACGACGTCATTGTTTGTTAGGGTTTTAGAGTCACCGAAAATTGTTG 5′

EN Cleavage site: 5′ GATT/G 3′ pA Length: 49 TSD Length: 6

5′ ATLAS only (Pre-Ta element)

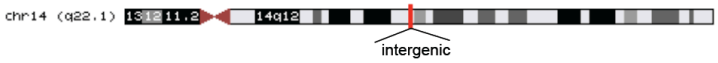
Allele frequency is at **0.35**% (Badge Lab)

Typed on HGDP

Activity- 74% of L1.3

Characterization of Fosmid #3-(2-1)

Clone #ABC10_2_1_000044551300_P12



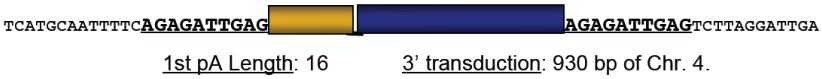
Within an intergenic region. Does not occur within an element. Transduction locale below.

Not in dbRIP

Empty site:

5′ AAATTTTAAAATATCATGCAATTTTCAGAGATTGAGTCTTAGGATTGATGAAAAATATAT 3′
3′ TTTAAAATTTTATAGTACGTTAAAAGTCTCTAACTCAGAATCCTAACTACTTTTTATATA 5′

EN Cleavage site: 5′ CTCT/G 3′ 2nd pA Length: 24 TSD Length: 10



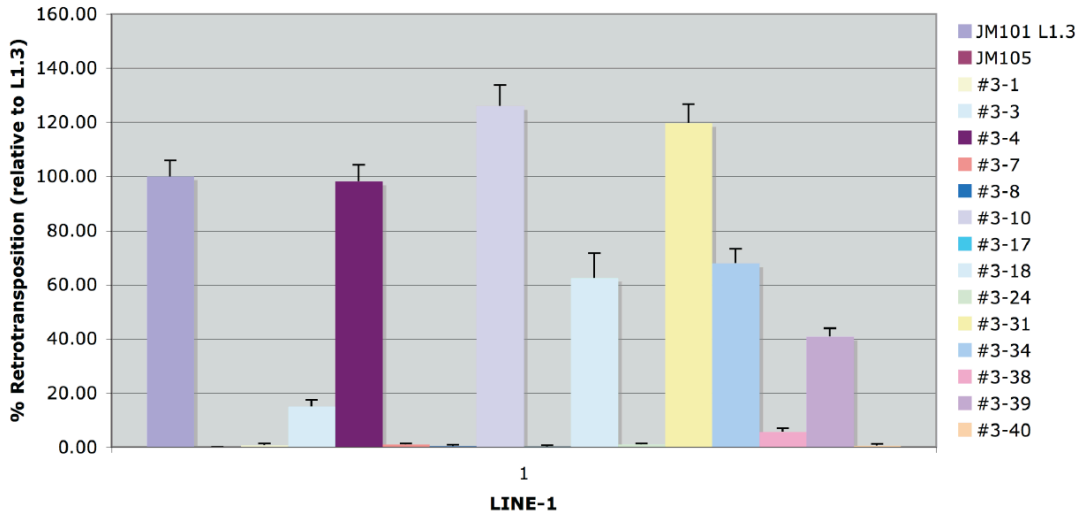
1st pA Length: 16 3′ transduction: 930 bp of Chr. 4.

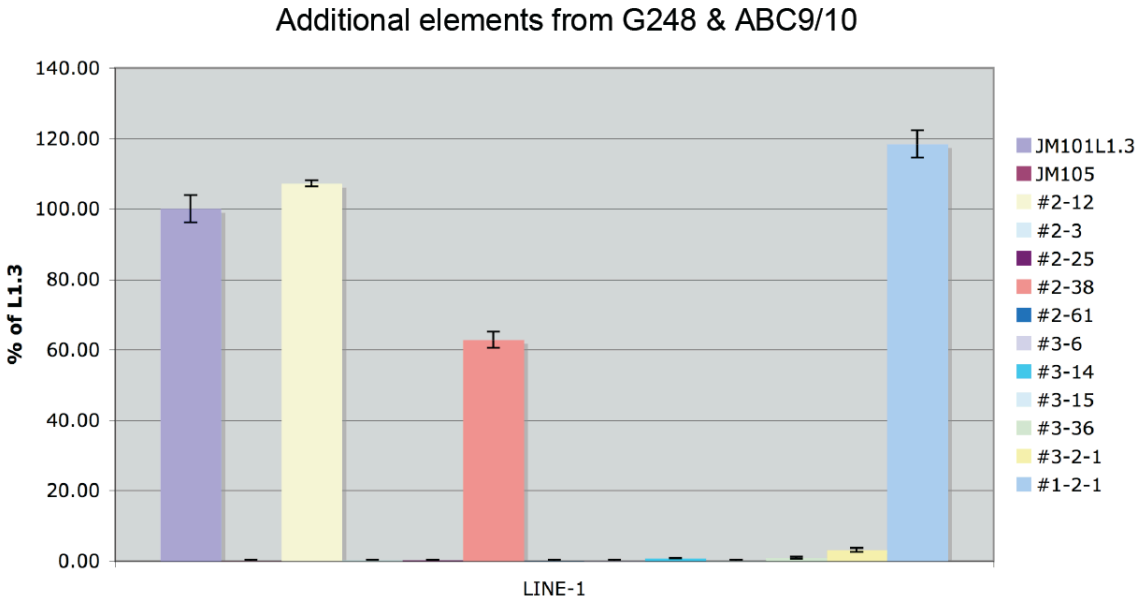
3′transduction: consists of an L1M1 followed by part of an LTR retrotransposon and the end of an old LINE (L1MEc). Location of progenitor also not in dbRIP, and not in cohort of found elements

Found from Eichler lab sequencing

Activity- 3% of L1.3

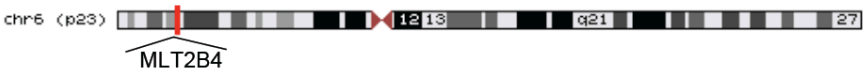
ABC10 Fosmid LINE-1 Elements





ABC11

Characterization of Fosmid #4-1



Inserted into an old ERVL, about 15,000 bp upstream of the *GFOD1* gene, opposite orientation.

Repeat rich region- IN dbRIP

Empty site:

5′ GAAGACAGCAGTGGTTAATTTTACATGTC**AAATGTAAAAATGA**CTGGGCCAGAAAGTGTCCAGATATTT 3′
3′ CTTCTGTCGTCAGGAATTAAAATCTACAG**TTTACATTTTTACT**GACCCGGTCTTTCACAGGTCTATAAA 5′

EN Cleavage site: 5′ ATTT/G 3′ pA Length: 16 bp TSD Length: 14

acatgtc**AAATGTAAAAATGA**GAGTCGAGGA **AAATGTAAAAATGA**ctgggcc

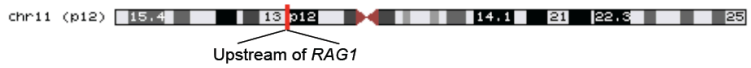
Untemplated

Screened via sequencing

Allele Freq. is at **58**% (Myers *et al.* 2002)

Activity- 0% of L1.3

Characterization of Fosmid #4-5



Inserted upstream of *RAG1*, in an alternate transcript predicted by unigene, and into an AluY
Opposite Orientation, not in dbRIP

Empty site:

5′ CGGTGAAACCCTGTCTCTACT**AAAAATACAAAAAATTAG**CCGGGCGTGGTGGCAGGCGCCTGT 3′
3′ GCCACTTTGGGACAGAGATGA**TTTTTATGTTTTTTAATC**GGCCCGCACCACCGTCCGCGGACA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 40 bp TSD Length: 18

Interrupted by a TA_3 repeat ($A_9(TAAA)_6TA_5T$)

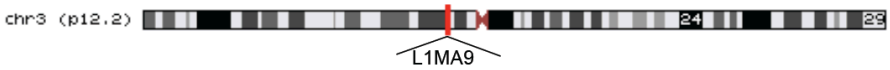
CTACT**AAAAATACAAAAAATTAGA**  **AAAAATACAAAAAATTAG**CCGGGCGTGG

Untemplated

Screened via sequencing

Activity- 1% of L1.3

Characterization of Fosmid #4-7



Inserted into an old LINE-1, distal to any known genes
Repeat rich region, not in dbRIP

Empty site:

5′ TCTGAAAACTAATGATGATG**AAATAAGATGATGA**AATAAATTGAAGTCACAAGTAAATGGAAAGATAT 3′
3′ AGACTTTTGATTACTACTAC**TTTATTCTACTACT**TTATTTAACTTCAGTGTTCATTTACCTTTCTATA 5′

EN Cleavage site: 5′ ATTT/C 3′ pA Length: 36 bp TSD Length: 14

Interrupted- $A_8TA_5TA_{21}$

ATGATGATG**AAATAAGATGATGA**  **AAATAAGATGATGA**AATAAATTGAAGT

Screened via sequencing

Activity- 69% of L1.3

Characterization of Fosmid #4-8

Inserted ~450 bp downstream of *SPEF2*, in opposite orientation
Gene involved in sperm flagellum, in opposite orientation- not in dbRIP

Empty site:

5′ AGGAAATATATTCTATT**AAGAAATAATTATGA**ATAGATGATGGGAACAAAGCTTTACTAGCAGT 3′
3′ TCCTTTATATAAGATAA**TTCTTTATTAATACT**TATCTACTACCCTTGTTTCGAAATGATCGTCA 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 57 bp TSD Length: 15

one A to G change

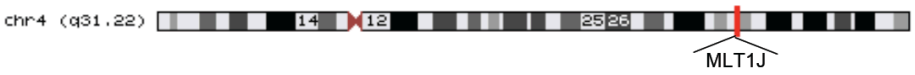
tctatt**AAGAAATAATTATGA**GAGA  **AAGAAATAATTATGA**atagatgat

Untemplated

Screened via sequencing

Activity- 161% of L1.3

Characterization of Fosmid #4-15



Inserted into an old MaLR LTR retrotransposon, in an annotated mRNA and a repeat rich region- not in dbRIP

Empty site:

5′ GAAACTGTGCCTCCACACTT**AAGAGAGAAAATTA**AAGTGTCTTTATTCATTAAGTTTCTGTTA 3′
3′ CTTTGACACGGAGGTGTGAA**TTCTCTCTTTTAAT**TCACAGAAATAAGTAATTCAAAGACAAT 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 95 bp TSD Length: 14

$A_{17}GAATTGTA_{12}TTATA_3TA_4CA_3GA_2GAATATGA_{32}$

TCCACACTT**AAGAGAGAAAATTA** **AAGAGAGAAAATTA**AGTGTCTT

Screened via sequencing

Activity- 54% of L1.3

Characterization of Fosmid #4-17

Inserted distal to known genes, not within a repetitive element- not in dbRIP

Empty site:

5′ TGAGCAGAAAAGAACATGTGCAACACTATATT**AAAAAAGAAGAGTA**TGAATTAGAAAAATATTCC 3′
3′ ACTCGTCTTTTCTTGTACACGTTGTGATATAA**TTTTTTCTTCTCAT**ACTTAATCTTTTTATAAGG 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 35 bp TSD Length: 14

Interrupted $A_7TTA_{18}TA_6T$

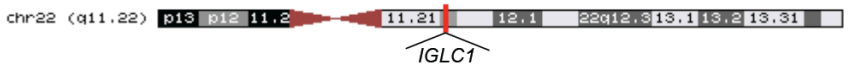
tatatt**AAAAAAGAAGAGTA**GGAGA **AAAAAAGAAGAGTA**tgaatta

Untemplated ACG

Screened via sequencing

Activity- 0% of L1.3

Characterization of Fosmid #4-19



Inserted within a region that potentially codes for regions of immunoglobulin lambda (Opp. Orientation)- no RefSeq gene. Did not insert into a repetitive element. Not in dbRIP

Empty site:

5′ GAATGCATCATGGGATGCTCATC**AAGAAATATGC**AACCCCTGCTGACACAGCTCCCTACATGT 3′
3′ CTTACGTAGTACCCTACGAGTAG**TTCTTTATACG**TTGGGGACGACTGTGTCGAGGGATGTACA 5′

EN Cleavage site: 5′ TCTT/G 3′ pA Length: 13 bp TSD Length: 11

GCTCATC**AAGAAATATGC**GATTCAGATGGTT **AAGAAATATGC**AACCCCT

Untemplated

Screened via sequencing

Activity- 0% of L1.3

Characterization of Fosmid #4-20

Inserted distal to known genes, within the end of an AluSc and close to an AT repeat region. Not in dbRIP

Empty site:

5′ CTGGGTGACAGTGCAAGACTCCATTT**AAAAAAAAAATTTA**TATATATATATAAAATCTGTATGTGT 3′
3′ GACCCACTGTCACGTTCTGAGGTAAA**TTTTTTTTTTAAAT**ATATATATATATTTTAGACATACACA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 58 bp TSD Length: 14

Interrupted $A_{12}TTA_{12}CA_5CA_5CA_4CA_4CA_9$

CCATTT**A**AAAAAAAAAATTTA[ ][ ]**AAAAAAAAAATTTA**TATATATATATAAA

Screened via sequencing Activity- 0% of L1.3

Characterization of Fosmid #4-27

Inserted within intron 1 of *FOXO3*, close to exon 2, in opposite orientation of the gene

FOXO3- Transcription of genes involved in apoptosis. Also known to be a breakpoint for MLL translocations that cause secondary acute leukemia.

Not in dbRIP

Empty site:

5′ GCAGGCCAGATGAGTTTTTATTTGCACGTT**AAAGACAAAAGATA**CTCCTGATGCTACTGTCCATATATTAAAT 3′
3′ CGTCCGGTCTACTCAAAAATAAACGTGCAA**TTTCTGTTTTCTAT**GAGGACTACGATGACAGGTATATAATTTA 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 82 bp TSD Length: 14bp

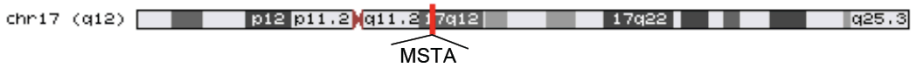
Interrupted- $A_{34}TAAAGTATCTCATAAACTTA_{29}$

tgcacgtt**AAAGACAAAAGATA**GATA[ ][ ]**AAAGACAAAAGATA**ctcctgat

Untemplated **7 bp 5′ trunc..**

Screened via sequencing Activity- 79% of L1.3

Characterization of Fosmid #4-29



Inserted distal to known genes, within an LTR retrotransposon (MSTA, Family MaLR).

Not in dbRIP

Empty site:

5′ GAAGGATGGTGCTAAACAAGTCAT**AAGAAACTGCCCCC**ATGATCCAATCTCTTCCCACCAGCC 3′
3′ CTTCCTACCACGATTTGTTCAGTA**TTCTTTGACGGGGG**TACTAGGTTAGAGAAGGGTGGTCGG 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 46 bp TSD Length: 14

Interrupted $A_{12}GA_3GGA_{28}$

AACAAGTCAT**AAGAAACTGCCCCC**[ ][ ]**AAGAAACTGCCCCC**ATGATCCAAT

Screened via sequencing Activity- 38% of L1.3

Characterization of Fosmid #4-32

Inserted within intron 4 of *TDRD5* (tudor domain containing 5) in the opposite transcriptional orientation. Inserted within an older LINE element, L1MC1. Not in dbRIP

Empty site:

5′ CAAAATACCTGAACAGACCTCTCACT**AAAGAAGATATACCAG**ATGGCAAGTAAGCATATGAAAAGAT 3′
3′ GTTTTATGGACTTGTCTGGAGAGTGA**TTTCTTCTATATGGTC**TACCGTTCATTCGTATACTTTTGTA 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 68 bp TSD Length: 16

Interrupted $A_{22}GA_{11}TTATCTGTTAGAATTCTGA_{16}$

Screened via sequencing Activity- 11% of L1.3

Characterization of Fosmid #4-34

Inserted distal to known genes, within the LTR retrotransposon PABL_A (an ERV1), and in a repeat-rich region. Not in dbRIP. HS oligo gave equivocal results

Empty site:

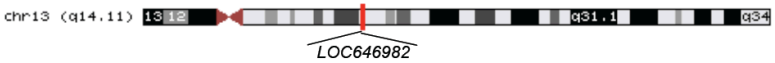
5′ AAAAATAAAAGTGCTTAAATACTTCAAC**AGAAAAAGGAAAGA**CTAGTCAAATGCTTTCTCAAGTTT 3′
3′ TTTTTATTTTCACGAATTTATGAAGTTG**TCTTTTTCCTTTCT**GATCAGTTTACGAAAGAGTTCAAA 5′

EN Cleavage site: 5′ TTCT/G 3′ pA Length: 31 bp TSD Length: 14

Interrupted- three interspersed ts

Screened via sequencing Activity- 57% of L1.3

Characterization of Fosmid #4-35



Inserted within intron 5 of *LOC646982*, not within a repeat (opposite orientation). This validated gene is the Twelve-thirteen translocation leukemia gene (non-coding RNA)

Not in dbRIP. Progenitor of this is likely on Chr. 4, 112848409- near an unannotated RIP

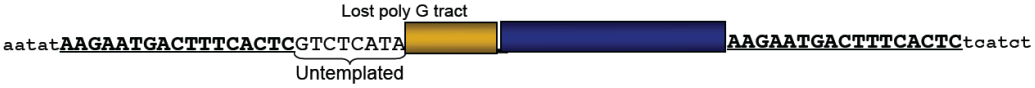
Empty site:

5′ TGAGACCAGATTTAGCTACTCTAATAT**AAGAATGACTTTCACTC**TCATCTTATCCGGGCCTCTTAGCA 3′
3′ ACTCTGGTCTAAATCGATGAGATTATA**TTCTTACTGAAAGTGAG**AGTAGAATAGGCCCGGAGAATCGT 5′

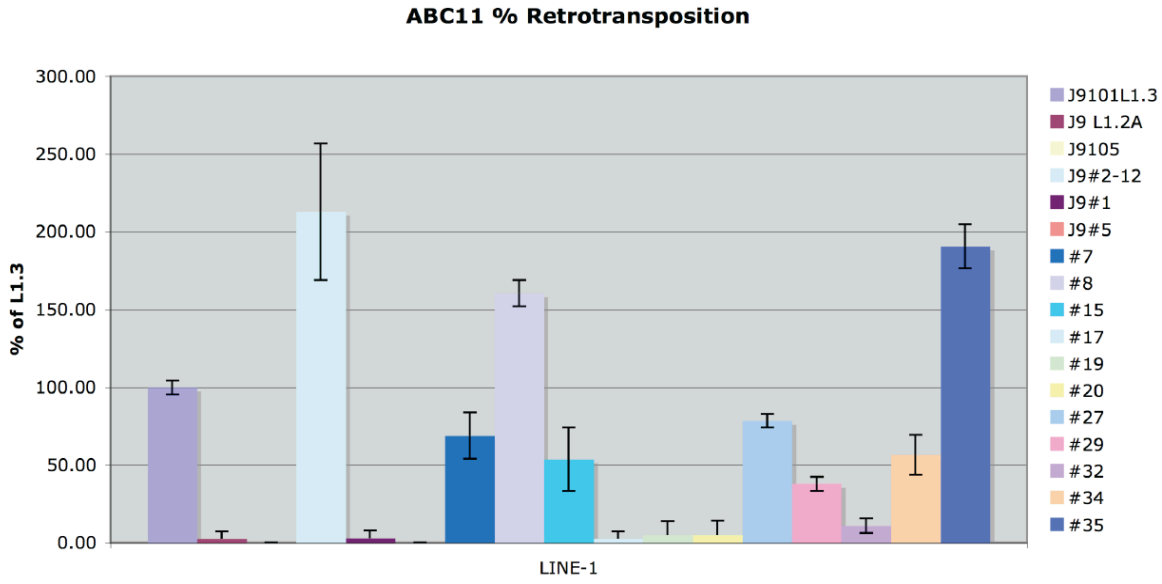
EN Cleavage site: 5′ TCTT/A 3′ pA Length: ~102 bp TSD Length: 17

Interrupted $A_{32}TAAAGTATCTCATAAACTTA_{50}$

Lost poly G tract



Screened via sequencing Activity- 191% of L1.3



ABC12

Characterization of Fosmid #5-36

Inserted within a non-human ref seq gene, same orientation. Inserted within an LTR retrotransposon. Also, ~2000 bp downstream of ref seq *C10ORF120* in opposite orientation.

Not in dbRIP

Empty site:

5′ AGCTTTTACATAGATAAGACCCTTGTAT**AAGAAAAACTTAAAG**ATGATGCATTCCTCTGCTTGCTTTCTGAG 3′
3′ TCGAAAATGTATCTATTCTGGGAACATA**TTCTTTTTGAATTTC**TACTACGTAAGGAGACGAACGAAAGACTC 5′

EN Cleavage site: 5’ TCTT/A 3’ pA Length: 83 bp TSD Length: 15

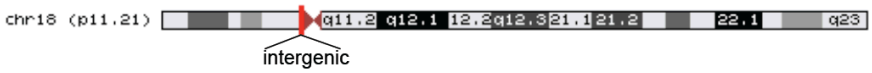
Interrupted A_{16}TTTAACAGATCTCTCCACACA$_{20}$GA_3TGA_{21}

CCCTTGTAT**AAGAAAAACTTAAAG**  **AAGAAAAACTTAAAG**ATGATGCA

Screened via sequencing

Activity- 1% of L1.3

Characterization of Fosmid #5-54



Inserted within an intergenic region, not within a repeat. In a Segmental Duplication.

No HS oligo sequence. Not in dbRIP

Empty site:

5′ TAATACTTCATTT**AAGAATAGTGTCCTC**TTCAAAATCGATTTGGGAATACTTACAGCTTG 3′
3′ ATTATGAAGTAAA**TTCTTATCACAGGAG**AAGTTTTAGCTAAACCCTTATGAATGTCGAAC 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 15 bp TSD Length: 15bp

Screened via sequencing

Activity- 0% of L1.3

Characterization of Fosmid #5-55

Inserted within intron 9 of predicted gene (further 3′ alt. transcript of) *ROR2* in the same orientation. Insertion occurs ~50kb downstream of RefSeq *ROR2*.

Not in dbRIP

Empty site:

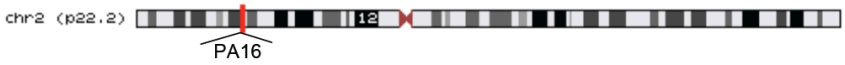
5′ TATTCTGTTCTTGCATTGCTGT**AAAGAAATACCTGAG**ACTGGGTTATTTATTTATTTATTATGAGAT 3′
3′ ATAAGACAAGAACGTAACGACA**TTTCTTTATGGACTC**TGACCCAATAAATAAATAAATAATACTCTA 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 32 bp TSD Length: 15

Screened via sequencing

Activity- 21% of L1.3

Characterization of Fosmid #5-58



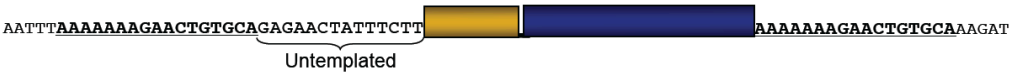
Inserted within an intergenic region, near the poly A tail of an L1PA16.

Not in dbRIP, however an unresolved L1 is close

Empty site:

5′ TGAAATTATAAAAAAAATAAAAATTT**AAAAAAAGAACTGTGCA**AAGATGTAAACATCTTTCTTCTT 3′
3′ ACTTTAATATTTTTTTTATTTTTAAA**TTTTTTTCTTGACACGT**TTCTACATTTGTAGAAAGAAGAA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 24 bp TSD Length: 17



Screened via sequencing

Activity- 13% of L1.3

Characterization of Fosmid #5-66

Inserted in an intergenic region, within an LTR retrotransposon.

Not in dbRIP, however an unresolved L1 is close

Empty site:

5′ CAATTAAAGTTCTTTCCTTT**ATAAATTACCCAGTCTC**AGGTATGTCCTAATAGCAGAGGGAGAAT 3′
3′ GTTAATTTCAAGAAAGGAAA**TATTTAATGGGTCAGAG**TCCATACAGGATTATCGTCTCCCTCTTA 5′

EN Cleavage site: 5' TTAT/A 3' pA Length: 41 bp TSD Length: 17

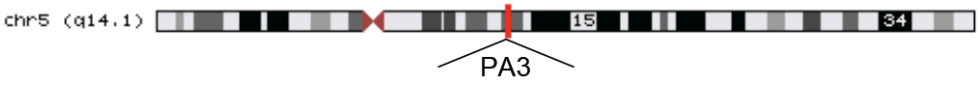
Interrupted by G and T

TTTCCTTT**ATAAATTACCCAGTCTC** **ATAAATTACCCAGTCTC**AGGTATG

Screened via sequencing

Activity- 23% of L1.3

Characterization of Fosmid #5-68



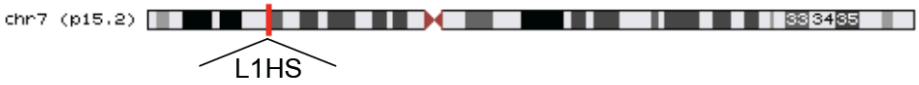
Older, L1PA3 in this location. In genome seq.

No 3'UTR out sequence

In this individual (ABC12), there is a lack of an *Alu* Ya5 element that interrupted the 5'UTR of this PA3 in HGR.

i.e., this individual is lacking a polymorphic *Alu* present at this site in the HGR.

Characterization of Fosmid #5-70



This element is in the reference sequence. Is a full-length Human Specific LINE-1. The element in the genome has no deletions/insertions and is 0.6% diverged, however, element has a nonsense mutation in ORF2- therefore not in Brouha paper. No Accession # for this fosmid sequence. The element from ABC12 is intact and jumps at ~8% of L1.3

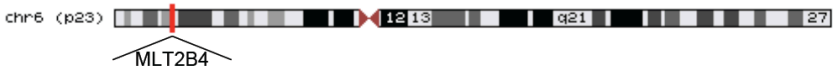
Not in dbRIP

5' Flanking Sequence into element-

TGCCAATCAAGCAGCCAGACAAAGTGACACACAACTGCAGTTCCCAGCTGCCTGGGNAGGCTGAGGCTAGGGGATCA
CTTGAGCTCAGAAGTTTGAGGCTATAGTGCACTATGATCTCTCCTATGAATAGCCACTGCACACCAGCCTCAGCAGCA
TAGTGAGACCTTGTCTCTAAAAACAACGATACACCAGCCAAAATCAATTATGACACTGGACATTTTCATGAGCCCAAGT
GTCCCTGTCATAGTTAAATACATAGTTTGTATTTCAGTTTTTGGAAGTATTGGAAACTAACTGGAGTACACTCACATCTG
CCTTTGCGGATACTCAATAGTCCAGAACAATTGACCTCCTTAACCATTATTCCCATTTGAAGGACAGATGGTAACTTTC
GAACAGCTTCTTCTCCCACATCTTCCCTCTGAGTTTCAGTTTCCAGGGAGAGGGACACTGAGTTTATAAAAAAGTTTGA
GTTGGGGGAGGAGCCAAGATGGCCGAATACGAACAGCTCCGGTCTACAGCTCCCAGGGTGAGCGACGCAGAAGACG
GGTGATTTCTGCATTTCC

Activity- 8% of L1.3

Characterization of Fosmid #5-77 **(same as #4-1)**



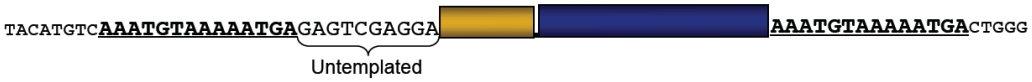
Inserted into an old ERVL, about 15,000 bp upstream of the *GFOD1* gene, opposite orientation.

Repeat rich region- IN dbRIP

Empty site:

5′ CAGCAGTGGTTAATTTTACATGTC**AAATGTAAAAATGA**CTGGGCCAGAAAGTGTCCAGATATTTGGTC 3′
3′ GTCGTCACCAATTAAAATGTACAG**TTTACATTTTTACT**GACCCGGTCTTTCACAGGTCTATAAACCAG 5′

EN Cleavage site: 5′ ATTT/G 3′ pA Length: 16 bp TSD Length: 14



Screened via sequencing

Activity- 0% of L1.3

Allele Freq. is at **58**% (Myers *et al.* 2002)

Characterization of Fosmid #5-82

Inserted within intron 49 of *RYR3*, not within a repetitive element and about 300bp upstream of exon 50 in the opposite transcriptional orientation as the gene.

Not in dbRIP

Empty site:

5′ TGGTGGAAATGCAGGTATCCTTGGAAAAT**GAAATAACACTATA**AATCCTCAAGTCACCATTTGCTTTA 3′
3′ ACCACCTTTACGTCCATAGGAACCTTTTA**CTTTATTGTGATAT**TTAGGAGTTCAGTGGTAAACGAAAT 5′

EN Cleavage site: 5′ TTTC/A 3′ pA Length: 21 bp TSD Length: 14 bp

Screened via sequencing

Activity- <1% of L1.3

Characterization of Fosmid #5-86

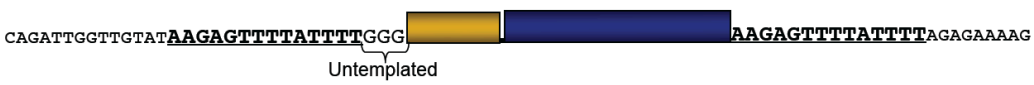
Inserted within *ADAMTS12* in intron 2, in same orientation, not within a repetitive element.

Not in dbRIP, unresolved L1s are close

Empty site:

5′ TCTTTCTCCTGTGTTTCAGATTGGTTGTAT**AAGAGTTTTATTTT**AGAGAAAAGATTTATCAAAGTGCTAG 3′
3′ AGAAAGAGGACACAAAGTCTAACCAACATA**TTCTCAAAATAAAA**TCTCTTTTCTAAATAGTTTCACGATC 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 24 bp TSD Length: 14 bp

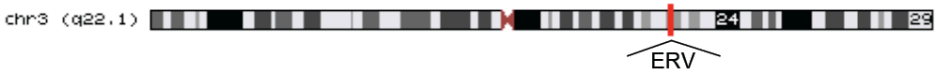


Screened via sequencing

Activity- 0% of L1.3

ABC13

Characterization of Fosmid #6-91



Inserted within the predicted gene *BC043572* in the opposite orientation, within an ERV1 (MER31B).

Directly downstream of *NEK11* (~500 bp away) in the same orientation.

Not in dbRIP

Empty site:

5′ ACTGTAATCAGGACTATTTTGATAGGTATC**AACAATATTGCA**ATATGGGAGAGGCATTGGGCTCAACTCTGAAG 3′
3′ TGACATTAGTCCTGATAAAACTATCCATAG**TTGTTATAACGT**TATACCCTCTCCGTAACCCGAGTTGAGACTTC 5′

EN Cleavage site: 5′ TGTT/G 3′ pA Length: 110 bp TSD Length: 12bp

$A_{12}GA_3GGA_{24}$GAACTTATACATCTTTCTTTTTGGTAGTTTTTGAAATAAAT
TTAACATAGCTATAGA$_{12}$

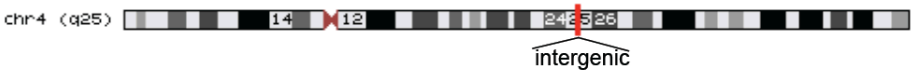
GATAGGTATC**AACAATATTGCA**GGA TD **AACAATATTGCA**ATATGGGAGAGGCAT

Untemplated

Screened via sequencing

Activity- 157% of L1.3

Characterization of Fosmid #6-97



Inserted within an intergenic region, not within a repetitive element, ~22kb upstream of *COL25A1* in the same orientation as the gene.

Not in dbRIP

Empty site:

5′ CAGATCCCTCAAATATAATGATGAG**AAAAAAAGGAGTACCTT**TATATGAAGTTAAAAAAATATAAA 3′
3′ GACTAGGGAGTTTATATTACTACTC**TTTTTTTCCTCATGGAA**ATATACTTCAATTTTTTTATATTT 5′

EN Cleavage site: 5′ TTTT/C 3′ pA Length: 38 bp TSD Length: 17

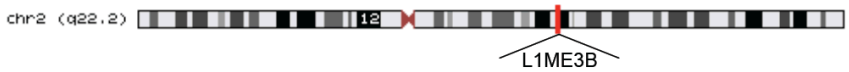
TAATGATGAG**AAAAAAAGGAGTACCTT**G **AAAAAAAGGAGTACCTT**TATATG

Untemplated

Screened via sequencing

Activity- 36% of L1.3

Characterization of Fosmid #6-100



Two L1HS elements in one fosmid. 'New' element is present in ABC13 only and not in dbRIP. 'Old' element is in HGR, and is polymorphic according to Boissinot *et al.*, 2004 paper.

New element- Occurred in an old L1, in an intergenic region.

Empty site:

5′ AAGACAAATTGTTAGAATATAT**AAGAAAATGTA**ATGAAGTTAATATATACAAGTGGAACATGAATTAA 3′
3′ TTCTGTTTAACAATCTTATATA**TTCTTTTACAT**TACTTCAATTATATATGTTCACCTTGTACTTAATT 5′

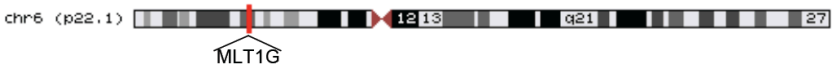
EN Cleavage site: 5′ TCTT/A 3′ pA Length: 249 bp TSD Length: 11

(transduction- A_{29}-common TD (same as 2-53)-A_{15})

AATATAT**AAGAAAATGTA** TD **AAGAAAATGTA**ATGAAGTTAA

Screened via sequencing Activity- both 0-1% of L1.3

Characterization of Fosmid #6-102



Inserted within predicted olfactory gene in the same orientation. Within an older LTR retrotransposon

Not in dbRIP

Empty site:

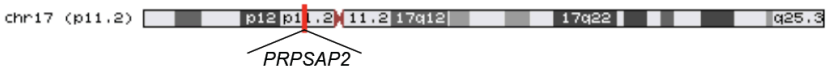
5′ TTAACTAACCAGATCAAAGTTTAGTAGCAT**AAAACTATATATATA**TATATATATGTGTGTGTGTATTTTTTTTA 3′
3′ AATTGATTGGTCTAGTTTCAAATCATCGTA**TTTTGATATATATAT**ATATATATACACACACACATAAAAAAAAT 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 39 bp TSD Length: 15

AGTTTAGTAGCAT**AAAACTATATATATA** **AAAACTATATATATA**TATATATATG

Screened via sequencing Activity- 129% of L1.3

Characterization of Fosmid #6-107



Inserted within intron 5 of *PRPSAP2* (phosphoribosyl pyrophosphate), in opposite orientation to the gene and only 500bp downstream of exon 5, not directly within a repetitive element.

Not in dbRIP

Empty site:

5′ CAATACTTGGTTATTTCCACCCGTT**AAGAACACTGAGTGGCTGG**ACACGGTGGCTCATACCTGTAATCCCA 3′
3′ GTTATGAACCAATAAAGGTGGGCAA**TTCTTGTGACTCACCGACC**TGTGCCACCGAGTATGGACATTAGGGT 5′

EN Cleavage site: 5′ TCTT/A 3′ pA Length: 34 bp TSD Length: 19

SMALL, 7BP 5′ TRUNCATION (loss of poly G tract and the AG that follows)

GTT**AAGAACACTGAGTGGCTGG**GTGGATT **AAGAACACTGAGTGGCTGG**ACA

Untemplated

Screened via sequencing Activity- 139% of L1.3

Characterization of Fosmid #6-109

Inserted within an intergenic region, within a L1MA1. Distal to any known genes.

Not in dbRIP

Empty site:

5′ TTTTAATTATTTCAATCTCTTTTTTTAAAAAAAATTATCTGTTAGAATTCTGAATTCCTTCTCTGTAT 3′
3′ AAAATTAATAAAGTTAGAGAAAAAAATTTTTTTTAATAGACAATCTTAAGACTTAAGGAACACAGATA 5′

EN Cleavage site: 5′ TTTT/A 3′ pA Length: 20 bp TSD Length: 15

CTCTTTTTTTAAAAAAAATTATCTG[gene][gene]AAAAAAAATTATCTGTTAGAATTCT

Screened via sequencing

Activity- 73% of L1.3

Characterization of Fosmid #6-113

Inserted in an intergenic region, within an LTR retrotransposon. Close to the the 3′ UTR of GenBank mRNA BX648305, downstream ~1.5kb in the same orientation. 5′ UTR AS sequencing failed due to deletion (see below)

Not in dbRIP

Empty site:

5′ GACCCTCACAGTGAGTGTTACAGCTCTTAAAGATTGTGTGTCCGGAGTTTGTTCCTTCAGATGTTCAGATGTG 3′
3′ CTGGGAGTGTCACTCACAATGTCGAGAATTTCTAACACACAGGCCTCAAACAAGGAAGTCTACAAGTCTACAC 5′

EN Cleavage site: 5′ CTTT/A 3′ pA Length: 26 bp TSD Length: 17 bp

~530 bp Deletion in the 5′UTR of bp 92 -622

AGCTCTT[AAAGATTGTGTGTCCGG][gene][gene][AAAGATTGTGTGTCCGG]AGTTTGTTCC

Screened via sequencing

Activity- 4% of L1.3
(tested without a CMV promoter)

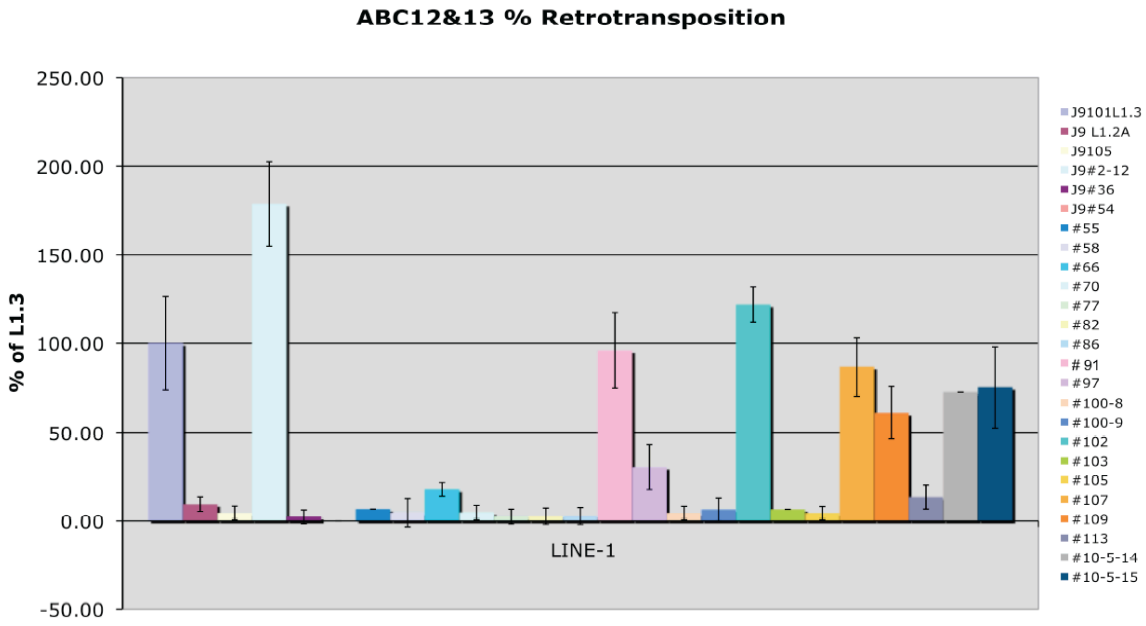


Table 2.4: L1 Insertions in Genes, Related to Figure 2.3

First column: the identifier of each element (as defined in Table 2.2 and Table 2.3). Second column: the RefSeq gene name. Third column: the intron containing the insertion, or the distance (5′ or 3′) from the L1 insertion to the nearest gene. Fourth column: the transcriptional orientation of the L1 with respect to the gene. Fifth column: if annotated, a short summary about whether mutations in the gene containing the insertion are implicated in human disease (http://www.ncbi.nlm.nih.gov/omim/).

Table 2.4: L1 Insertions in Genes, Related to Figure 2.3

L1 ID	Gene	Intron	Transcription Orientation	OMIM data
1-3	*RIMS1*	2	Same	Synaptic vesicle exocytosis- Cone-rod dystrophy
1-4	*NOS*	Upstream 15kb	Same	Neuronal/neuromuscular signaling
1-5	*ABCA1*	5	Opposite	Tangier disease-HDL deficiency
1-(2-1)	*NPAS3*	4	Opposite	Psychiatric illness- schizophrenia?
2-1	*PHACTR1*	5	Same	Expressed in brain
2-6	*SBF2*	6	Same	CMT 4B2
2-7	*MAMDC2*	8	Opposite	
2-12	*PRKCA*	3	Same	Numerous phenotypes associated-cancer, myocyte contraction, etc.
2-14	*SCN5A*	13	Opposite	LQT, Brugada syndrome, cardiac arrhythmia
2-30	*TTC28*	11	Same	
3-10	*SDK1*	5	Opposite	Neurite laminar targeting
3-17	*SLC10A7*	7	Opposite	
3-25	*KIF5C*	9	Opposite	Neuronal kinesin heavy chain
3-34	*RAD51L1*	5	Same	tumor-suppressor
3-39	*SPIRE1*	8	Opposite	Actin nucleation/ axis formation in oocytes and embryos
3-40	*CTNNA2*	2	Same	Motility in dendrite spines
3-15	*HOMER2*	2	Same	KO mice- cocaine sensitization, neuronal gene
4-1	*GFOD1*	Upstream 15kb	Opposite	
4-5	*RAG1*	Upstream	Opposite	
4-8	*SPEF2*	Downstream	Opposite	sperm flagellum
4-27	*FOXO3*	1	Opposite	Apoptotic gene transcription
4-32	*TDRD5*	4	Opposite	
4-35	*LOC646982*	5	Opposite	Twelve-thirteen translocation leukemia gene
5-36	*C10ORF120*	2 kb downstream	Opposite	
5-55	*ROR2*	9 (alt. transcript)	Same	Chondrocyte growth and patterning
5-82	*RYR3*	49	Opposite	
5-86	*ADAMTS12*	2	Same	Cell adhesion- implicated in cancer
6-91	*NEK11*	1 kb downstream	Same	DNA repair, S-phase checkpoint
6-97	*COL25A1*	22 kb upstream	Same	Brain-specific collagen, Alzheimer disease
6-107	*PRPSAP2*	5	Opposite	De novo synthesis of nucleotides/some amino acids

Table 2.5: **Allele Frequencies of L1s, Related to Figure 2.4**

First column: identifier of the L1. Second column: whether the element is active in the retrotransposition assay. Third column: Allele Frequency (calculated across all relevant genotyping panels). Fourth column: whether the allele frequency was determined in this study, and if not, the source of the allele frequency data, as well as whether the element was typed in the Zimbabwean panel or the HGDP. Fifth column: data from the *in silico* genotyping of each element. L1Hs elements in red text were either not assayed for retrotransposition in this study (5′ ATLAS only: #2-17 and 3-30), or were genotyped in other studies (Badge et al., 2003; Myers et al*.,* 2002; Xing et al., 2009).

Table 2.5: Allele Frequencies of L1s, Related to Figure 2.4

L1 ID	Active	Allele Freq.	Genotyping done here	*In Silico* Genotyping
1-2	N	0.11	yes	Insertion in G248, One End Anchored in ABC12, 8, Transchromosomal in ABC8
1-3	Y	0.21	yes	Insertion in G248 & ABC14, Inversion in ABC8
1-4	Y	0.44	yes	Insertion in G248, ABC9, ABC11 ABC12, 14
1-5	Y	0	Yes- HGDP	Insertion and Inversion in G248. Insertion in ABC14 (not there when PCR genotyped)
1-(2-1)	Y	-	-	Inversion G248, ABC8, 9
2-1	Y	0.178	yes	Insertion ABC9 & 11 Inversion ABC8
2-6	N	0.012	yes- Badge has .07	Insertion in ABC9 & 11
2-7	N	-	-	Insertion in ABC9 only
2-12	Y	0.223	yes	Insertion in ABC9 & 12 Inversion in ABC8
2-14	Y	0.266	yes	Insertion in ABC9 8, 11, 12, 13
2-17	-	0.605	yes	Insertion in ABC8, 9, 12, 14
2-21	Y	0.209	yes	Insertion in ABC9 & 11
2-24	N	-	-	Insertion in ABC9, 11, 12 Inversion and Deletion in ABC8
2-30	Y	0.0903	yes	Insertion in ABC9 only
2-32	Y	0.207	yes	Insertion in ABC9& 12, Inversion in ABC8
2-39	N	0.051	yes	Insertion in ABC9, 8, 10, Inversion in ABC13
2-42	-	-	-	Inversion in ABC12, 13 NO Insertion in ABC9
2-53	Y	0.063	yes	Insertion in ABC9, 11, 14
2-59	Y (low)	-	-	Insertion in ABC9, 11, 14
2-25	N	-	-	Insertion in G248, ABC9, 8, 12, 13 Inversion in ABC7
2-38	Y	-	-	Insertion in ABC9, 11, 12, Inversion in ABC7, 8, Deletion in G248
2-61	N	-	-	Insertion in ABC9 & 8
3-1	N	-	-	Insertion in ABC10 & 12
3-3	Y	0.264	yes	Insertion in ABC10, 9, 11, 13, Inversion in ABC8, One End Anchored in G248
3-4	Y	0.00249	Yes- Z	Insertion in ABC10, Inversion in ABC8

L1 ID	Active	Allele Freq.	Genotyping done here	*In Silico* Genotyping
3-7	N	-	-	Insertion in ABC10 and 13
3-10	Y	0.00249	Yes- Z	Insertion in ABC10, Transchromosomal (X) in ABC7
3-17	N	High	No- Xing	Insertion G248, ABC10, 8, 12, 13, 14, Inversion ABC8
3-18	Y	-	-	Insertion in ABC8, 10, 13, Deletion in ABC14
3-24	N	0	Yes- HGDP	Insertion in ABC10 only. Inversion in ABC7
3-25	Y	0.00249	Yes-Z	Insertion in ABC10. Inversion in ABC8.
3-30	-	0?	Yes- CEPH only	Insertion in ABC10 only, Inversion in ABC7, 8
3-31	Y	0.00174	Yes- HGDP	Insertion in ABC7, 10, Inversion in ABC12
3-34	Y	.078	Yes- Z	Insertion in ABC10, 8, inversion in ABC8
3-38	Y (low)	-	-	Insertion in ABC10, 11, 12, 14 Inversion in ABC9
3-39	Y	0.0155	yes	Insertion in ABC10, 11, 12, 13, Deletion in ABC14, Inversion in ABC7
3-40	N	.091	Yes- Z	Insertion in ABC10, 7
3-14	N	0.0944	yes	Insertion in ABC10, 13, Deletion in ABC8
3-15	N	0.43-0.45	No- Myers & Badge	Insertion in ABC10, 14, Inversion in ABC9, 11, 12, 13,14
3-36	N	-	-	Insertion in ABC9, 10, 13, 14
3-6	N	-	-	Insertion in ABC10, 13
3-(2-1)	Y (low)	-	-	Insertion in ABC10, 12, 13, G248
3-5	Y	0.00349	Yes- HGDP	Insertion in ABC10, Inversion in ABC7, Deletion in ABC12, Transchromosomal ABC8
4-1	N	0.58	No- Myers	Insertion in ABC8,10,11,12,13,14 Inversion in ABC8
4-5	Y (low)	-	-	Insertion in G248, ABC11, 13, Deletion in ABC14, Inversion in ABC7
4-7	Y	-	-	Insertion in ABC9, 11, 12
4-8	Y	-	-	Insertion in ABC11, 13, Deletion in ABC13, Inversion in ABC9
4-15	Y	-	-	Insertion in ABC11, Inversion in ABC8
4-17	N	-	-	Insertion in ABC11
4-19	N	-	-	Insertion in ABC8, 11, 14 Deletion in ABC9, 11, 12, 14 Inversion ABC8, 13
4-20	N	-	-	Insertion in ABC11 and G248

L1 ID	Active	Allele Freq.	Genotyping done here	*In Silico* Genotyping
4-27	Y	-	-	Insertion in ABC11
4-29	Y	-	-	Insertion in ABC11, Inversion in G248
4-32	Y	-	-	Insertion in ABC11, 10, 13, 14
4-34	Y	-	-	Insertion in ABC11, 7, 9, Inversion in G248, ABC7, 13
4-35	Y	-	-	Insertion in ABC11
5-36	Y (low)	-	-	Insertion in ABC8, 10, 12, 14, One End Anchored in ABC13, Inversion in ABC7
5-54	N	-	-	Insertion in ABC11, 12, 13, 14
5-55	Y	-	-	Insertion in ABC12, inversion in ABC10
5-58	Y	-	-	Insertion in ABC11, 12, 13, Inversion in G248
5-66	Y	-	-	Insertion in G248, ABC11, 12, 13, 14
5-77	N	0.58	No- Myers	Insertion in ABC8, 10, 11, 12, 13, 14, Inversion in ABC8
5-82	N	-	-	Insertion in ABC12, 13, Inversion in ABC12
5-86	N	-	-	Insertion in ABC12
6-91	Y	-	-	Insertion in ABC13, Deletion in ABC9
6-97	Y	-	-	Insertion in ABC10, 13, Inversion in ABC7
6-100	N	-	-	Insertion in ABC13 only
6-102	Y	-	-	Insertion in ABC9, 13
6-107	Y	-	-	Insertion in ABC10, 13, Inversion in G248, ABC7, 8, 12, 14
6-109	Y	-	-	Insertion in ABC13, Inversion in G248, ABC9
6-113	Y (low)	-	-	Insertion in ABC11, 13, 14

Table 2.6- Accession Numbers, Related to Figure 2.5

Fosmid Number	Accession Number	Fosmid Number	Accession Number	Fosmid Number	Accession Number
1-2	AC193155.1	3-4	AC203651	4-17	AC216905
1-3	AC195775.1	3-7	AC203650	4-19	AC226753
1-4	AC193146.1	3-10	AC203619	4-20	AC220069
1-5	GU477636	3-17	AC208506	4-27	AC215798
1-(2-1)	AC213207.1	3-18	AC225391	4-29	AC226751
2-1	AC206597	3-24	AC214167	4-32	AC216813
2-6	AC210891	3-25	AC206420	4-34	AC216112
2-7	AC212493	3-31	AC209305	4-35	AC217325
2-12	AC209294	3-34	AC210873	5-36	AC214986
2-14	AC211854	3-38	AC208067	5-54	AC226067
2-21	AC207709	3-39	AC206473	5-55	AC226114
2-24	AC208581	3-40	AC209201	5-58	AC209560
2-30	AC215719.2	3-14	AC204956	5-66	AC229893
2-32	AC214812	3-15	AC203593	5-77	AC208509
2-39	AC209235	3-36	AC207480	5-82	AC206103
2-42	AC208589	3-6	AC204965	5-86	AC236757
2-53	AC207969	3-5	AC204967	6-91	AC226062
2-59	AC209341	3-(2-1)	AC203592	6-97	AC236929
2-25	AC210912	4-1	AC217408	6-100	AC219161
2-38	AC207965	4-5	AC217815	6-102	GU477637
2-61	AC209421	4-7	AC217244	6-107	AC216987
3-1	AC203662	4-8	AC215801	6-109	AC216964
3-3	AC203635	4-15	AC216136	6-113	AC225317

Table 2.6- Accession Numbers, Related to Figure 2.5

Accession numbers refer to the fosmid sequence generated at the Washington University Genome Sequencing Center (St. Louis) (Kidd et al., 2008). Each L1 also was sequenced in its entirety as part of this study. Three elements (#1-3, 2-12, and 4-5) were found to contain changes relative to the NCBI entries. These sequences are available upon request. Two new L1 sequences recently deposited in GenBank are displayed in red text.

References

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 Dev *20*, 210-224.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

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 Res *32*, 3846-3855.

Babushok, D.V., and Kazazian, H.H., Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat *28*, 527-539.

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

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 Res *18*, 1875-1883.

Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S., and Devine, S.E. (2004). Natural genetic variation caused by transposable elements in humans. Genetics *168*, 933-51.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R*., et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Boeke, J.D. (2003). The unusual phylogenetic distribution of retrotransposons: a hypothesis. Genome Res *13*, 1975-1983.

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

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

Boissinot, S., and Furano, A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol *18*, 2186-2194.

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. Am J Hum Genet *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. Proc Natl Acad Sci U S A *100*, 5280-5285.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A.*, et al.* (2002). A human genome diversity cell line panel. Science *296*, 261-262.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M.*, et al.* (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol *311*, 17-40.

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

Cordaux, R., Hedges, D.J., and Batzer, M.A. (2004). Retrotransposition of Alu elements: how many sources? Trends Genet *20*, 464-467.

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.

Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd, and Edgell, M.H. (1992). Master genes in mammalian repetitive DNA amplification. Trends Genet *8*, 307-311.

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

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci U S A *90*, 6513-6517.

Donahue, W.F., and Ebling, H.M. (2007). Fosmid libraries for genomic structural variation detection. Curr Protoc Hum Genet *Chapter 5*, Unit 5 20.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Felsenstein, J. (1989). PHYLIP- Phylogeny Interference Package (Version 3.2). Cladistics *5*, 164-166.

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.

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques *17*, 46, 48-49, 52.

Frikke-Schmidt, R. (2010). Genetic variation in the ABCA1 gene, HDL cholesterol, and risk of ischemic heart disease in the general population. Atherosclerosis *208*, 305-316, Published online June 11, 2009.

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

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

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

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

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 Res *11*, 321-338.

Han, K., Xing, J., Wang, H., Hedges, D.J., Garber, R.K., Cordaux, R., and Batzer, M.A. (2005). Under the genomic radar: the stealth model of Alu amplification. Genome Res *15*, 655-664.

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. Nat Genet *7*, 143-148.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics *8*, 460.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

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.

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.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F.*, et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet *8*, 1557-1560.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol *16*, 111-120.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L.*, et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420-426.

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

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.

Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res *14*, 2253-2260.

Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A., and Moran, J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am J Hum Genet *73*, 1431-1437.

Marchani, E.E., Xing, J., Witherspoon, D.J., Jorde, L.B., and Rogers, A.R. (2009). Estimating the age of retrotransposon subfamilies using maximum likelihood. Genomics *94*, 78-82.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C.*, et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res *19*, 1527-1541.

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., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208-212.

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. Nat Genet *31*, 159-165.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V.*, et al.* (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet *71*, 312-326.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res *11*, 2050-2058.

Penzkofer, T., Dandekar, T., and Zemojtel, T. (2005). L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res *33*, D498-500.

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

Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet *70*, 841-847.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Second Edition edn (Cold Spring Harbor, New York, Cold Spring Harbor Laboratory press).

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. Nat Genet *16*, 37-43.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A *103*, 6611-6616.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., and Swergold, G.D. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res *10*, 1496-1508.

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

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

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.

Tarraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldon, T., Dopazo, J., and Dopazo, H. (2007). Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. Nucleic Acids Res *35*, W38-42.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.*, et al.* (2005). Fine-scale structural variation of the human genome. Nat Genet *37*, 727-732.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., and Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat *27*, 323-329.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y.*, et al.* (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60-65.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem *284*, 435-438.

Weichenrieder, O., Repanas, K., and Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure (Camb) *12*, 975-986.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T.*, et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872-876.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res *19*, 1516-1526.

Chapter 3

A Natural LINE-1 Mutation Spectrum

Abstract

We previously characterized 68 distinct full-length L1 polymorphisms, 37 of which were highly active in a cell culture retrotransposition assay. Here, we examined the 31 elements from this study that were unable to retrotranspose or had low-level activities. From the sequences of these L1s, we determined that ~50% contained chain terminating frame shift or nonsense mutations in one or both of the two L1-encoded open reading frames (ORFs). The additional 16 L1s contained intact ORF1p and ORF2p coding sequences. However, these elements consisted of a 5′UTR splicing mutation (#6-113), an L1 with an endonuclease mutation (#3-24), and 14 L1s with missense mutations in one or both ORFs that may impact retrotransposition. We examined the amino acid sequences of these 14 L1s to determine potentially causative differences, and have tested the elements in a functional assay to elucidate the endonuclease function of the L1s. Amino acid changes potentially responsible for low-level retrotransposition activity have been identified for at least 2 elements, and additional analysis examined the importance of conserved RNA changes and the utility of transduction sequences in the identification of highly active L1s.

Introduction

Long interspersed element-1 (LINE-1 or L1) sequences comprise ~17% of human DNA and play an important part in the evolution of the human genome (Lander et al., 2001). Moreover, it is estimated that ~80-100 L1s per human genome retain the ability to mobilize via a copy and paste mechanism known as retrotransposition (Brouha et al., 2003). The proteins encoded by retrotransposition-competent L1s are also responsible for the mobility of other non-autonomous retrotransposons, including Alu, SVA, and U6 (Buzdin et al., 2002; Dewannieux et al., 2003; Garcia-Perez et al., 2007; Hancks et al., 2011; Ostertag et al., 2003), and for processed pseudogene formation by the mobilization of cellular mRNAs (Esnault et al., 2000; Wei et al., 2001). Recently, a number of groups have identified many L1s and Alus that are polymorphic in human populations (Beck et al., 2010; Ewing and Kazazian, 2011, 2010; Huang et al., 2010; Iskow et al., 2010; Mills et al., 2011; Xing et al., 2009). Additionally, L1-mediated insertions have resulted in ~65 known cases of human disease (Babushok and Kazazian, 2007; Goodier and Kazazian, 2008; Kazazian et al., 1988). Therefore, L1s and the elements they mobilize constitute ~1/3 of our genome, and continue to affect the DNA of modern humans.

Retrotransposition-competent L1s (RC-L1s) are ~6kb in length, contain two non-overlapping open reading frames (ORFs) and terminate in a 3′UTR that is punctuated by a poly(A) tail (Dombroski et al., 1991; Scott et al., 1987). The 5′UTR contains an internal RNA polymerase II promoter (Swergold, 1990) that drives transcription of the L1. ORF1 encodes an ~40 kDa nucleic acid binding

protein (ORF1p) with a coiled coil domain (Holmes et al., 1992), RNA recognition motif (Khazina and Weichenrieder, 2009), and a carboxyl-terminal basic domain (Khazina et al., 2011; Moran et al., 1996). ORF2 encodes an ~140 kDa protein (ORF2p) (Ergun et al., 2004) with endonuclease (EN), reverse transcriptase (RT), and cysteine-rich (C) domains that are required for retrotransposition (Fanning and Singer, 1987; Feng et al., 1996; Martin et al., 1995; Mathias et al., 1991; Moran et al., 1996) (Figure 3.1).

Upon transcription from the 5'UTR, the L1 RNA is exported to the cytoplasm where translation of ORF1p and ORF2p occurs (Alisch et al., 2006; Leibold et al., 1990; McMillan and Singer, 1993). The two L1-encoded proteins then bind back to their encoding RNA by the process of *cis*-preference (Esnault et al., 2000; Kulpa and Moran, 2006; Wei et al., 2001). The resultant ribonucleoprotein particle (RNP) is a presumed retrotransposition intermediate (Hohjoh and Singer, 1996; Kulpa and Moran, 2005; Martin, 1991). The L1 RNP then enters the nucleus through a process that may occur independent of nuclear envelope break down (Kubo et al., 2006). Once in the nucleus, the EN domain of ORF2p cleaves the genome at a loose consensus site: 5′-TTTT/A-3′, with the "/" representing the scissile phosphate (Cost and Boeke, 1998; Feng et al., 1996; Morrish et al., 2002). This cleavage exposes a free 3′-OH that serves as a primer for the reverse transcription of L1 RNA and the formation of a cDNA L1 copy (Cost et al., 2002; Feng et al., 1996; Kulpa and Moran, 2006; Luan et al., 1993). In addition to the cleavage and first-strand cDNA synthesis steps, the more poorly understood steps of second strand cDNA synthesis and the integration of

the nascent L1 into a genomic location must also occur. This mechanism of retrotransposition is termed target-site primed reverse transcription (TPRT), and it results in a new L1 insertion flanked by variable length and sequence target-site duplications (TSDs) (Cost et al., 2002; Feng et al., 1996; Luan et al., 1993).

L1s are present in every mammalian genome studied to date. In primates, L1s have arisen in a single lineage over the last ~40 million years of evolution, with the newer active subfamily subsequently replacing the previously active subfamily (Boissinot and Furano, 2001; Khan et al., 2006). Human specific L1 (L1Hs) elements can be stratified into several subfamilies (pre-Ta, Ta-0, Ta-1, Ta1-d, Ta1-nd) based upon the presence of diagnostic sequence variants contained within their 5′ and 3′ UTRs (Boissinot et al., 2000; Scott et al., 1987; Skowronski et al., 1988; Smit et al., 1995). The classification of human L1s (Boissinot et al., 2000) and the development of consensus sequences from older elements present in the human genome reference sequence (HGR) have allowed investigation into L1 evolution (Boissinot and Furano, 2001; Khan et al., 2006; Lander et al., 2001).

Mutations in conserved residues of ORF1p and ORF2p are associated with a loss or decrease in retrotransposition (Doucet et al., 2010; Kulpa and Moran, 2005; Moran et al., 1996). Indeed, many of the critical functions of the L1-encoded proteins have been defined through the mutation of highly conserved residues (Feng et al., 1996; Khazina et al., 2011; Khazina and Weichenrieder, 2009; Moran et al., 1996; Weichenrieder et al., 2004). However, it is also possible that non-conserved amino acid sequences that separate the highly

conserved and well-defined regions of ORF1p and ORF2p may be crucial for efficient retrotransposition. In support of this hypothesis, phylogenetic analyses of L1s in the HGR indicated that the coiled coil domain of ORF1p displays a strong signature of positive selection in the recent evolutionary history of primate L1 elements, indicating that this region may interact with host proteins (Boissinot and Furano, 2001; Khan et al., 2006). In addition, some amino acid residues of ORF1p and ORF2p that drastically affect retrotransposition when mutated are not highly conserved in mammals (*e.g.,* the putative leucine zipper domain of ORF1p) (Doucet et al., 2010). We therefore hypothesized that profiling the natural mutations present in the L1-encoded proteins of inactive elements in conjunction with evolutionary comparisons could potentially indicate functionally relevant domains of ORF1p and ORF2p.

We previously generated a data set of 68 L1Hs elements that are polymorphic in human populations with respect to the HGR (Beck et al., 2010). Here, we have investigated the inactive elements from this study using nucleotide and amino acid sequence alignments. We identified 16 inactive or low-level activity elements that contained intact ORF1p and ORF2p, and in one of the L1s identified a splicing-mediated deletion of 5'UTR sequence. We also documented potentially deleterious amino acid differences in the two ORFs of 15 remaining L1s. The alignments developed here may allow identification of highly active elements from either the nucleotide or amino acid sequences of ORF1p and ORF2p. Additionally, a neighbor-joining tree indicated that highly active L1s

may be overrepresented within transduction families, and we show that transduction sequences can be utilized to identify highly active progenitor L1s.

Results

Identification of Inactive Polymorphic L1s with Intact ORFs

To identify mutations that were potentially responsible for the inactivity of polymorphic L1s, we utilized a previously developed dataset of 68 elements from fosmid libraries of 6 geographically diverse individuals (Beck et al., 2010; Kidd et al., 2008). High quality capillary sequencing of the 68 L1s was generated and compared to the sequence of the fosmid from which each was cloned to ensure an accurate representation of each element. The L1s were cloned from fosmids without the use of PCR and tested for their ability to retrotranspose in a cultured cell assay (Beck et al., 2010; Moran et al., 1996). Of the 68 L1s, 37 were highly active (displayed >10% the activity of a known, highly active element, L1.3-accession number L19088 (Dombroski et al., 1993)) and 6 additional elements displayed low-level activities (retrotransposition efficiencies <6% of L1.3). The 68 L1s in this dataset were relatively young in comparison to other studies, both with respect to the allele frequency of the elements in the population (Brouha et al., 2002), and via a maximum likelihood estimate of L1 age (Marchani et al., 2009). Therefore, the 31 elements with low-level or no activity (Beck et al., 2010) may contain mutations in crucial amino acid residues of the two L1-encoded proteins that rendered them unable to retrotranspose.

We first determined if any of the inactive or low-level activity L1s contained intact open reading frames. One L1, #2-42, contained an extra restriction site

within ORF2p that rendered the element unable to be cloned in the same manner as the other 67, and was therefore not tested for retrotransposition activity and is excluded from this study. Fourteen of the 30 remaining elements were rendered inactive due to in frame nonsense or frame shift mutations in one or both of the L1-encoded ORFs (Figure 3.1). The remaining 16 L1s potentially contain missense mutations that inactivated ORF1p or ORF2p.

Next, we developed a nucleotide alignment of the 37 highly active elements and the 16 additional L1s that contained intact ORFs (data not shown). The alignments of the 53 elements were also compared to a consensus sequence of highly active elements (Beck et al., 2010) and to consensus sequences of PA2-PA5 elements in the HGR (Boissinot and Furano, 2001). This L1 sequence alignment identified a large 524 bp deletion in the 5′UTR of element #6-113 (from bp 98 to 621 of L1.3 (Dombroski et al., 1993)).

A 5'UTR Splicing Mutation Responsible for Low Retrotransposition Activity

We next sought to determine how the #6-113 5'UTR deletion was generated, and if the low level activity of the L1 was due to the loss of promoter activity. Intriguingly, a nucleotide alignment revealed the GT immediately following the upstream junction in the 57 other L1s to be a previously described splice donor (SD) (Figure 3.2A) (Belancio et al., 2006; Belancio et al., 2008). Upstream of the 3' junction resides a potential splice acceptor (SA) site. The putative SD and SA sequences were examined using the Berkeley Drosophila Genome Project splice site prediction webpage (http://www.fruitfly.org/seq_tools/splice.html) that uses the NNSPLICE 0.9

prediction program trained with human sequences (Reese et al., 1997). The SD and SA at bp 98-99 and 620-621 scored above the default 0.4 cut off value, indicating that these positions are good candidates for functional splice sites (see Table 3.1 for all sites above 0.2). The SD and SA sequences flanking the junction in the 5'UTR of #6-113 suggested that a full-length L1 with intact 5'UTR was transcribed, spliced in the suggested manner, and inserted into the genome by TPRT at a new location with the 5'UTR deletion (Figure 3.2A). In this model, the ORFs of the L1 would be competent for retrotransposition. *In silico* genotyping analysis (Beck et al., 2010) showed that insertion fosmid(s) mapped to the location of the #6-113 insertion in 3 of the 9 individuals in the human genome structural variation project (including the library in which it was discovered, ABC13). Therefore, this L1 is not likely to be a private insertion (Beck et al., 2010; Kidd et al., 2008).

To determine if the 5'UTR deletion observed in element #6-113 was a common L1 splice variant, we conducted BLAT searches using either 60 or 400 bp flanking the 5'UTR splice junction to query the UCSC genome browser (hg19/GRCh37 human genome release http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) (Kent, 2002; Kent et al., 2002). The >100 potential spliced L1 sequences and their genomic flanks were obtained from the RepeatMasker track (Jurka et al., 2005) of the HGR for additional analysis. Sequences were then aligned to #6-113 to determine whether they contained the same splicing event and to assess the presence of TSDs. Ninety-six additional L1s contained the same junction as in #6-113, and at least 49 of the 96 L1s

contained TSDs (Table 3.2). An alignment of 30 validated 5'UTRs with the #6-113 splice junction is shown in Figure 3.2B. Interestingly, one of the L1Hs elements found in the HGR was previously found to contain two intact ORFs, was present at an allele frequency of ~0.87, and retrotransposed with low-level activity in a cultured cell assay (~0.3% of a reference element) (Brouha et al., 2003).

The L1s found in the BLAT search were all L1Hs or PA2-6 elements, which prompted examination of the SD and SA site conservation through primate L1 evolution. The SD is conserved in consensus sequences of L1Hs through L1PA10 elements (Khan et al., 2006), whereas the SA is conserved through L1PA6 elements. Therefore, the evolutionary conservation of the #6-113 splice site is consistent with the range of L1s found to contain the same junction (data not shown). Additionally, we compared the number of L1s with the splice junction to the known number of full-length elements from L1Hs or PA2-6 present in the HGR (Lander et al., 2001; Myers et al., 2002; Song and Boissinot, 2007). Spliced L1s from PA2, PA4 and PA5 families constituted ~1% of the full-length members of a given family of L1. There were 8 L1Hs elements that contained the junction, constituting ~2.7% of the full-length PA1 elements in the HGR. Additionally, there were more spliced PA3 L1s than any other family, and these 46 elements corresponded to ~3.0% of the total full-length PA3s. Conversely, the one L1PA6 element we found represented only ~0.1% of the total full-length PA6 elements in the HGR. However, though the PA1 estimate may be inaccurate due to L1Hs polymorphisms, and the PA6 estimate may be low due to other 5'UTR changes

making identification of spliced sequences difficult, the overrepresentation of PA3 elements is intriguing and warrants further study.

Element #6-113 was previously tested with its truncated 5'UTR driving transcription and was found to retrotranspose at ~4% of L1.3 (Beck et al., 2010). Next, we used an exogenous promoter to determine if the ORFs of this L1 were capable of high levels of retrotransposition. Cloning a CMV promoter upstream of #6-113 increased the retrotransposition efficiency of this element to >20% of our reference element L1.3 in the same vector (Figure 3.2C) (pCEP4, see Experimental Procedures). With a CMV promoter upstream of an L1, the retrotransposition efficiency of an element completely lacking a 5'UTR is significantly lower than an L1 containing a full-length 5'UTR (~40% decrease- Peter Larson and Aurelien Doucet unpublished data). Therefore, the low-level activity of #6-113 is primarily due to a promoter deletion rather than a deficiency in ORF1p or ORF2p function. The 15 additional L1s that were intact and inactive had similar retrotransposition efficiencies in CMV- and + contexts, and therefore may contain amino acid changes responsible for their inactivity in cell culture.

A Primary Screen for Causative Amino Acid Changes in Inactive L1s

To determine the residues potentially responsible for the inactivity or low-level activity of some L1Hs elements, we developed ORF1p and ORF2p amino acid alignments for the 37 highly active, 16 intact inactive elements, L1.3, and 5 consensus sequences (59 elements in total- see Figure 3.3). Highly active elements (*e.g.,* L1.3) may contain many differences from a “hot” L1 consensus sequence (Beck et al., 2010), and some amino acid changes were present in a

number of L1s. Therefore, amino acid changes present in both highly active L1s and inactive elements were unlikely to be the cause of an element's inability to retrotranspose in the cell culture assay, and have been eliminated from further study.

Amino acid sequences of ORF1p and ORF2p were examined for residues in the inactive/low-level activity elements that were different from the highly active L1s in the study. Amino acid residues identified in this comparison represent the subset of changes in 15 L1s that is potentially responsible for the lack of retrotransposition in cultured cells (Table 3.3). To prioritize the changes that were potentially most deleterious to protein function, differences identified in the 15 L1s were compared to the evolutionary conservation of ORF1p and ORF2p derived from diverse L1 and L1-like elements (Moran and Gilbert, 2002). As previous alignments lacked regions of ORF2p, some of the inactive L1s did not contain amino acid changes in conserved residues. Therefore, new alignments of ORF1p and ORF2p were created from evolutionarily diverse species (Figure 3.4). The new alignments contained more species than those previously generated, and when available, were created with consensus sequence files or the sequences of L1s active in cell culture (Experimental Procedures). Comparison of all amino acid differences in the intact L1s to the alignments in Figure 3.4 yielded 1-7 differences in conserved residues (through L1 *Canis lupus familiaris* - L1_Cf) (Table 3.3) that were potentially responsible for element inactivity. Interestingly, many of these residues are in the endonuclease domain of ORF2p.

Inactive L1s Encode Potential Endonuclease Domain Mutations

From the comparison of the 15 intact ORF L1 elements to the 37 highly active L1s, we determined that at least 7 elements contained mutations that could potentially render the L1 EN domain inactive (Table 3.3- conserved residue changes from 1-240 in ORF2p). L1s with mutations in conserved catalytic and structural residues of this domain are unable to jump in HeLa cells, yet retain the ability to retrotranspose in cells deficient in the non-homologous end-joining (NHEJ) pathway of DNA repair (Morrish et al., 2002). Endonuclease independent (ENi) retrotransposition was examined in NHEJ-incompetent XR-1 cells lacking the *XRCC4* gene product, and the parental (NHEJ competent) 4364a CHO cell line (Morrish et al., 2002). Retrotransposition was detected as a function of the reverse transcription, integration, and expression of the *mblastI* reporter cassette; which confers blasticidin resistance to cells with a new L1 integrant (Moran et al., 1996; Morrish et al., 2002; Wei et al., 2000) (see Experimental Procedures).

We tested the 15 inactive elements in the endonuclease-independent (ENi) retrotransposition assay to discern whether the L1s contained deficiencies in the EN domain. Remarkably, this modified retrotransposition assay indicated that 11 elements were able to jump with high efficiency (>20% of L1.3) in an ENi manner (Figure 3.5 and Table 3.3) (Morrish et al., 2002). To confirm that a given mutation was causative of the ENi retrotransposition in Figure 3.5, the mutation needs to be isolated from other changes in the L1 by introduction into a known retrotransposition-competent element. One of the 15 L1s had already been validated in this manner (#3-24). In this case, introduction of the S228P mutation

into ORF2 of L1.3 conferred the inability to jump in HeLa cells and ENi retrotransposition activity to this L1 when tested in XR-1 cells (Beck et al., 2010) (Figure 3.5).

Predictions of Functional Amino Acid Changes

Using both the alignment data and the results of genetic and biochemical functional assays, determining the amino acid changes likely responsible for L1 inactivity may be possible. Combining this data showed that 7 of the inactive L1s contained differences in conserved residues of the EN domain, 2 others contained changes in non-conserved EN residues; 6 of these 9 L1s retrotransposed in the ENi retrotransposition assay. Additional elements that would be predicted to retrotranspose in an ENi manner (*e.g.,* #3-36 or 4-20) also contained mutations in conserved residues of the RT or C domains. Thus, some L1s from this study may contain mutations in two or more domains important for retrotransposition. In contrast, some of the 14 L1s with amino acid changes from highly active L1s may have only one difference that is responsible for their inactivity in HeLa cells (similar to #3-24). For example, elements #2-59 (T44I) and #3-38 (T192P) are good candidates for EN mutations. Therefore, analysis of amino acid changes in conjunction with functional assays can identify potential causative mutations in inactive L1 sequences.

Amino Acid Changes and Highly Active L1s

Although L1 alignments were created to examine amino acid mutations that have inactivated intact elements, these resources can potentially identify changes that are shared between highly active L1s. In Figure 3.3, L1s that

retrotranspose at greater than 115% the efficiency of a reference element (L1.3, shown at top) are highlighted in yellow. Notably, these elements tend to be the most dissimilar from PA2-PA5 L1s, and are very close to a consensus sequence of active L1 elements (Beck et al., 2010). However, there is no consistent difference that is indicative of an adaptive change in the amino acid sequences of highly active L1s. Conversely, RNA changes in the 5' and 3'UTRs have been used as diagnostic sequences to identify human-specific L1s, and thus differ from older elements that are now inactive in human genomes. Therefore, conserved RNA changes may have been advantageous to L1Hs elements.

Highly Active L1s and Conserved RNA Changes

A consensus of L1s in the human genome contains a GAG at bp 5929-5931 of the 3'UTR (nucleotide position from L1.3) (Scott et al., 1987). However, most L1Hs elements contain an ACA at this position of the 3'UTR, which characterizes the transcribed, subset A or L1 Ta subfamily (Skowronski et al., 1988). In the dataset examined here, 35 of the 37 highly active elements contain an ACA at this position, and the other two contain the intermediate ACG, indicative of the human-specific pre-Ta subfamily L1s. The ACA/GAG distinction between human L1s and older elements, and the prevalence of ACA-containing L1s in disease-causing mutations (only one mutagenic human L1 insertion is a pre-Ta element (Kazazian et al., 1988)), led us to question if the difference in the 3'UTR is an advantage that allowed L1Hs elements to proliferate. Alternately, the three-nucleotide ACA difference may be a passenger mutation that happened to occur

in an element in a permissive expression context, or with other advantageous amino acid differences (see above).

To test this hypothesis, we created allelic versions of L1.3 with an ACA or a GAG in the 3'UTR. Other than this difference, there is no other nucleotide change between the elements. Assays comparing the two elements (n=4) showed that L1.3 GAG retrotransposed at ~95% of the activity of L1.3 ACA in HeLa cells (this difference is not significant). Therefore, in this assay the ACA 3'UTR difference is not necessarily advantageous for retrotransposition (data not shown).

Phylogenetic Trees Indicate Clustering of Related L1 Sequences

We next developed a nucleotide alignment of the 53 L1s with intact ORFs. These alignments were subsequently used to generate a phylogenetic tree using MEGA4 (Tamura et al., 2007). This tree was generated with the complete deletion option, which eliminates all gaps from the comparison and allowed #6-113 to be placed on the phylogeny. A bootstrap consensus neighbor-joining tree of the elements in this study is shown in Figure 3.6A.

Interestingly, this tree showed clustering of L1s with the same 3′ transductions or from the same transduction "subfamily" (*e.g.,* #2-12, 1-2-1, and 2-1) (Beck et al., 2010). Transductions occur by the read-through of the endogenous L1 poly (A) tail and the retrotransposition of 3′ flanking genomic sequence (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000). The phylogeny construction was independent of 3′ transduction sequence, but recapitulated element relationships that can be inferred from the

likely transfer of transductions from progeny to daughter elements. This clustering of transduction subfamilies is indicative of correct ordering of taxa, and unlike other alignments is constructed with the inclusion of #6-113 (Beck et al., 2010).

Each of the transduction clusters shown in Figure 3.6A contains one or more highly active L1s (lightly shaded L1s). Indeed, highly active elements are over represented within transduction subfamilies (17/26 vs. 37/68 or 68% of L1s with transductions vs. 54% of the total). Therefore, examining genomes for L1s from active transduction subfamilies is a potentially useful method for highly active element discovery.

TS-ATLAS Identifies a Potential RP Progenitor Element

Transduction subfamilies include some of the most highly active L1 elements (Beck et al., 2010; Brouha et al., 2002; Kidd et al., 2010). Moreover, we found that of 37 highly active L1 elements, 17 contained transductions. Therefore, we sought to identify the polymorphic L1s from known active transduction subfamilies. A transduction-specific modification of the suppression PCR-based amplification typing of L1 active subfamilies technique or TS-ATLAS was recently developed to specifically locate L1s with a given transduction sequence (Richard Badge Laboratory, manuscript in preparation) (Badge et al., 2003). With a 3' transduction-specific primer, elements were isolated from a subfamily that includes a full-length mutagenic insertion into the *retinitis pigmentosa-2* gene (the $L1_{RP}$ mutagenic insertion contains an 11 bp 3' transduction of 5'- A*n*GTTTTAAATTTA*n* -3'). Therefore, these elements belong to

the RP subfamily of L1s (Kimberland et al., 1999). RP-specific TS-ATLAS identified two full-length and two 5' truncated L1s in 9 genomic DNAs from unrelated individuals. One of the full-length RP elements was flanked by target site duplications that included the transduction sequence and lacked a second poly(A) tail 3' of the downstream TSD, indicating that it is the putative progenitor L1 for the RP subfamily.

The putative RP progenitor L1 contained two intact ORFs, and therefore may be competent for retrotransposition. The L1, termed AL050308 for the accession number of its insertion location on the X chromosome, was PCR amplified from the genomic DNA of the blood donor with primers adapted to the 5' and 3' flanking regions. The element was then digested and cloned into a vector that allowed the L1 to be tested in a retrotransposition assay. Sequencing the L1 in the vector verified that it was the same as the genomic L1, with no nucleotide changes. The L1 RP progenitor was then tested in the cultured cell retrotransposition assay, and was found to retrotranspose at ~170% of L1.3 and ~150% of the mutagenic $L1_{RP}$ insertion itself (Figure 3.6 B). These two RP family members contain 11 sequence changes from one another, and no amino acid differences. The isolation of an active progenitor element provides proof-of-principle that transduction-specific PCR-based techniques may indeed be a novel method to identify highly active polymorphic L1s.

Discussion

Recent studies have focused on the prevalence of L1 polymorphisms (Beck et al., 2011), or the affect that mutations in conserved residues have on

retrotransposition (Doucet et al., 2010). Here, we used a recently developed data set of 68 elements, 53 of which have been tested for retrotransposition efficiency and contained intact ORFs (Beck et al., 2010), to explore the cohort of sequence differences present in genomic L1s.

Examination of the nucleotide sequence of #6-113 allowed the discovery of a prevalent alternative splice variant of the 5'UTR. The 5'UTR variant was present in 96 genomic L1 elements, 30 of which were found in the introns of known RefSeq genes, and 21 of the 30 were in the opposite transcriptional orientation. This 2:1 preference for the antisense orientation is similar to other cohorts of L1 elements (Beck et al., 2010; Ovchinnikov et al., 2001; Smit et al., 1995), and indicates that intronic insertions in the sense orientation to a gene may be more deleterious. Full-length L1s also appear to be more deleterious to truncated elements, as they tend to be depleted from the genome over millions of years (Boissinot et al., 2001; Boissinot et al., 2004). The deletion of 524 bp in the central region of the 5'UTR affects the RUNX3 binding site that overlaps the splice donor, the SRY binding sites at bp 472 and 572, and also deletes the critical region for the L1 antisense promoter (Athanikar et al., 2004; Speek, 2001; Swergold, 1990; Tchenio et al., 2000; Yang et al., 2003). Although upstream promoters may transcribe some intronic 5'UTR deletion-containing L1s, many of the sequences in Table 3.2 are likely to have been generated by a splicing event in the 5'UTR of an intact L1. The prevalence of L1PA3 elements with this junction suggests that some L1 families may contain better contexts for a given splice donor or acceptor. This splice variant shows that some L1 mRNA transcripts can

still readily retrotranspose after the spicing-mediated removal of a large portion of the 5'UTR.

Analysis of the 16 L1s with intact ORFs and low-level or no retrotransposition activity allowed discovery of potentially deleterious mutations. Although the splicing mutation in #6-113 and the EN mutation of #3-24 are the likely cause of their inactivity in HeLa cells, amino acid changes are potentially responsible for the inactivity of 14 additional elements. Swapping ORF1p and ORF2p from inactive L1s with the corresponding proteins of L1.3 would be valuable for determining which changes might be deleterious to retrotransposition. Other functional assays, including the examination of protein localization to RNPs and the stability of ORF1p and ORF2p or an *in vitro* assessment of L1 reverse transcriptase activity (L1 element amplification protocol, or LEAP) could also aid in the determination of a cause for L1 inactivity in HeLa cells (Doucet et al., 2010; Kulpa and Moran, 2005, 2006). Finally, to pinpoint the mutation in the 14 L1s responsible for their low-level activities, introduction of putatively causative amino acid changes into a known, active element (*i.e.* L1.3) will be required.

Interestingly, 5 of the L1s that retrotranspose with high efficiency in NHEJ deficient CHO cells lack mutations in the canonical EN domain, which consists of the first 240 bp of ORF2p (Figure 3.5 and Table 3.3- elements #2-25, 3-2-1, 4-17, 4-19, and 5-36) (Feng et al., 1996; Weichenrieder et al., 2004). While mutations directly affecting EN domain function may occur past amino acid 240, other differences within ORF2p could potentially affect the function of the

endonuclease domain indirectly. Alternately, the 5 elements may be deficient in additional domains of ORF2p or ORF1p that could render the L1 active in an ENi assay, but inactive in HeLa cells (Kopera et al., 2011). Although further study will be required to pinpoint the mutations responsible for the inactivity of many of the L1s displayed in Table 3.3, this study has provided an important framework for future analysis of both these L1s and new datasets.

In summary, we have conducted a thorough investigation of the coding regions of 53 intact, polymorphic L1Hs elements. These studies have informed us about the nature of the mutations present in human L1 sequences. Moreover, we have identified numerous amino acid residues that may be important for L1 activity. This study was the first to analyze an extensive data set of intact and inactive L1 elements, and expansion of this analysis to new L1s identified via next generation sequencing and the 1000 genomes project is now possible (Ewing and Kazazian, 2011, 2010; Huang et al., 2010; Iskow et al., 2010; Mills et al., 2011; Xing et al., 2009). The ability to identify defects in L1s unable to retrotranspose in HeLa cells has now been shown for two L1s, #3-24 and #6-113. Element #3-24 was previously determined to contain a mutation in the EN domain. Element #6-113 was shown here to have a deletion of the 5'UTR, rendering the L1 unable to retrotranspose without an exogenous promoter. Although some of the L1s in this study and in further investigations of intact and inactive elements may contain numerous domains affected by amino acid differences (*e.g.,* #3-36 or 4-20), many may have simple causes for their inactivity in HeLa cells and will inform us about novel aspects of L1 biology (*e.g.,*

#6-113). Therefore, the mutational spectrum gleaned from these inactive elements will likely lead the L1 field in novel directions.

Experimental Procedures

L1Hs Sequencing

Full-length L1Hs elements were isolated, sequenced, and tested with respect to retrotransposition activity as previously described (Beck et al., 2010; Kidd et al., 2008; Tuzun et al., 2005). Briefly, intermediate-sized insertions (~4-8 kb) were identified in the genomes of 6 individuals using a fosmid paired-end sequencing approach (Kidd et al., 2008; Tuzun et al., 2005). The insertion-containing fosmids were then screened to find clones with full-length L1s that were absent from the HGR (Beck et al., 2010). L1-containing fosmids were then sequenced at the Washington University Genome Sequencing Center (St. Louis) and/or the L1 itself was sequenced at the University of Michigan Sequencing Core Facility using an Applied Biosystems ABI Model 3730XL (Beck et al., 2010; Kidd et al., 2010).

Additional L1 Sequence Resources

L1PA2, L1PA3b, L1PA4, and L1PA5 sequences (comprising the last ~25 million years of L1 evolution in the human genome) were obtained from fasta format alignments via FTP from ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000165.dat (Boissinot and Furano, 2001). Previously described consensus sequences (Khan et al., 2006)

(Boissinot and Furano, 2001) were generated using full-length elements of a given primate L1 family from the HGR (Lander et al., 2001).

Many of the sequences for the alignments of ORF1p and ORF2p from evolutionary diverse species were obtained from RepBase (http://www.girinst.org/repbase/) and then deposited into a MEGA 4 alignment file. These include consensus nucleic acid sequences for L1 Rn (*Rattus norvegicus*), Ss (*Sus scrofa*), and Md (*Monodelphis domestica*), and elements that are similar to consensus sequences and flanked by TSDs from Mm (*Mus musculus*), Bt (*Bos taurus*), Cf (*Canis lupus familiaris*), Tx1 (*Xenopus laevis*), Sw1 (*Oryzias latipes*), Dr (*Danio rerio*), CIN4 (*Zea Mays*) DRE (*Dictyostelium discoideum*), Tal1 (*Arabidopsis thaliana*), Zepp (*Chlorella vulgaris*), and Zorro (*Candida albicans*). Other sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/guide/), including the polymorphic *Homo sapiens* L1.3 nucleic acid sequence (accession number L19088) and the *Nycticebus coucang* (slow loris) ORF2p amino acid sequence (accession number P08548). Amino acid sequences were generated from nucleotide files by translating the sequences using the lasergene software package from DNASTAR (http://www.dnastar.com/).

Sequence Alignments

Alignments were created using MEGA4 (Tamura et al., 2007). CLUSTALW (Thompson et al., 1994) was used to find the best alignment of the 53 L1s with intact ORFs and both consensus sequences for L1PA2-5 and a 'hot' or highly active L1 consensus (Beck et al., 2010; Boissinot and Furano, 2001). Regions of

the aligned elements (UTRs, ORFs, and the 63 bp intergenic spacer- Figure 3.1) were annotated to allow subsequent ORF1p and ORF2p amino acid alignments to be generated from the data file in MEGA4. ORF1p and ORF2p alignments with the 53 L1Hs elements and the 5 consensus sequences are shown in Figure 3.3. CLUSTALW was also used to generate alignments of evolutionarily diverse ORF1p and ORF2p amino acid sequences (Figure 3.4). The ORF2p evolutionary alignment was then hand curated according to known important residues in Zorro (Dong et al., 2009).

Plasmid Constructs

L1Hs-containing plasmids in this study are derived from those published previously (Beck et al., 2010). L1s in JCC9, a pBluescript-based (Stratagene) plasmid, were subjected to restriction digest with NotI and BstZ17i (New England Biolabs). Resultant ~6 kb L1 fragments were then ligated into either JM105 L1.3 or JJ105 L1.3 backbones lacking the same NotI to BstZ17i fragment of an RT defective L1.3 (Accession number L19088 (Dombroski et al., 1993)) containing a D702A mutation in ORF2p (Wei et al., 2001). JM105 L1.3 consists of a pCEP4 (Invitrogen) backbone containing an exogenous CMV promoter driving transcription of a full-length L1 with a neomycin retrotransposition indicator cassette tagging the 3′UTR of the element (Freeman et al., 1994; Moran et al., 1996). JJ105 L1.3 is similar to JM105L1.3, but contains a blasticidin indicator cassette in the 3′UTR of the L1 (Morrish et al., 2002). The genomic L1s present in the vectors are as sequenced except for the first ~35bp of the 5′UTR, and the last ~50 bp of the 3′UTR, which are from L1.3.

JM101 and 105 L1.3 ACA are L1.3 constructs described above that contain either wild type L1.3 or L1.3 with a D702A mutation in ORF2p, respectively. To construct JM101 L1.3 GAG, we created a primer extending from 7 bp downstream of the BstZ17i site through the desired GAG mutation, and continuing 5' of the original ACA nucleotides by 10 bp (spanning bp 5919-5975 of L1.3). PCR with this oligonucleotide and a 5' primer that spans bp 5287-5313 of L1.3 (ORF2K) was used to amplify a 688 bp fragment of linearized JCC5 L1.3 (pBluescript-based plasmid with full-length L1.3 sequence) that contained the GAG mutation when sequenced. Digestion of the fragment with SpeI and BstZ17i allowed cloning of a 532 bp GAG-containing fraction of L1.3 into JCC5 L1.3 ACA, and sequencing confirmed that no additional nucleotide changes were introduced in the cloning process. The L1.3 GAG allele was then cloned into JM105 L1.3 ACA using NotI and BstZ17i as described above.

Cell Culture

HeLa and Chinese Hamster Ovary (CHO) cells and maintenance were previously described (Beck et al., 2010; Moran et al., 1996; Morrish et al., 2002; Wei et al., 2000). Briefly, cells were maintained in an incubator at 37°F with 100% humidity and 7% CO_2. HeLa cells were maintained in DMEM high glucose media with 10% FBS, 20 U/ml penicillin/streptomycin, and 0.4 mM glutamine (Gibco). CHO cells are auxotrophic for proline and glycine, and were maintained in low-glucose DMEM with 10% FBS, 20 U/ml penicillin/streptomycin, 0.4 mM glutamine, and non-essential amino acids (Gibco).

Retrotransposition Assays

Retrotransposition assays were conducted as previously described (Moran et al., 1996; Morrish et al., 2002; Wei et al., 2000). Briefly, 3 wells (per construct to be tested) of a 6-well tissue culture plate were seeded with $2x10^3$, $2x10^4$, and $2x10^5$ HeLa cells the day before transfection. Additionally, 3 wells per construct were seeded with $2x10^4$ and $2x10^5$ cells to test for transfection efficiency. Each experimental well is transfected with 1 µg of plasmid (L1 constructs with retrotransposition indicator cassettes are prepared with the Qiagen tip-100 midi prep kit and analyzed on agarose gel) using Fugene 6 transfection reagent (Roche) and Opti-mem media (Gibco). Efficiency is obtained for each construct by transfection with 0.5 µg of the plasmid of interest, and an additional 0.5 µg of pCEP green fluorescent protein (GFP) per well (Alisch et al., 2006). Twenty-four hours post transfection, cell culture media is aspirated and replenished. Seventy-two hours post transfection, cell culture media is aspirated and replaced with the original media plus 400 µg/ml G418 (for JM backbone plasmids). Thirteen-fourteen days post transfection, cells are fixed with a solution of 2% formaldehyde/0.2% glutaraldehyde in PBS and stained with 0.1% crystal violet. For Blasticidin-resistant constructs (JJ backbone), cells are fed with fresh media at 72 hours, and at 120 hours media is aspirated and replaced with media plus 10 µg/ml Blasticidin. Twelve-thirteen days post transfection cells are fixed and stained as above. GFP-containing cells are harvested at 72 hours and subjected to FACS analysis for % GFP-positive cells, which is used as an estimate for the efficiency of the accompanying retrotransposition indicator cassette-containing

construct. CHO cells are seeded with $1x10^3$, $1x10^4$, and $1x10^5$ cells/well, transfected 6-8 hours later, and assays are then performed as above.

Putative $L1_{RP}$ Progenitor Cloning and Retrotransposition Assays

TS-ATLAS is a modification of Amplification Typing of L1 Active Subfamilies, or ATLAS (Badge et al., 2003). The protocol is similar to the previously published method, with the replacement of the 3' primer for one that specifically anneals to a transduction sequence of interest (RP, in this case). RP-specific TS-ATLAS was used to isolate the putative progenitor of the lineage that includes $L1_{RP}$ (Kimberland et al., 1999) (insertion AL050308). Sequencing the genomic DNA flanking the L1 identified the insertion location and allowed the generation of primers. The L1 was then amplified from 50ng genomic DNA of a volunteer sperm donor by long-range PCR using flanking genomic primers JM0308D (5'- TTTGGATTAAAAAGTTTTAAATTGGG- 3', which includes the 5 Gs at the 5' end of the L1) and CM0308A (5'- GACTCTTTCAGTTGCCAGATGC - 3', to the 3' flanking region of the insertion site in the HGR). Direct sequencing of the PCR products confirmed this allele of AL050308 had intact open reading frames and was likely to be retrotransposition competent. To facilitate successful amplification of relatively error-free PCR products, long range PCR using the Expand Long Range polymerase system (Roche) was employed using 50ng of template DNA and buffer 2. PCR products were then cloned by digesting PCR fragments with *Acc*I, performing PCR clean up with the Zymoclean DNA gel recovery kit from Zymo Research, and ligation of fragments into appropriate restriction sites in the pBluescript-based vector JCC9 containing the *mneoI*

retrotransposition indicator cassette. Cloned $L1_{RP}$ progenitor elements were sequenced in their entirety to identify clones with minimal sequence variation from the directly determined consensus. The RP progenitor clone that was tested for activity in Figure 3.6B contained no nucleotide changes from the genomic DNA PCR sequence of the L1. Retrotransposition assays comparing L1.3 to the RP progenitor and $L1_{RP}$ mutagenic insertion were conducted as discussed above, however cell concentrations of $5x10^3$, $2x10^4$, and $2x10^5$ are used.

Acknowledgements

I would like to thank all members of the Moran lab for helpful discussions that contributed to this Chapter. Notably, John Moran identified the 5′UTR splicing, Aurelien Doucet previewed evolutionary alignments, Peter Larson and Aurelien Doucet shared unpublished data, and John Moldovan and Peter Larson were sounding boards for multiple conversations involved in this Chapter. The laboratory of Richard Badge conducted the TS-ATLAS experiments. I cloned, tested, and analyzed the RP progenitor element, with the help of Amanda Day, Swathi Yadlapalli, and Diane Flasch. The GAG allele of L1.3 was cloned with the help of Ashley Tan.

Figure 3.1: The Coding Potential of 68 Full-Length L1s

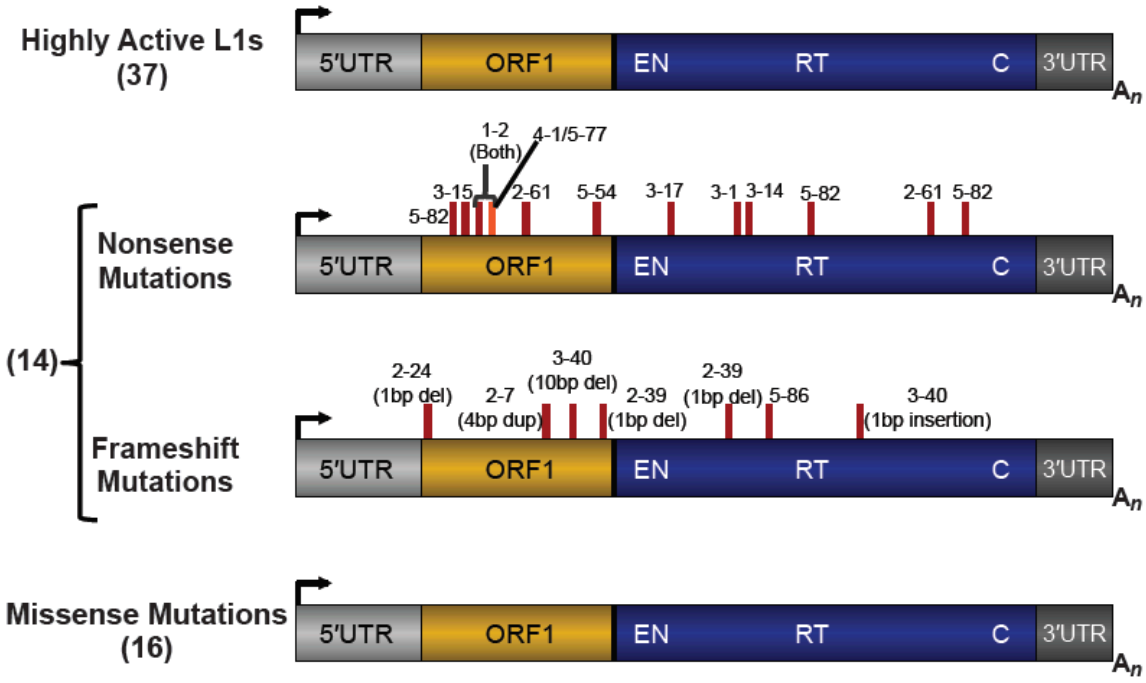


Figure 3.1: The Coding Potential of 68 Full-Length L1s

Full-length L1s are ~6 kb in length, encode their own promoter within the 5′UTR (grey rectangle), end in a 3′UTR (dark grey rectangle), and are punctuated by a poly(A) tail (A*n*). Retrotransposition-competent elements also encode two protein-coding ORFs separated by a short (63 bp) intergenic region: ORF1 (yellow rectangle) and ORF2 (blue rectangle). ORF2 contains endonuclease (EN), reverse transcriptase (RT), and cysteine-rich (C) domains. Thirty-seven of the L1s in this study were previously found to retrotranspose at >10% of a reference element (L1.3) (Beck et al., 2010). The remaining 30 L1s had low-level retrotransposition efficiency in the cell culture assay: 14 elements contain nonsense or frameshift mutations resulting in the premature termination of one or both ORFs, and 16 elements potentially encode amino acid differences. Red lines indicate the approximate location of the differences in ORF1 and ORF2, and are labeled with the previously described element name (Beck et al., 2010). The orange line indicates a CpG to TpG change present in two separate L1 elements changing a CGA (R residue) to a TGA (stop codon).

Figure 3.2: Splicing Within the 5′UTR of Element #6-113

(A) Element #6-113 contains a 524 bp deletion in the 5′UTR that is consistent with a splicing event. **(B)** The putative splicing event in #6-113 is a common splice junction in many genomic L1s. Thirty of the 96 L1s from the human genome reference sequence that contain the same splice as #6-113 are depicted with the original element. The red arrow indicates the splice junction for each of the 31 L1s. **C)** The deletion in the 5′UTR of #6-113 causes a loss of transcriptional activity that can be partially rescued by cloning an exogenous CMV promoter upstream of the L1. Two independent clones of #6-113 with a CMV promoter were tested. CMV #6-113 retrotransposes at ~20% of CMV L1.3 (JM101 L1.3), and an RT- negative control is shown (JM105 L1.3- D702A mutation in ORF2p) (percent retrotransposition efficiency relative to L1.3 is shown at the bottom of the panel).

Figure 3.2: Splicing Within the 5′UTR of Element #6-113

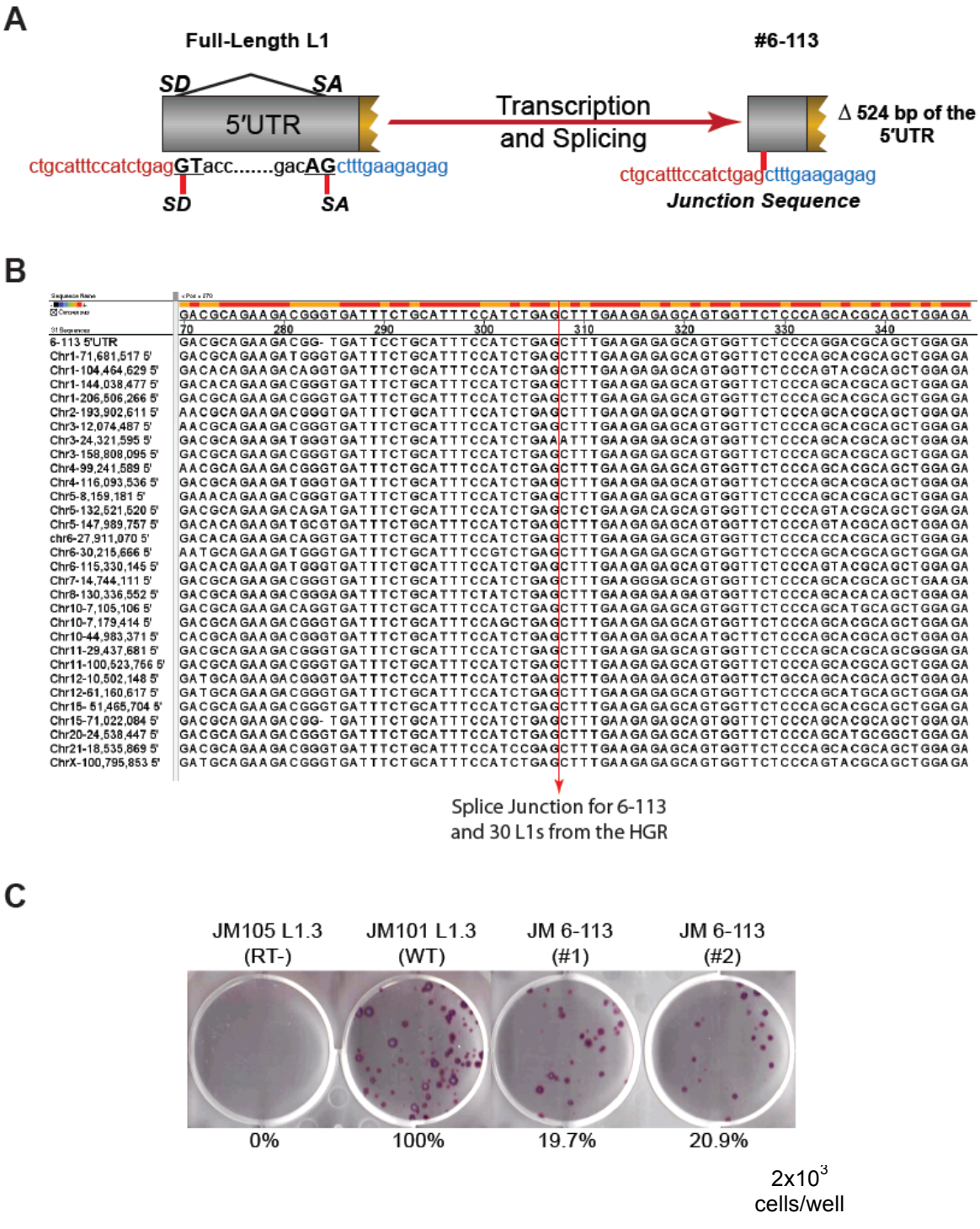
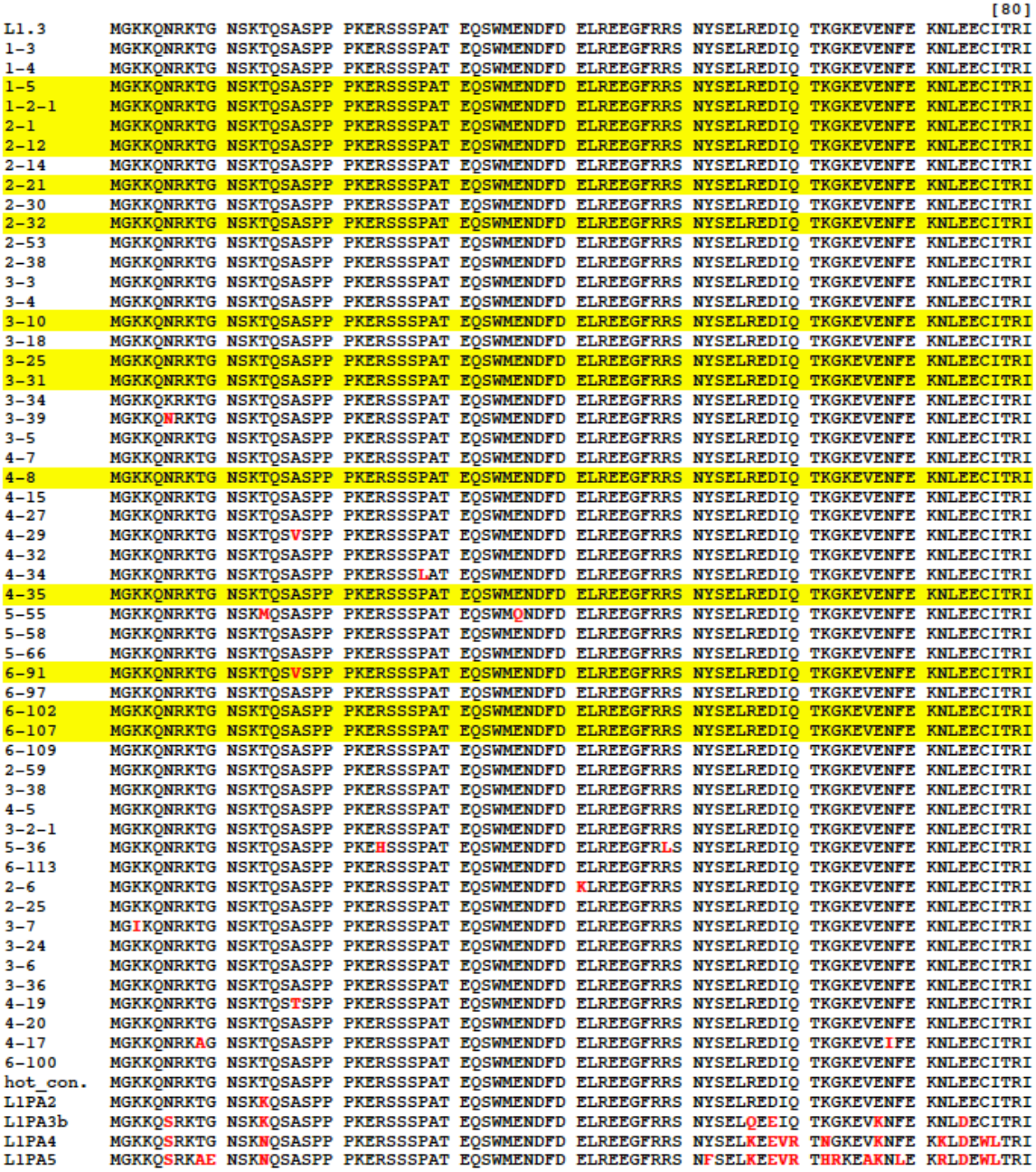


Figure 3.3: Amino Acid Alignments of ORF1p and ORF2p

(A) The figure depicts an alignment of ORF1p from 53 L1s as well as consensus sequences of older L1s derived from the human genome and L1.3 (Beck et al., 2010; Boissinot and Furano, 2001; Dombroski et al., 1993; Khan et al., 2006). Amino acid residues in bold red differ from a consensus amino acid sequence. Yellow highlighting denotes the L1s that retrotranspose >115% of L1.3. **(B)** An alignment of ORF2p from the 59 elements with the same demarcations for sequence differences and highly active elements as in ORF1p.

Figure 3.3A: Amino Acid Alignment of ORF1p



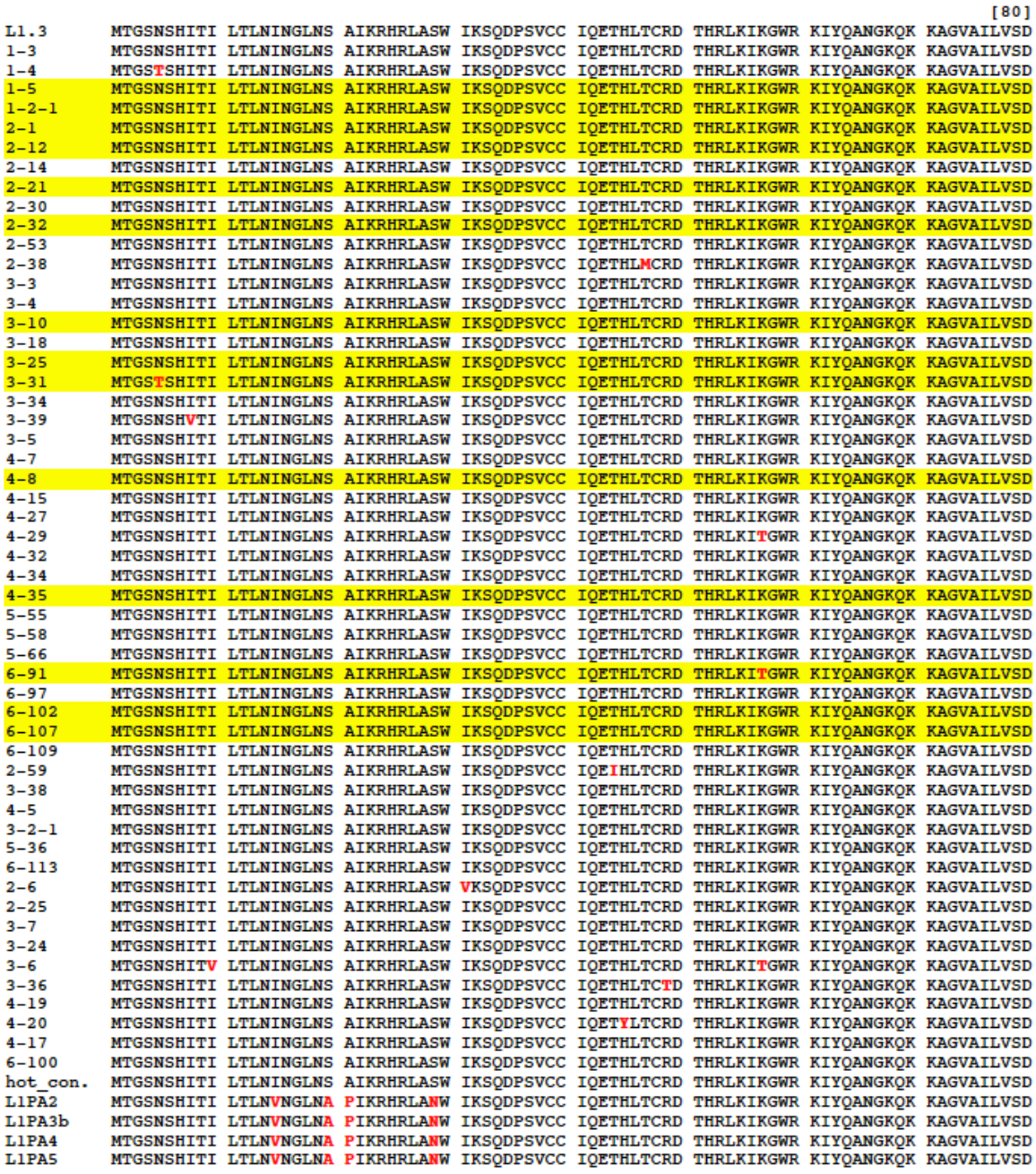
```
                                                                                                                                [160]
L1.3        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
1-3         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
1-4         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
1-5         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
1-2-1       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-1         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-12        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-14        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-21        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-30        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-32        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-53        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-38        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-3         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-4         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-10        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-18        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-25        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-31        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-34        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-39        TNTEKCLKEL MELKTKAREL REECRSLRSR CNQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-5         TNTENCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-7         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-8         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-15        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-27        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKQ EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-29        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-32        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-34        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-35        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
5-55        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
5-58        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
5-66        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-91        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-97        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-102       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-107       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-109       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-59        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-38        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-5         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-2-1       TNTEKCLKEL MELKIKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
5-36        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-113       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQNLQEIW DYVKRPNLRL
2-6         TNTEKCLKEL MELKTKAREL REECRSLRSR CNQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
2-25        TNTEKCLKEL MELKTKAREL REECRSLRSR YNQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-7         TNTEKCLKEL IELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-24        TNTENCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-6         TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
3-36        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-19        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-20        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
4-17        TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
6-100       TNTEKCLKER MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
hot_con.    TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
L1PA2       TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSA MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
L1PA3b      TNTEKCLKEL MELKTKAREL REECRSLRSR CDQLEERVSV MEDEMNEMKR EGKFREKRIK RNEQSLQEIW DYVKRPNLRL
L1PA4       TNAEKSLKDL MELKTKAREL RDECTSLSSR CDQLEERVSV MEDEMNEMKR EEKFREKRIK RNEQSLQEIW DYVKRPNLRL
L1PA5       TSVEKSLNDL MELKTMAREL RDECTSFSSR FDQLEERVSV IEDQMNEMKR EEKFREKRIK RNEQSLQEIW DYVKRPNLRL
```

								[240]
L1.3	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
1-3	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
1-4	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
1-5	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
1-2-1	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-1	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-12	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-14	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-21	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-30	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-32	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-53	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRATREKGRV
2-38	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	KQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-3	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGWV
3-4	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-10	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-18	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-25	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-31	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-34	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-39	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-5	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRFSSR	RATPRHIIVR	FTKVEMKEKI	LRAAREKGRV
4-7	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-8	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-15	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-27	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-29	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-32	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-34	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHILVR	FTKVEMKEKM	LRAAREKGQV
4-35	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
5-55	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRFSSR	RATPRHIIVR	FTKVEMKEKI	LRAAREKGRV
5-58	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
5-66	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRKPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-91	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-97	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-102	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-107	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-109	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-59	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-38	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-5	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQKI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-2-1	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
5-36	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQTQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-113	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKLEMKEKM	LRAAREKGRV
2-6	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
2-25	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-7	IGVPESDVEN	GTKLENTLQD	NIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-24	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRFSSR	RATPRHIIVR	FTKVEMKEKI	LRAAREKGRV
3-6	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
3-36	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FNKVEMKEKM	LRAAREKGRV
4-19	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-20	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
4-17	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
6-100	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
hot_con.	IGVPESDVEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
L1PA2	IGVPESDGEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
L1PA3b	IGVPESDGEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
L1PA4	IGVPESDGEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV
L1PA5	IGVPESDGEN	GTKLENTLQD	IIQENFPNLA	RQANVQIQEI	QRTPQRYSSR	RATPRHIIVR	FTKVEMKEKM	LRAAREKGRV

```
                                                                                                      [320]
L1.3      TLKGKPIRLT VDLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
1-3       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
1-4       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
1-5       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
1-2-1     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-1       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-12      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-14      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-21      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-30      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-32      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-53      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-38      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SFPAKLSFIS EGERKYFTDK QMLRDFVTTR PTLKELLKEA
3-3       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-4       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-10      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-18      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-25      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-31      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-34      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-39      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-5       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLNFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEV
4-7       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGERKYFTDK QMLRDFVTTR PTLKELLKEA
4-8       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-15      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-27      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-29      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEV
4-32      TLKGKPIRLT ADLSAETLQA RREWGPIFNN LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-34      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKES
4-35      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
5-55      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEV
5-58      TLKGKPIRLT VDLSAETLQA RREWGPIFNI LKENNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
5-66      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-91      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEV
6-97      TLKGKPIRLT VDLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-102     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-107     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-109     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-59      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-38      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-5       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-2-1     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
5-36      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SCPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-113     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-6       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
2-25      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFHPRI SFPAKLSFIS EGERKYFTDK QMLRDFVTTR PTLKELLKEA
3-7       TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQTRI SYPAKLSFIN EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-24      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEV
3-6       TLKGKPIRLT ADLLAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
3-36      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-19      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
4-20      TLKGKPIRLT ADLSAETLQA RREWGPILNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVATR PALKELLKEA
4-17      TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
6-100     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
hot_con.  TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFIDK QMLRDFVTTR PALKELLKEA
L1PA2     TLKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFTDK QMLRDFVTTR PALKELLKEA
L1PA3b    THKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFTDK QMLRDFVTTR PALKELLKEA
L1PA4     THKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKYFTDK QMLRDFVTTR PALKELLKEA
L1PA5     THKGKPIRLT ADLSAETLQA RREWGPIFNI LKEKNFQPRI SYPAKLSFIS EGEIKSFTDK QMLRDFVTTR PALQELLKEA
```

```
                                [338]
L1.3        LNMERNNRYQ PLQNHAKM*
1-3         LNMERNNRYQ PLQNHAKM*
1-4         LNMERNNRYQ PLQNHAKM*
1-5         LNMERNNRYQ PLQNHAKM*
1-2-1       LNMERNNRYQ PLQNHAKM*
2-1         LNMERNNRYQ PLQNHAKM*
2-12        LNMERNNRYQ PLQNHAKM*
2-14        LNMERNNRYQ PLQNHAKM*
2-21        LNMERNNRYQ PLQNHAKM*
2-30        LNMERNNRYQ PLQNHAKM*
2-32        LNMERNNRYQ PLQNHAKM*
2-53        LNMERNNRYQ PLQNHAKM*
2-38        LNMERNNRYQ PLQNHAKM*
3-3         LNMERNNRYQ PLQNHAKM*
3-4         LNMERNNRYQ PLQNHAKM*
3-10        LNMERNNRYQ PLQNHAKM*
3-18        LNMERNNRYQ PLQNHAKM*
3-25        LNMERNNRYQ PLQNHAKM*
3-31        LNMERNNRYQ PLQNHAKM*
3-34        LNMERNNRYQ PLQNHAKM*
3-39        LNMERNNRYQ PLQNHAKM*
3-5         LNMERNNRYQ PLQNHAKM*
4-7         LNMERNNRYQ PLQNHAKM*
4-8         LNMERNNRYQ PLQNHAKM*
4-15        LNMERNNRYQ PLQNHAKM*
4-27        LNMERNNRYQ PLQNHAKM*
4-29        LNMERNNRYQ PLQNHAKM*
4-32        LNMERNNRYQ PLQNHAKM*
4-34        LNMERNNRYQ PLQNHAKM*
4-35        LNMERNNRYQ PLQNHAKM*
5-55        LNMERNNRYQ PLQNHAKM*
5-58        LNMERNNRYQ PLQNHAKM*
5-66        LNMERNNRYQ PLQNHAKM*
6-91        LNMERNNRYQ PLQNHAKM*
6-97        LNMERNNRYQ PLQNHAKT*
6-102       LNMERNNQYQ PLQNHAKM*
6-107       LNMERNNRYQ PLQNHAKM*
6-109       LNMERNNRYQ PLQNHAKM*
2-59        LNMERNNRYQ PLQNHAKM*
3-38        LNMERNNRYQ PLQNHAKM*
4-5         LNMERNNRYQ PLQNHAKM*
3-2-1       LNMERNKRYQ PLQNHAKM*
5-36        LNMESNNRYQ PLKNHAKM*
6-113       LNMERNNRYQ PLQNHAKM*
2-6         LNMERNNRYQ PLQNHAKM*
2-25        LNMERNNRYQ PLQNHAKM*
3-7         LNMERNNRYQ PLQNHAKM*
3-24        LNMERNNRYQ PLQNHAKM*
3-6         LNMERNNWYQ PLQNHAKM*
3-36        LNMERNNRYQ PLQNHAKM*
4-19        LNMERNNRYQ PLQNHAKM*
4-20        LNMERNNRYQ PLQNHAKM*
4-17        LNMERNNRYQ PLQNHAKM*
6-100       LNMERNNRYQ PLQNHAKM*
hot_con.    LNMERNNRYQ PLQNHAKM*
L1PA2       LNMERNNRYQ PLQNHAKM*
L1PA3b      LNMERNNRYQ PLQNHAKL*
L1PA4       LNMERNNRYQ PLQKHAKL*
L1PA5       LNMERNNRYQ PLQKHAKL*
```

Figure 3.3B: Amino Acid Alignment of ORF2p



								[160]
L1.3	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
1-3	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
1-4	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
1-5	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
1-2-1	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-1	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-12	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-14	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-21	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	HLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-30	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-32	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-53	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	HLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-38	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	SILDRSTRQK
3-3	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-4	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-10	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-18	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-25	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-31	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-34	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-39	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-5	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSMRQK
4-7	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	SILDRSTRQK
4-8	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-15	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-27	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-29	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-32	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-34	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	ELQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-35	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
5-55	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSMRQK
5-58	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
5-66	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-91	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-97	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-102	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-107	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-109	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-59	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-38	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-5	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-2-1	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
5-36	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-113	KTDFKPTKIK	RDKEGHHIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
2-6	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSMRQK
2-25	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	SILDRSTRQK
3-7	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-24	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSMRQK
3-6	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
3-36	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQKDLDSHT	LIMGDFSTPL	STLDRSTRQK
4-19	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-20	KTDFKPTKIK	RDKEGHYLMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
4-17	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
6-100	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	HLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
hot_con.	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
L1PA2	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	LIMGDFNTPL	STLDRSTRQK
L1PA3b	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	IIMGDFNTPL	STLDRSTRQK
L1PA4	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLS	DLQRDLDSHT	IIMGDFNTPL	STLDRSTRQK
L1PA5	KTDFKPTKIK	RDKEGHYIMV	KGSIQQEELT	ILNIYAPNTG	APRFIKQVLR	DLQRDLDSHT	IIMGDFNTPL	STLDRSTRQK

								[240]
L1.3	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
1-3	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
1-4	VNKDTQELNS	ALHQADLIDI	YRTLQPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
1-5	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
1-2-1	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-1	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-12	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-14	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-21	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-30	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-32	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-53	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-38	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-3	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-4	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-10	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-18	VNKDTQEMNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-25	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-31	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-34	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-39	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-5	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-7	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-8	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-15	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-27	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-29	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLEIRIK
4-32	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-34	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-35	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
5-55	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
5-58	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
5-66	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-91	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLEIRIK
6-97	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-102	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-107	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-109	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-59	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIE
3-38	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YPFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-5	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSASHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	TAIKLELRIK
3-2-1	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
5-36	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-113	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-6	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
2-25	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-7	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHILGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
3-24	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLPDH	SAIKLELRIK
3-6	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIIRNYLSDH	SAIKLEIRIK
3-36	VNKDTQELNS	ALHQADLIDI	CRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRA	EIITNYLSDH	SAIKLELRIK
4-19	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-20	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
4-17	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
6-100	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
hot_con.	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
L1PA2	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHILGS	KALLSKCKRT	EIITNYLSDH	SAIKLELRIK
L1PA3b	VNKDTQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNCLSDH	SAIKLELRIK
L1PA4	VNKDIQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNCLSDH	SAIKLELRIK
L1PA5	VNKDIQELNS	ALHQADLIDI	YRTLHPKSTE	YTFFSAPHHT	YSKIDHIVGS	KALLSKCKRT	EIITNCLSDH	SAIKLELRIK

```
                                                                                                                [320]
L1.3        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
1-3         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
1-4         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
1-5         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
1-2-1       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-1         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-12        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-14        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA LKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-21        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-30        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-32        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-53        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-38        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-3         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-4         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-10        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-18        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-25        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-31        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-34        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-39        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-5         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-7         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-8         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-15        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA LKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-27        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDILTSQ
4-29        NLTQSHSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDSLTSQ
4-32        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-34        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-35        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
5-55        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
5-58        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
5-66        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
6-91        NLTQSHSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDSLTSQ
6-97        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
6-102       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYQRKQE RSKIDTLTSQ
6-107       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
6-109       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-59        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-38        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-5         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE GSKIDTLTSQ
3-2-1       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNESK DTTYQNLWDT FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
5-36        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
6-113       NLTQSCSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-6         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWNA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
2-25        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-7         NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQK RSKIDTLTSQ
3-24        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
3-6         NLTQSHSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDSLTSQ
3-36        NLTQSRSTTW KLNNLLPNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-19        NLTQSHSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENN DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-20        NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
4-17        NLTQSHSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAHKRKQE RSKIDTLTSQ
6-100       NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
hot_con.    NLTQSRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAYKRKQE RSKIDTLTSQ
L1PA2       NLTQNRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAHKRKQE RSKIDTLTSQ
L1PA3b      KLTQNRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAHKRKQE RSKIDTLTSQ
L1PA4       KLTQNRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAHKRKQE RSKIDTLTSQ
L1PA5       KLTQNRSTTW KLNNLLLNDY WVHNEMKAEI KMFFETNENK DTTYQNLWDA FKAVCRGKFI ALNAHKRKQE RSKIDTLTSQ
```

								[400]
L1.3	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
1-3	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
1-4	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKINRPLARL	IKKKREKNQI	DTIKNDKGDI
1-5	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
1-2-1	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-1	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-12	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-14	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-21	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-30	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-32	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-53	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-38	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-3	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-4	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-10	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKNLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-18	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-25	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-31	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-34	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-39	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKLN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-5	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-7	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-8	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-15	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-27	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-29	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-32	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-34	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-35	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
5-55	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKMN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKKDKGDI
5-58	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
5-66	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-91	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-97	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-102	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-107	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-109	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-59	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-38	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-5	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRLLARL	IKKKREKNQI	DTIKNDKGDI
3-2-1	LKELEKQEQT	HSKASRRQEV	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DRIKNDKGDI
5-36	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
6-113	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-6	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
2-25	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-7	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-24	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-6	LKDLEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
3-36	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-19	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
4-20	LKELEKQEQT	HSKASRRQEI	TKIRAEWKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDEGDI
4-17	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDGPLARL	IKKKREKNQI	DTIKNDKGDI
6-100	LKELEKQEQT	QSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
hot_con.	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
L1PA2	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DTIKNDKGDI
L1PA3b	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFERI	NKIDRPLARL	IKKKREKNQI	DAIKNDKGDI
L1PA4	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFEKI	NKIDRPLARL	IKKKREKNQI	DAIKNDKGDI
L1PA5	LKELEKQEQT	HSKASRRQEI	TKIRAELKEI	ETQKTLQKIN	ESRSWFFEKI	NKIDRPLARL	IKKKREKNQI	DAIKNDKGDI

```
                                                                                                [480]
L1.3      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
1-3       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
1-4       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
1-5       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
1-2-1     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-1       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-12      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-14      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-21      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-30      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGLDGFTAEF
2-32      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA VINSLPTKKS PGPDGFTAEF
2-53      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTVPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-38      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LKQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-3       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-4       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-10      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-18      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-25      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-31      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-34      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-39      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-5       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-7       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-8       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-15      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-27      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-29      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-32      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-34      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-35      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
5-55      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
5-58      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
5-66      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-91      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-97      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-102     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-107     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-109     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-59      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-38      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-5       TTDPTEIQTT IREYYKHLYA NKLENLEEMD KFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-2-1     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPT LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
5-36      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-113     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-6       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
2-25      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-7       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-24      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-6       TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
3-36      TTNPTEIQTT IREYYKHLYA NKLENLEEMD TFLNTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-19      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
4-20      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGLTAEF
4-17      TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
6-100     TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
hot_con.  TTDPTEIQTT IREYYKHLYA NKLENLEEMD TFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
L1PA5     TTDPTEIQTT IREYYKHLYA NKLENLEEMD KFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
L1PA3b    TTDPTEIQTT IREYYKHLYA NKLENLEEMD KFLDTYTLPR LNQEEVESLN RPITGSEIVA IINSLPTKKS PGPDGFTAEF
L1PA4     TTDPTEIQTT IREYYKHLYA NKLENLEEMD KFLDTYTLPR LNQEEVESLN RPITGSEIEA IINSLPTKKS PGPDGFTAEF
L1PA5     TTDPTEIQTT IREYYKHLYA NKLENLEEMD KFLDTYTLPR LNQEEVESLN RPITGSEIEA IINSLPTKKS PGPDGFTAEF
```

								[560]
L1.3	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
1-3	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
1-4	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
1-5	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
1-2-1	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-1	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-12	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-14	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-21	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-30	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-32	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-53	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-38	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKL	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-3	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILAKR	IQQHIKKLIH
3-4	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-10	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	DRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-18	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKTENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-25	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-31	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-34	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-39	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-5	YQRYKEELVP	FLLKLFQSIE	REGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-7	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-8	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-15	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-27	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASITLIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-29	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-32	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-34	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-35	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
5-55	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
5-58	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQLHIKKLIH
5-66	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-91	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-97	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-102	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-107	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-109	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTEKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-59	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-38	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-5	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-2-1	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
5-36	YQRYKEELVP	FLLKLFQTIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-113	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-6	YQRYMEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
2-25	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKL	GRDTAKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-7	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKL	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-24	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-6	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILTPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
3-36	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-19	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
4-20	YQRYKEELVP	FLLKRFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILSKILANR	IQQHIKKLIH
4-17	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
6-100	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
hot_con.	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
L1PA2	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
L1PA3b	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
L1PA4	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH
L1PA5	YQRYKEELVP	FLLKLFQSIE	KEGILPNSFY	EASIILIPKP	GRDTTKKENF	RPISLMNIDA	KILNKILANR	IQQHIKKLIH

```
                                                                                                  [640]
L1.3      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
1-3       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYLKII RAIYDKPTAN
1-4       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
1-5       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
1-2-1     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-1       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-12      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-14      HDQVGFIPGM QGWFNTRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-21      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
2-30      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-32      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-53      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
2-38      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-3       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-4       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTFFKII RAIYDKPTAN
3-10      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-18      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPRAN
3-25      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYLKII RAIYDKPTAN
3-31      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-34      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-39      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
3-5       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLIKL GIDGTYFKII RAIYDKPTAN
4-7       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-8       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-15      HDQVGFIPGM QGWFNTRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RGIYDKPTAN
4-27      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-29      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-32      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
4-34      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-35      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
5-55      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLIKL GIDGTYFKII RAIYDKPTAN
5-58      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
5-66      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-91      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-97      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-102     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-107     HDQVGFIPGM QGWFNIRKSI NVIQHVNRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTFFKII RAIYDKPTAN
6-109     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-59      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-38      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-5       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
3-2-1     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
5-36      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-113     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-6       HDQVGFIPGM QGWFNIRKSI NIIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
2-25      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYGKPTAN
3-7       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ RFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-24      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLIKL GIDGTYFKII RAIYDKPTAN
3-6       HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
3-36      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-19      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-20      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ RFMLKTLNKL GIDGTYFKII RAIYDKPTAN
4-17      HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ RFMLKTLNKL GIDGTYFKII RAIYDKPTAN
6-100     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGMYFKII RAIYDKPTAN
hot_con.  HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
L1PA2     HDQVGFIPGM QGWFNIRKSI NVIQHINRAK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYFKII RAIYDKPTAN
L1PA3b    HDQVGFIPGM QGWFNIRKSI NVIQHINRTK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYLKII RAIYDKPTAN
L1PA4     HDQVGFIPGM QGWFNIRKSI NVIQHINRTK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYLKII RAIYDKPTAN
L1PA5     HDQVGFIPGM QGWFNIRKSI NVIQHINRTK DKNHMIISID AEKAFDKIQQ PFMLKTLNKL GIDGTYLKII RAIYDKPTAN
```

```
                                                                                                                [720]
L1.3        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
1-3         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
1-4         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
1-5         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
1-2-1       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-1         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-12        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-14        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-21        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-30        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-32        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-53        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-38        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-3         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-4         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-10        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-18        IILNGQILEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-25        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIISAQNLLK
3-31        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-34        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-39        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-5         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-7         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PITSAQNLLK
4-8         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-15        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-27        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-29        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-32        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-34        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIISAQNLLK
4-35        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
5-55        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
5-58        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
5-66        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-91        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-97        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-102       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-107       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-109       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-59        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-38        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-5         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-2-1       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
5-36        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-113       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-6         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
2-25        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-7         IILNGQKLEV FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLTLF ADDMIVYLEN PIVSAQNLLK
3-24        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-6         IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
3-36        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAMR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-19        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-20        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
4-17        IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
6-100       IILNGQKLEA FPLKTGTRQG CPLSPLLFII VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
hot_con.    IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
L1PA2       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
L1PA3b      IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
L1PA4       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
L1PA5       IILNGQKLEA FPLKTGTRQG CPLSPLLFNI VLEVLARAIR QEKEIKGIQL GKEEVKLSLF ADDMIVYLEN PIVSAQNLLK
```

```
                                                                                           [800]
L1.3       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFV IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
1-3        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
1-4        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
1-5        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
1-2-1      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-1        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-12       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-14       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-21       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-30       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKRK
2-32       LISTFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-53       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-38       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
3-3        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-4        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-10       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-18       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-25       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-31       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-34       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-39       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-5        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-7        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-8        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-15       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-27       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-29       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-32       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQVMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-34       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-35       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
5-55       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
5-58       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFV IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
5-66       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
6-91       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
6-97       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFV IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
6-102      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
6-107      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
6-109      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-59       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
3-38       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-5        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASRRIKYLG IQLTRDVKDL FKENYKPLLN EIKEDTNKWK
3-2-1      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
5-36       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT TASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
6-113      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-6        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
2-25       LISNFSKVSG YKINVQKSQA FLYNNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
3-7        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKDNYKPLLK EIKEDTNKWK
3-24       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
3-6        LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
3-36       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
4-19       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-20       LISNFSEVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT ITSKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
4-17       LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTWDVKDL FKENYKPLLK EIKEDTNKWK
6-100      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
hot_con.   LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMGELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEETNKWK
L1PA2      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLK EIKEDTNKWK
L1PA3b     LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLN EIKEDTNKWK
L1PA4      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLN EIKEDTNKWK
L1PA5      LISNFSKVSG YKINVQKSQA FLYTNNRQTE SQIMSELPFT IASKRIKYLG IQLTRDVKDL FKENYKPLLN EIKEDTNKWK
```

```
                                                                                                                [880]
L1.3        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
1-3         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
1-4         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
1-5         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
1-2-1       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-1         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-12        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-14        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-21        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-30        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-32        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-53        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-38        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPYFKLYYK
3-3         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-4         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-10        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-18        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-25        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-31        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-34        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-39        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-5         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMPFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-7         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-8         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-15        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-27        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-29        NIPCSWVGRI NIVKMAILPK VIYRFNAIPV KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-32        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-34        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-35        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
5-55        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
5-58        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
5-66        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-91        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-97        NIPCSWVGRI NIVKMAILPK VIYTFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-102       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-107       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-109       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-59        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-38        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-5         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-2-1       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRACIAKSI LSQKNKAGGI TLPDFKLYYK
5-36        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
6-113       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-6         NIPCSWVGRI NIMKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
2-25        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-7         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-24        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-6         NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
3-36        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-19        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-20        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
4-17        NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFRLYYK
6-100       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
hot_con.    NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
L1PA2       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
L1PA3b      NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
L1PA4       NIPCSWVGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKSI LSQKNKAGGI TLPDFKLYYK
L1PA5       NIPCSWIGRI NIVKMAILPK VIYRFNAIPI KLPMTFFTEL EKTTLKFIWN QKRARIAKTI LSQKNKAGGI TLPDFKLYYK
```

```
                                                                                                                [960]
L1.3        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
1-3         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
1-4         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
1-5         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
1-2-1       ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-1         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-12        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-14        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-21        ATVTKTAWYW YKNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-30        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-32        ATVTKTAWYW YQNRDIDQWK RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-53        ATVTKTAWYW YKNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-38        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD LFLTPYTKIN
3-3         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-4         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-10        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-18        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWEIW LAICRKLKLD PFLTPYTKIN
3-25        ATVTKTAWYW YQNRDIDHWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-31        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-34        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPL IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-39        ATVTKTAWYW YKNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-5         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IHNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-7         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD LFLTPYTKIN
4-8         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-15        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-27        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-29        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-32        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-34        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-35        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
5-55        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IHNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
5-58        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
5-66        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-91        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-97        AIVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-102       ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-107       ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-109       ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-59        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-38        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-5         ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-2-1       ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
5-36        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
6-113       ATVTKTAWYW YQNRDVDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-6         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
2-25        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-7         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-24        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IHNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-6         ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
3-36        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-19        ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-20        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
4-17        ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWWWENW LALCRKLKLD PFLTPYTKIN
6-100       ATVTKTAWYW YKNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
hot_con.    ATVTKTAWYW YQNRDIDQWN RTEPSEIMPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
L1PA2       ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
L1PA3b      ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNYLIFDKP EKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
L1PA4       ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNYLIFDKP DKNKQWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
L1PA5       ATVTKTAWYW YQNRDIDQWN RTEPSEITPH IYNHLIFDKP DKNKKWGKDS LFNKWCWENW LAICRKLKLD PFLTPYTKIN
```

```
                                                                                                                [1040]
L1.3        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
1-3         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIQLKSFC TAKETTIRVN RQPTTWEKIF
1-4         SRWIKDLNIK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
1-5         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
1-2-1       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-1         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-12        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-14        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-21        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKHFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-30        SRWIKDLNVK PKTIKTLEEN VGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-32        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-53        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-38        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-3         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTEWEKIF
3-4         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
3-10        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-18        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-25        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-31        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-34        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-39        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-5         SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-7         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
4-8         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-15        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIQLKSFC TAKETTIRVN RQPTTWEKIF
4-27        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-29        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-32        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-34        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKTKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-35        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
5-55        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
5-58        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
5-66        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTEWEKIF
6-91        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
6-97        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
6-102       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
6-107       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
6-109       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-59        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTEWEKIF
3-38        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTEWEKIF
4-5         SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-2-1       SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
5-36        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
6-113       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
2-6         SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPIKWEKIF
2-25        SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-7         SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-24        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-6         SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
3-36        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-19        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKVKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-20        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
4-17        SRWIKDLNVR PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AIATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
6-100       SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
hot_con.    SRWIKDLNVK PKTIKTLEEN LGITIQDIGV GKDFMSKTPK AMATKDKIDK WDLIKLKSFC TAKETTIRVN RQPTTWEKIF
L1PA2       SRWIKDLNVR PKTIKTLEEN LGITIQDIGM GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
L1PA3b      SRWIKDLNVR PKTIKTLEEN LGITIQDIGM GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
L1PA4       SRWIKDLNVR PKTIKTLEEN LGNTIQDIGM GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTKWEKIF
L1PA5       SRWIKDLNVR PKTIKTLEEN LGNTIQDIGM GKDFMSKTPK AMATKAKIDK WDLIKLKSFC TAKETTIRVN RQPTEWEKIF
```

```
                                                                                                                [1120]
L1.3       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
1-3        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVKMA
1-4        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
1-5        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
1-2-1      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-1        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-12       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-14       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-21       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-30       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-32       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-53       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-38       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-3        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAVKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-4        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPFRMA
3-10       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-18       TTYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-25       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-31       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-34       ATYSYDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-39       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-5        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-7        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-8        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-15       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-27       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-29       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-32       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-34       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-35       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
5-55       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
5-58       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWVKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
5-66       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAVKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-91       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-97       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-102      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-107      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPFRMA
6-109      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-59       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAVKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-38       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAVKK HMKKCSSSLA IREMQIKTAM RYHLTPVRMA
4-5        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLV IREMQIKTTM RYHLTPVRMA
3-2-1      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHFTPVRMA
5-36       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-113      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
2-6        TTYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTV RYHLTPVRMA
2-25       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-7        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-24       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-6        ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
3-36       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-19       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-20       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
4-17       ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
6-100      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
hot_con.   ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
L1PA2      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
L1PA3b     ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
L1PA4      ATYSSDKGLI SRIYNELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAAKK HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
L1PA5      AIYSSDKGLI SRIYKELKQI YKKKTNNPIK KWAKDMNRHF SKEDIYAANR HMKKCSSSLA IREMQIKTTM RYHLTPVRMA
```

```
                                                                                                                [1200]
L1.3        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
1-3         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
1-4         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
1-5         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
1-2-1       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-1         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-12        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-14        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-21        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-30        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-32        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-53        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-38        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-3         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-4         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-10        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-18        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-25        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-31        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-34        IIKKSGNNRC WRGCREIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-39        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-5         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNDYKSCCYK DTCTRMFIAA
4-7         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
4-8         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
4-15        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
4-27        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
4-29        IIKKSGNNRC WRGCGEIGTL LHCWWDCQLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
4-32        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
4-34        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
4-35        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
5-55        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNDYKSCCYK DTCTRMFIAA
5-58        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
5-66        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
6-91        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
6-97        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
6-102       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
6-107       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
6-109       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-59        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
3-38        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
4-5         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLI QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-2-1       IIKKSGNYRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
5-36        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
6-113       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-6         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
2-25        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-7         IIKKSGNNRC WRGCGEIGTL SHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-24        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNDYKSCCYK DTCTRMFIAA
3-6         IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
3-36        IIKKSGNNRC WRGCGEIGKI LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTRTRMFIAA
4-19        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIVA
4-20        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYE DTCTRMFIAA
4-17        IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
6-100       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
hot_con.    IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNEYKSCCYK DTCTRMFIAA
L1PA2       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PNDYKSCCYK DTCTRMFIAA
L1PA3b      IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
L1PA4       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKSVWRF LRDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
L1PA5       IIKKSGNNRC WRGCGEIGTL LHCWWDCKLV QPLWKTVWRF LKDLELEIPF DPAIPLLGIY PKDYKSCCYK DTCTRMFIAA
```

[1275]

```
L1.3      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
1-3       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
1-4       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
1-5       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
1-2-1     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-1       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-12      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-14      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-21      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-30      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-32      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-53      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-38      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-3       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-4       LFTIAKTWNQ PKCPKMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LICGN*
3-10      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-18      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-25      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-31      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-34      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ENIILSKLSQ EQKTKHRIFS LIGGN*
3-39      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-5       LFTIAKTWNQ PKCPPMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHHIFS LIGGN*
4-7       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-8       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-15      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVATWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-27      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-29      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIIS LIGGN*
4-32      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-34      LFTIAKTWNQ PKCPTMIVWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHCIFS LIGGN*
4-35      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
5-55      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHHIFS LIGGN*
5-58      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
5-66      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-91      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIIS LIGGN*
6-97      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-102     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-107     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-109     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-59      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-38      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRVFS LIGGN*
4-5       LFTIAKTWNQ PKCPTMIDWI KKMWHIYSME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-2-1     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YHAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
5-36      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-113     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-6       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
2-25      LFTIPKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-7       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-24      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHHIFS LIGGN*
3-6       LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
3-36      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-19      LFTIAKTWNQ LKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-20      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
4-17      LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYATIKNDEF MSFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
6-100     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
hot_con.  LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF ISFVGTWMKL ETIILSKLSQ EQKTKHRIFS LIGGN*
L1PA2     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL EIIILSKLSQ EQKTKHRIFS LIGGN*
L1PA3b    LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL EIIILSKLSQ EQKTKHRIFS LIGGN*
L1PA4     LFTIAKTWNQ PKCPTMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ GQKTKHRMFS LIGGN*
L1PA5     LFTIAKTWNQ PKCPSMIDWI KKMWHIYTME YYAAIKNDEF MSFVGTWMKL ETIILSKLSQ GQKTKHRMFS LIGGN*
```

Figure 3.4: Conservation of ORF1p and ORF2p

(A) An alignment of 10 ORF1p amino acid sequences from human through zebrafish (*danio rerio*). Highlighted amino acids are those that have a strong to moderate affect on RNA binding, and residues in red indicate highly conserved amino acids (Khazina and Weichenrieder, 2009). **(B)** An alignment of 17 ORF2p amino acid sequences from human through Zorro (*candida albicans*). Residues in red are highly conserved, and the grey shading indicates the regions of the RT domain (Fanning and Singer, 1987; Moran and Gilbert, 2002; Weichenrieder et al., 2004). Species in both of the alignments and their corresponding abbreviations in the Figure are listed in the Experimental Procedures. Alignments of ORF1p and ORF2p have been published previously (Moran and Gilbert, 2002).

Figure 3.4A: Conservation of ORF1p

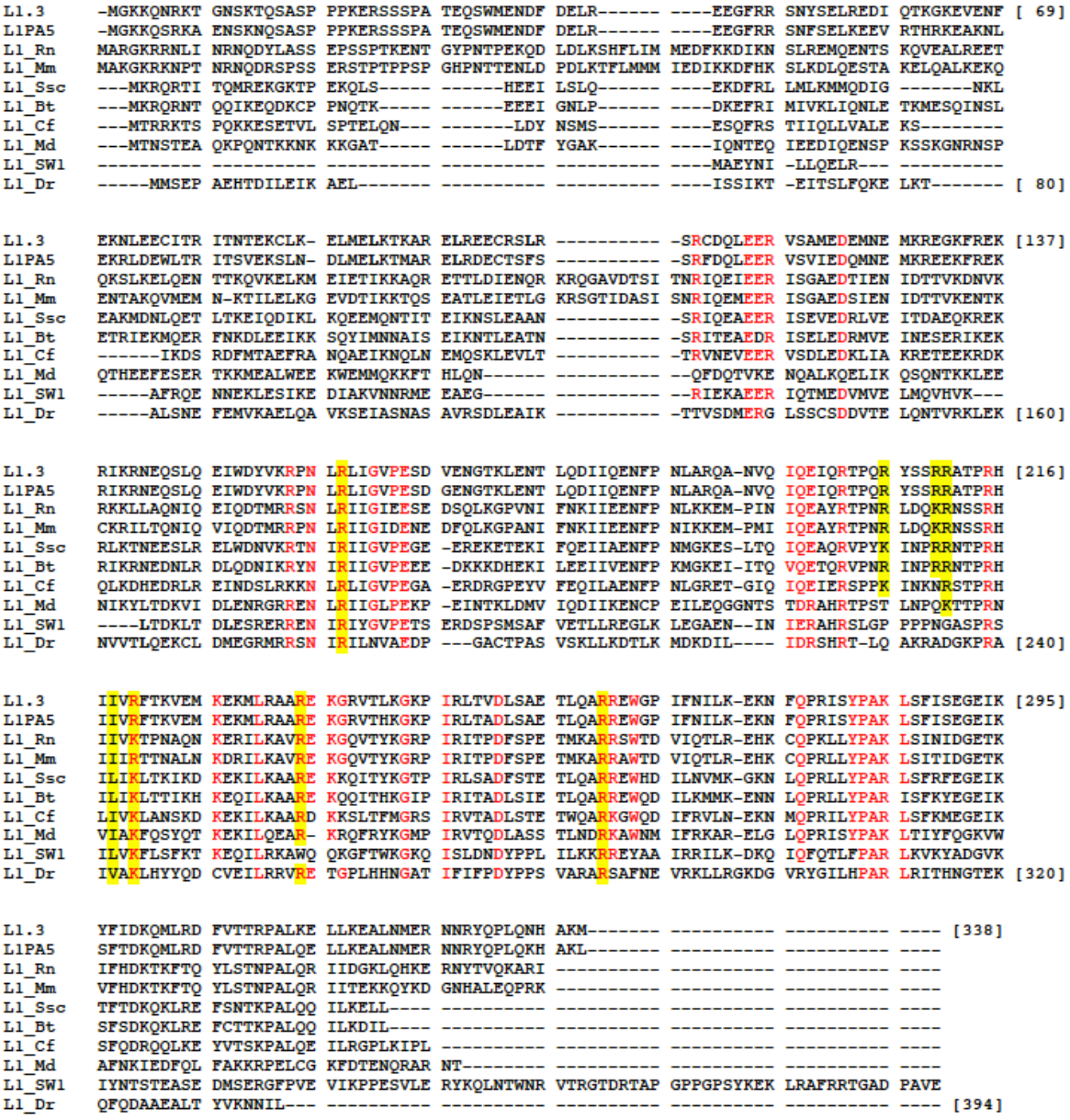
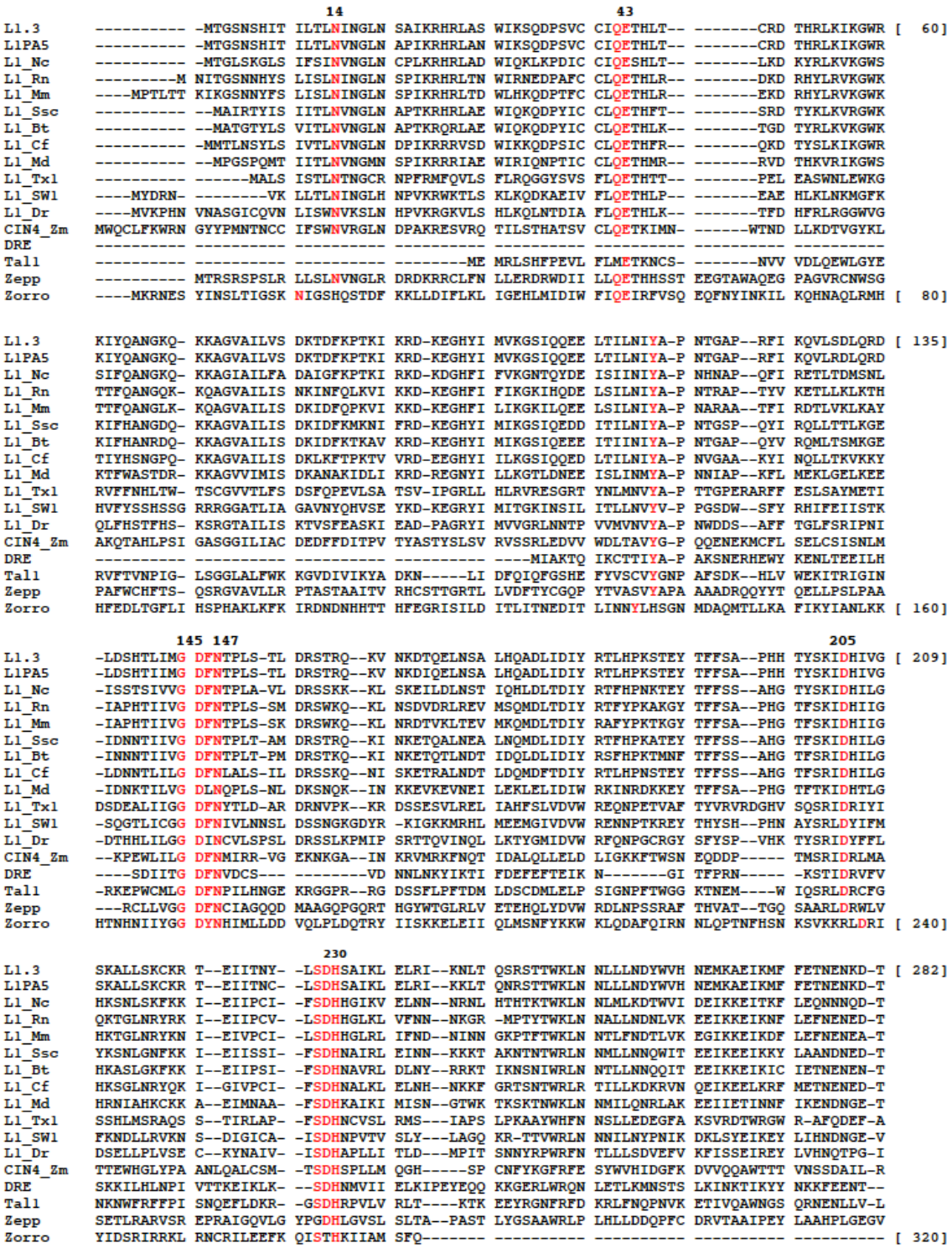
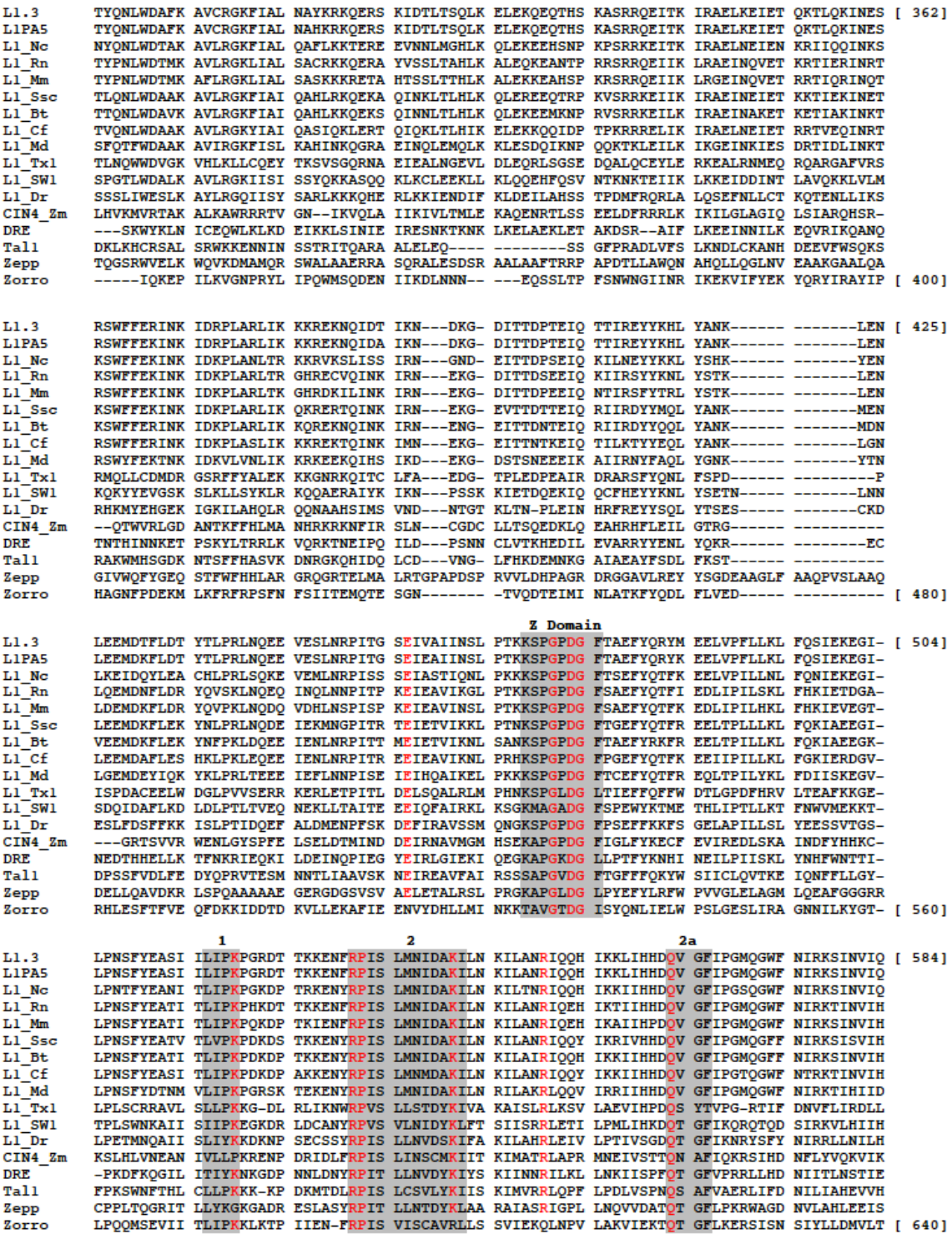
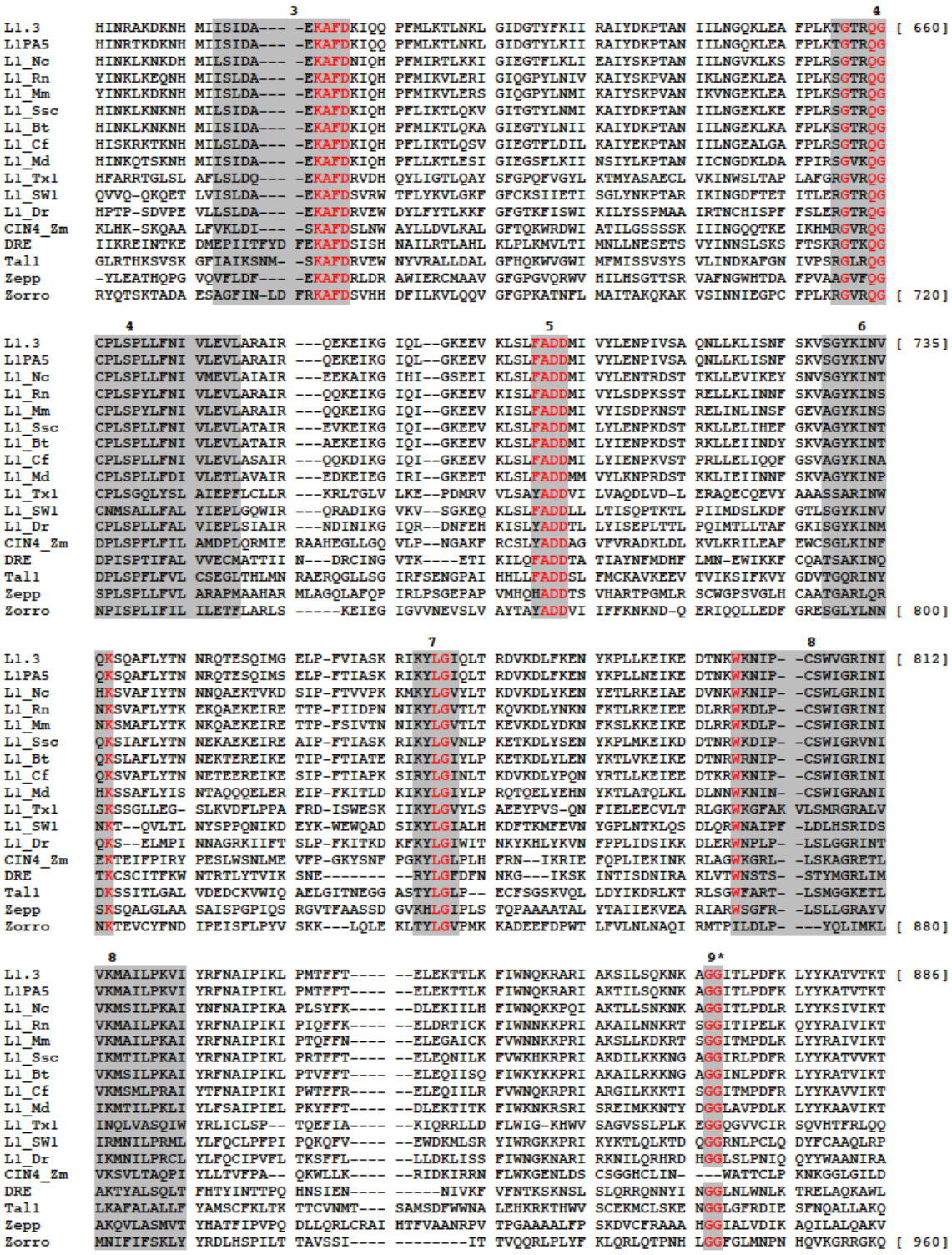


Figure 3.4B: Conservation of ORF2p




Z Domain
1
2
2a



```
Ll.3      AWYWYQNRDI DQ-------W NRTEP---SE IMPHIYNYLI FDKPEKNKQW GKD--SLFNK WCWENWLAIC RKLK------ [ 948]
LlPA5     AWYWYQNRDI DQ-------W NRTEP---SE ITPHIYNHLI FDKPDKNKKW GKD--SLFNK WCWENWLAIC RKLK------
Ll_Nc     AWYWHKNREV DV-------W NRIEN---QE MDPATYHYLI FDKPIKNIQW GKD--SLFNK WCWVNWLAIC RRLK------
Ll_Rn     AWYWYRDRQI DQ-------W NRIED---PE MNPHTYGHLI FDKGAKTIQW KKD--SIFSK WCWFNWRATC RRMQ------
Ll_Mm     AWYWYRDRQV DQ-------W NRIED---PE MNPHTYGHLI FDKGAKTIQW KKD--SIFNN WCWHNWLLSC RRMR------
Ll_Ssc    AWYWHKDRHI DQ-------W NRIES---PE LNPRTYSQLI YDKGGKNIQW RKD--SLFNK WCWENWTATW KRMK------
Ll_Bt     VWYWHKDRNM DQ-------W NKIES---PE INPRTYGHLI FDKGGKDIQW IKD--NLFNK WCWEIWSTTC KRMK------
Ll_Cf     VWYWHKNRHI DQ-------W NRIEN---PE VDPELYGQLI FDKGGKTIHW KKD--SLFNK WCWENWTSTC RRMK------
Ll_Md     IWYWLRNRKE DQ-------W NRLGE---ND LS-----KTV YDKPKDPSFW DKN--PLFDK NCWENWKTVW ERLG------
Ll_Txl    IQRYLYADPS PQ-------W CTLASSFYRQ VRNMGYDRQL FIIEPEGFLR NLSTLPAYYQ DTLKTWSMVS VLRQGATEGE
Ll_SWl    LICMCSPVYT AG-------W KDLEL---KT FEKIPLKALL ADLKLQGELL LQD--DPLLS MMIKTWNDTV KKCN------
Ll_Dr     MLHWSNPSYD SGP-----NW LSLEN---TS NFSTSLHALL CSNFPTPEPL SKYSLNPVVK HSLKIWAQFR RSFA------
CIN4_Zm   LERFAR---- ---------- ---------- ---------- ---------- -----ALRLR WLWLRWTNRD KAWT------
DRE       FERYLHQRVS NTP------- ---------- ---------- ---------- ---------S SYIKLWEEEL KNNN------
Tall      AWRLLQFPNS LFARFFKSRY YDEEDFLDAE LKATPSYAWR SILHGRDLLI KGFRKKVGNG SSTSVWMDPW IYDNDPRLPL
Zepp      VGRLLEPEQL AWK------- --------AN FDHWLYRSTA WLAAQEPGTC QRG------G STSGSWEDSC SSPP------
Zorro     IYLLYTQEDD LIIK------ ---------- ---------- --------FM RTKIQDILDN IAKDYYTMPT EPDK------ [1040]

Ll.3      --LDPFLTPY TKINSRWIKD LNVKPKTIKT LEENLGITIQ DIGVGKDFM- SKTPKAMATK DKIDKWDLIK LKSFCTAKET [1025]
LlPA5     --LDPFLTPY TKINSRWIKD LNVRPKTIKT LEENLGNTIQ DIGMGKDFM- SKTPKAMATK AKIDKWDLIK LKSFCTAKET
Ll_Nc     --LDPHLSPL TKIDSHWIKD LNLRHETIKI LEESAGKTLE GISLGEYFM- RRTPQAIEAV SKIHYWDLIK LKSFCTAKNI
Ll_Rn     --IDPCLSPC TKLKSKWIKD LHIKPDTLKL IEEKLGKHLE HMGTGKNFL- NKTPMAYALR SRIDKWDLIK LQSFCKAKDT
Ll_Mm     --IDPYLSPC TKVKSKWIKE LHIKPETLKL IEEKVGKSLE DMGTGEKFL- NRTAMACAVR SRIDKWDLMK LQSFCKAKDT
Ll_Ssc    --LEHSLTPY TKINSKWIKD LDIRPDTIKL LEENIGQTLS DINDSNIFS- DPPIRVLTIK RKINKWDLIK LQSFCTAKET
Ll_Bt     --LDHFLTPY TKINSKWIKD LNVRPETIKL LEENIGKTLS DIYHSRILY- DPPPRIMEIK AKINKWDLIN LKSFCTSKET
Ll_Cf     --LDHSLSPY TKINSKWMKD LNVRQDSIKI LEKNTGNTLF ELGHSNFLQ- DTSTKAKETK AKMNYWDFIK IRSFCTAKDT
Ll_Md     --IDQHLTPY TKINSKWVSD LNIKKETISK LGKHRIVYMS DLWEGKGFK- TKQDIERITK CKINNFDYIK LKSFCTNKTN
Ll_Txl    DILNEPLLYN PSFKTRMLES ISIRRRLCQA QLTRVGDLLD FEKSDWVDSQ AVMQRMGFLT TRVPHRLLKE IKDTISPDSH
Ll_SWl    --LMEDSKIL RWCTCDSEFT PNKYDGRFKL WIAKGLTDFN SFVHKGAFQT FDTLKKKHGL ISDDFFRYLQ VRHYFNQKIK
Ll_Dr     --LKGLS-AY APIARNHMFT PSTIDKTFDI WSMKGLKILK DMFIDGQFAS FQQVKVKFQI PNSHFFRYLQ LRSFVSSSMS
CIN4_Zm   ---------- -GLQLPCDKA DVDLFNASTI VTIGDGKMAD FWRSSWIQGQ APKNIAPTLF MKAKRKNISV CQALTNNRWM
DRE       ---------- ---NNKTTTK QNQLQLHWQC KQAWTQLKTP QNKQTHYEHL PKLKKIYEDM MTTQSPEHNK FIPTPGQKEI
Tall      QKHFSVNLDL RVHDLINVED RCRRRDRLEE LFYPADIEII VKRNPVVSMD DFWVWLHSKS GEYSVKSGYW LAFQTNKPEL
Zepp      -------SQR SMWRPRSAYG STCMPTSSCG PITCGSRKRS ATTQSWVSHC SSTGSSRMQS AALCLGGLGQ AWPGQDQPPA
Zorro     -------LIV YPWYLFGMGV SCHLSLQFRY LKERVYENLT KLEISWFEAW FQLVHYTGPP VETPIIIMLI EDYANLIIQP [1120]

Ll.3      TIRVNRQPTT --WEKIFATY SSDKGLISRI YNELKQIYKK KTNNPIKKWA KDMNRHFSKE DIYAAKKHMK KCSSSLAIRE [1103]
LlPA5     TIRVNRQPTE --WEKIFAIY SSDKGLISRI YKELKQIYKK KTNNPIKKWA KDMNRHFSKE DIYAANRHMK KCSSSLAIRE
Ll_Nc     VSKASRQPSE --WEKIFAGY TSDKGLITRI HRELKHINKK RTRDPISGWA RDLKRNFSKE DRHTIYKHMK KSSSSLIIRE
Ll_Rn     VVRTKRQPTD --WEKIFTNP TTDRGLISKI YKELKKLDRR ETNNPIKKWG SELNKEFTAE ECRMAEKHLK KCSTSLVIRE
Ll_Mm     VNKTKRPPTD --WERIFTYP KSDRGLISNI YKELKKVDFR KSNNPIKKWG SELNKEFSPE EYRMAEKHLK KCSTSLIIRE
Ll_Ssc    LNNTKRQPTE --WEKIFASE STDKGLISKI YKQLLQLHTK KTNNPIKKWA EDLNRQFSKE DIQMAEKHMK RCSASLIIRE
Ll_Bt     ISKVKRQPSE --WEKIIANE ATDKQLISKI YKRLLQLNSR KINDPIKKWA KELNRHFSKK DIQMANKHMK RCSTSLIIRE
Ll_Cf     VNKTQRQPTE --WEKIFAND ISDKGLVSKI YKELIKLNTK ETNNPIMKWA KDMNRNLTEE DIDMANMHMR KCSASLAIRE
Ll_Md     ITKIRRETTN --WEKIFIET S-DKGLITHI YNELNQLYKK SSHSPIDKWA REMDRQFSDK EIKTINKHMK KCSTSLIIRE
Ll_Txl    TFIDGVLHAG --EPRPPWNS SPPDIIIAPK TRQSPQAPPS PNLSQLENFP LTRFHDITRK LLYSLMLHTV HFLALISRYD
Ll_SWl    ISTDDPKFLK --TFKTLIKA ICPTKIISKL YNSILSHKGE NTYYVKERWE REGRLTITEE DWEQICRKQW ITTGSNIWRE
Ll_Dr     HYPSLPPPSL --LDSIMELS PYSKGLIGKI YSIINSHNLE PLVKLKRKWE VELEIELSED MWQSVLDNIH SSSICLKHRV
CIN4_Zm   HFCSPYTHED --EIKEFISL WQGINNTHKL NDFDDTILWR WTTDGSYSSS SAYKIQFTTN FCKIKISPIW KARIEPKCRF
DRE       MTKINSKHLP FKEIKKIINM KGRDLLWRYT LKALPKIYNM PCQQCGEDET SEHIFFNCKA HIKNTQEIFN YTLTKSGHTT
Tall      IREARVQPST NGLKEKIWST LTSPKIKLFL WRILSSALPV AYQIIRRGMP IDPRCQVCGE EGESINHVLF TCSLARQVWA
Zepp      RCRARPGASG --SDGARARA HPSGCLASSL ESANPGAATG RYVVRISDPS DRRVCLAQPK VAPTSQPHSK QHRSLLCQWT
Zorro     QTESIKLMSP -----NFEEQ SLSEQTFHHT SRKLCPQAPI IPEGWGKHFD TIKSLTISDW TEFWKNMNKV QKQSVGSLQD [1200]

                                                    Cysteine Rich Domain
Ll.3      MQIKTTMRYH LTPVRMAIIK KSGNNRCWRG ---------- ---------- -------CGE IGTLLHCWWD CKLVQPLWKS [1156]
LlPA5     MQIKTTMRYH LTPVRMAIIK KSGNNRCWRG ---------- ---------- -------CGE IGTLLHCWWD CKLVQPLWKT
Ll_Nc     MQIKTTLRYH LTPVRVAHIT KSPNQRCWRG ---------- ---------- -------CGG KGTLLHCWWE CPLIRSFWKD
Ll_Rn     MQIKTTLRFH LTPVRMAKIK NSGDSRCWRG ---------- ---------- -------CGE RGTLLHCWWD CRLVKPFWKS
Ll_Mm     MQIKTTLRFH LTPVRMAKIK NSGDSRCWRG ---------- ---------- -------CGE RGTLLHCWWE CRLVQPLWKS
Ll_Ssc    MQIKTTMRYH LTPARMAIIQ KSTNNKCWRG ---------- ---------- -------CGE KGTLVHCWWD CKLVQPLWKA
Ll_Bt     MQIKTTMRYH FTPVRMAAIQ KSTNNKCWRG ---------- ---------- -------CGE KGTLLHCWWE CKLVQPLWRT
Ll_Cf     IQIKTTMRYH LTPVRMGKIN KAGNNKCWRG ---------- ---------- -------CGE KGTLLHCWWE CELVQPLWKT
Ll_Md     MQIKTTLRYH LTPSRLANIT AKESNECWRG ---------- ---------- -------CGK VGTLIHCWWS CELIQPFWRA
Ll_Txl    TIWRRVLNEG ERPQWRAFYS SLVPRPTGDL SWKVLHGALS TGEYLARFTD SPAACPFCGK GESVFHAYFT CARLQPLLAL
Ll_SWl    FCWKSIMRFF TTPSQKKYLG ---NSKCWRC ---------- ---------- -------GNN GANHFHIFWD CVIIKKYWSD
Ll_Dr     IQFKVVHRLH WSKVKLAKFK ---PNIDPNC ---------- ---------- -------SIE PATLSHMFWA CSKLKKFWHL
CIN4_Zm   FAWTMLHNRI LTADNLQKRG WPCNPICCLC N---------- ---------- --------LS QETMPHLGKD CPFSVEVWNM
DRE       HTWNVKILNH LQIALVANLI AIIFDKIWHK ---------- ---------- ---------- RNKLIHDEKE --IIIHRQQV
Tall      LSGVPTSQFG FQNSSIFANI QYLLELKGKG LIP------- ---------- ----EQIKKS WPWVLWRLWK NRDKLFFEGT
Zepp      KALQPCTHSF LRPRSLHWFS TGTSPGPGIH ---------- ---------- -------AAA HVSSASTQSS CTQSSTCRWS
Zorro     YHLFILGYYS HYP-FYPNYK KLEIHFCQLC N---------- ---------- --------TG TDSIVHHIFE CGETMELWLR [1280]
```

```
L1.3      VWRFLR-DLE LEIPFDPAIP LLGIYPNEYK S------CCY KDTCTRMFIA ALFTIAKTWN QPKCPTMIDW IKKMWHIYT- [1228]
L1PA5     VWRFLK-DLE LEIPFDPAIP LLGIYPKDYK S------CCY KDTCTRMFIA ALFTIAKTWN QPKCPSMIDW IKKMWHIYT-
L1_Nc     VWRILR-DLK IDLPFDPIIP LLGLYPEDQK S------QYN KDICTRMFIA AQFIIAKSWK KPKCPSTHEW TSKLWYMYT-
L1_Rn     VWRFLR-KLD IELPEDPAIP LLGIYPKDAS -------TYK RDTCSTMFIA ALFIIARSWK EPRCPSTEEW IQKMWYIYT-
L1_Mm     VWRFLR-KLD IVLPEDPAIP LLGIYPEDAP -------TGK KDTCSTMFIA ALFIIARSWK EPRCPSTEEW IQKMWYIYT-
L1_Ssc    VWRFLR-KLN IELPFDPAIP LLGIYPEKT- -------TTR KDTCTPMFIA ALFTIAKTWK QPKCPSTEEW IQKMWYIYT-
L1_Bt     VWRFLK-KLE IEPPYDPAIP LLGIHTEKT- -------RRE RDTCTPMFIA ALFIIAKTWK QPRCPSADEW IRKLWYIYT-
L1_Cf     VWRFLK-QLK IYLPYDPAIA LLGIYPKDTN A------MKR RDTCTPMFLA AMATIAKLWK EPRCPTKDEW IKKMWFMYT-
L1_Md     IWNYAQRATK EYLPFDPAIA LLGLYPKEIM ---------D TKTCTKIFIA ALFVVAQNWK TRGCPSIGEW LNKLWYMLV-
L1_Tx1    LRKLYL---Q FWLHFSPHVY IFGRPVSRDN K-------EK DLLSNLLLAL AKLVIHKSRK QCLEGGNPLP AEVLFRVLVR
L1_SW1    IHEHLQNVFS -IVFPLSFES LFLSKIDGLD NK-----NKK LLYILLA--A SKKAITRKWL KPEQPTTEDW IDVVQRIYI-
L1_Dr     IFKFLSDALN TYVEPEAIIS IFGITPQSLC FN-----KSK INVIAFATLL ARRLILLKWK EKLPPTFKQW LMELLHHLT-
CIN4_Zm   ILSWVN---- ---------- ---------- ---------- ---LSFLSGV SNLGSLYDWW KRLRNRCCKE SKKIFDGLI-
DRE       IRELIKTQRA AWDRTQAVIN KTLRIKSKQR P-------EE QNKLDSLISL KLLQFSRQWN ---SPLHAIE LPKHLKKYN-
Tall      IFSPLKSIEK IRDDVQEWFL AQALVASVDA GETVCSAPCP SSWEPPPLGW VKCNISGVWS GKKRVCGGAW VLRDDHGKVL
Zepp      SRQGSSSQGS SWSGSSSCRS HCHHR----- ---------R PHLYFLCLWG SQVQDPQAWG LGTRPAHEFV VREATARRQ-
Zorro     HFNSQR---- ---------- ---------- ---------- ----TPQFII GNKHLHKRDL YALNEYIKEV VQKVKRRRG- [1360]

L1.3      ----MEYYAA IKND-EFISF VGTWMKLETI ILSKLSQEQK TKHRIFSLIG GN-------- ---------- ---------- [1275]
L1PA5     ----MEYYAA IKND-EFMSF VGTWMKLETI ILSKLSQGQK TKHRMFSLIG GN-------- ---------- ----------
L1_Nc     ----MEYYAA LKKDGDFTSF MFTWMELEHI LLSKVSQ--- ---------- ---------- ---------- ----------
L1_Rn     ----MEYYSA IKNN-EFMKF VGKWLELENI ILSELTQSQK DIHGMHSLIS GY-------- ---------- ----------
L1_Mm     ----MEYYSA IKKN-EFMKF LAKWMDLEGI ILSEVTHSQR NSHNMYSLIS GY-------- ---------- ----------
L1_Ssc    ----MEYYSA IKKN-EIPAF LATWMDLETI MLSEVSHTMR HQHQMLSLTC GI-------- ---------- ----------
L1_Bt     ----MEYYSA IKKN-TFESV LMRWMKLEPI IQSEVSQKEK HKYSILTHIY GI-------- ---------- ----------
L1_Cf     ----MEYYSA IRND-KYPPF ASTWMELEGI MLSEVSQSEK DKHYMFSFIW GI-------- ---------- ----------
L1_Md     ----MEYYCA QRNN-KVEKF HGDWNNLQEV MQSERSRTRR TLYTETNTLW YNRT------ ---------- ----------
L1_Tx1    SRIRAEYTQA VFTG-RLKEF ADQWAIDGVL CSVSPDLVSV QTILTLPYLS AL-------- ---------- ----------
L1_SW1    ----MERISH SLQI-RLDVF YATWSIWTEY VKPVRSDFI- ---------- ---------- ---------- ----------
L1_Dr     ----LEKIRY TFGG-CTDMF FLTWQPVLDH VKKMDPSVIL EE-------- ---------- ---------- ----------
CIN4_Zm   ---------- -------ISF ---------- ---------- ---------- ---------- ---------- ----------
DRE       ---------- -------NSL STFYK----- ---------- ---------- ---------- ---------- ----------
Tall      LHSRRAFSNL SVKK-DALFC CVKWALRVCL ATDSLRFCLL LSLGLTVRFC QAESLALFCF SCFRTNSFFG EDWRLEGLGG
Zepp      ---VLHRISK GEASPAGPLR PAIWADTTDD QRTPSGARGA MGGQGCSCWY LRRRS----- ---------- ----------
Zorro     ---------- -------SPI LNQGERENVD AGTNVLV--- ---------- ---------- ---------- ---------- [1440]

L1.3      ----
L1PA5     ----
L1_Nc     ----
L1_Rn     ----
L1_Mm     ----
L1_Ssc    ----
L1_Bt     ----
L1_Cf     ----
L1_Md     ----
L1_Tx1    ----
L1_SW1    ----
L1_Dr     ----
CIN4_Zm   ----
DRE       ----
Tall      KGRF
Zepp      ----
Zorro     ---- [1444]
```

Figure 3.5: Inactive L1s Contain Potential Endonuclease Mutations

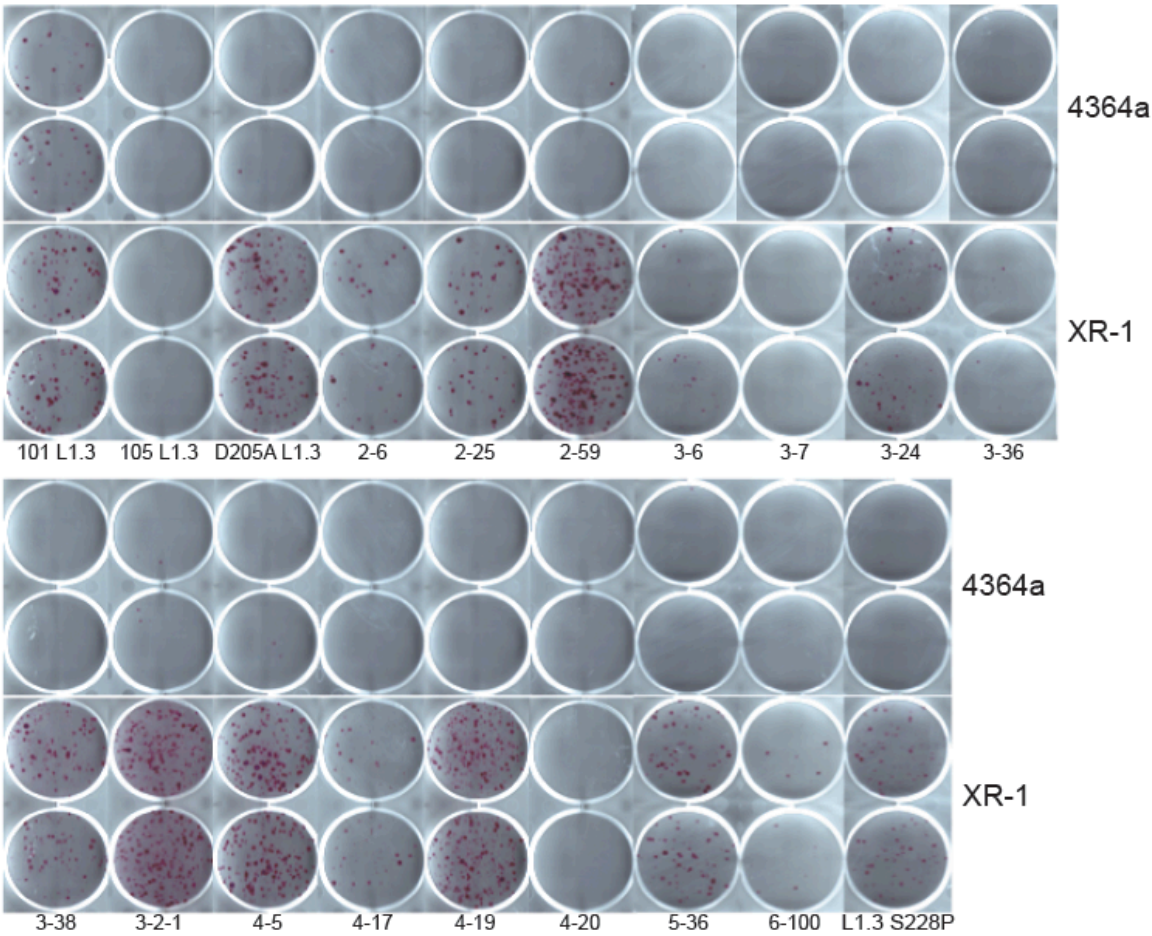


Figure 3.5: Inactive L1s Contain Potential Endonuclease Mutations

Endonuclease dependent (4364a) and independent (XR-1) L1 retrotransposition assays for all 15 elements containing putative amino acid mutations. The figure includes a positive control (JM101 L1.3), a negative RT- control with a D702A mutation (JM105 L1.3), and L1.3 with a previously characterized EN mutation (D205A) (Feng et al., 1996; Moran et al., 1996; Morrish et al., 2002; Wei et al., 2001). Retrotransposition events are detected as Blasticidin-resistant colonies.

Figure 3.6: L1Hs Subfamilies Cluster in a Phylogenetic Tree

(A) Relationships between the 53 L1Hs elements in this study were inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the L1s analyzed (Felsenstein, 1985). The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test are shown above the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Phylogenetic analyses were conducted in MEGA4 (Tamura et al., 2007). Brackets at right indicate individual transduction "subfamilies" that are shaded on the tree. Light shading indicates active transduction-containing L1s, and dark shading indicates inactive elements in the subfamily. **(B)** Putative RP progenitor retrotransposition assay results. Percentage retrotransposition efficiencies at the bottom of the figure are reported relative to L1.3.

Figure 3.6: L1Hs Subfamilies Cluster in a Phylogenetic Tree

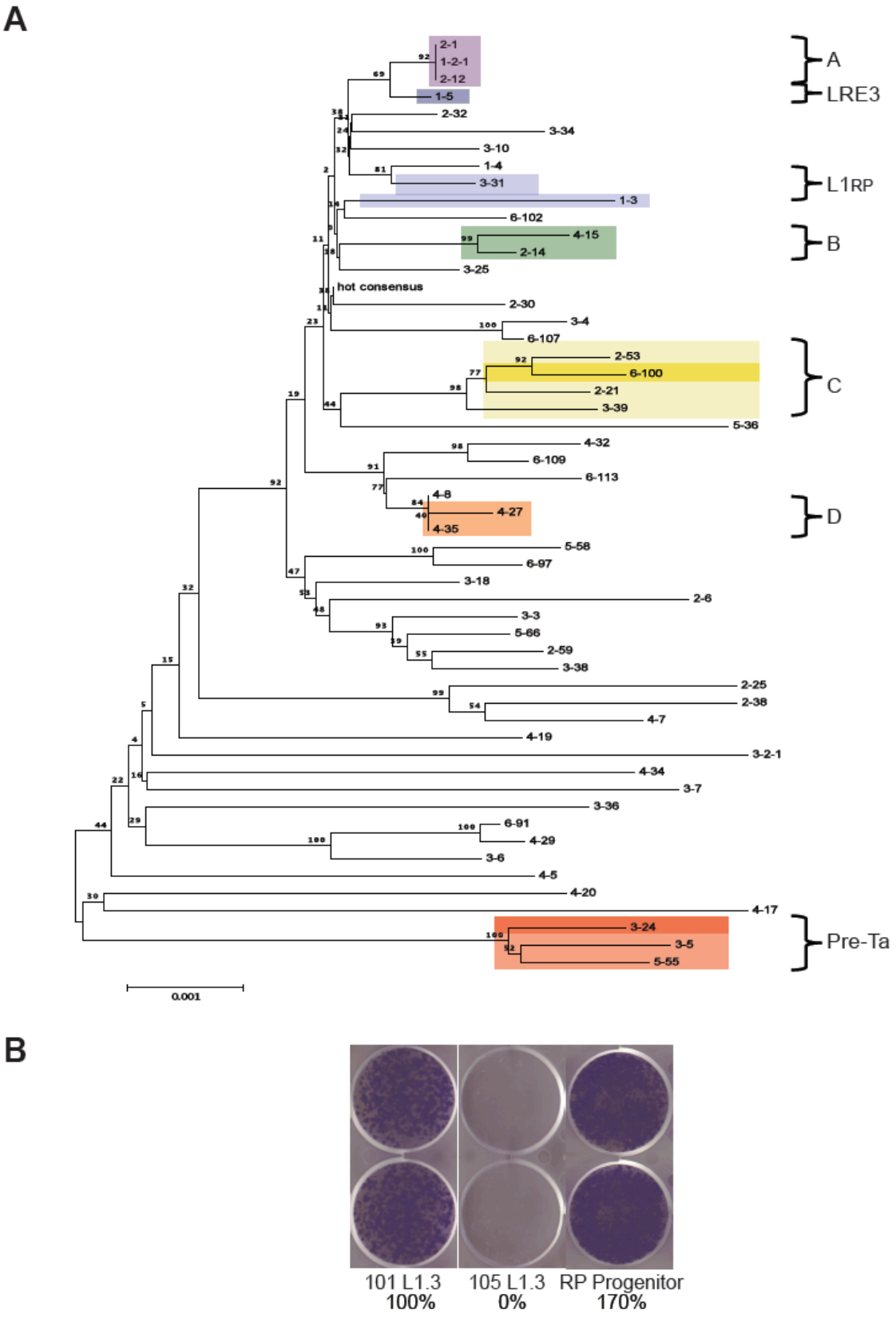


Table 3.1: 5′UTR Splice Site Predictions From BDGP

The Table depicts splice sites present in the 5′UTR of L1.3 within a 0.2 cutoff value from the Berkeley Drosophila Genome Project (BDGP) splice site website (http://www.fruitfly.org/seq_tools/splice.html). Both sense and antisense strands were examined. The first and second columns indicate the beginning and end of the donor or acceptor sequences. The third column is the score as given by the BDGP site. The fourth column is the sequence of the site with the donor GT or acceptor AG nucleotides indicated in capital letters. Nucleotides that would be spliced out of the final insertion are indicated in red.

Table 3.1: 5′UTR Splice Site Predictions From BDGP

Sense Strand

Donor Sites:

Start	End	Score	Sequence
48	62	0.52	tcccagc**GT**gagcga
67	81	0.24	gaagacg**GT**gatttc
91	105	0.81	atctgag**GT**accggg**
461	475	0.83	tgctttag**GT**aaacaa
688	702	0.54	tcctcaa**GT**gggtcc

Acceptor Sites:

Start	End	Score	Sequence
26	66	0.34	aggaacagctccggtctac**AG**ctcccagcgtgagcgacgca
98	138	0.66	gtaccgggttcatctcact**AG**ggagtgccagacagtgggcg
282	322	0.86	cccgaatattgcgcttttc**AG**accggcttaagaaacggcgc
447	487	0.27	ccattgcccaggcttgctt**AG**gtaaacaaagcagccgggaa
521	561	0.52	gaggcctgcctgcctctgt**AG**gctccacctctgggggcagg
601	641	0.46	ttaagtgtccctgtctgac**AG**ctttgaagagagcagtggtt**

Anti-Sense Strand

Donor Sites:

Start	End	Score	Sequence
798	784	0.22	gctgcag**GT**ctgttg
351	337	0.61	gagccag**GT**gtggga
266	252	0.95	tttccag**GT**gcgacc
203	189	0.99	ttcccag**GT**gaggca
166	152	0.59	gcgcacg**GT**gcgcac
108	94	0.44	gaacccg**GT**acctca

Acceptor Sites:

Start	End	Score	Sequence
846	806	0.86	gatgtcctttctggttgtt**AG**ttttccttctaacagacagg
831	791	0.89	tgttagttttccttctaac**AG**acaggaccctcagctgcagg
812	772	0.27	agacaggaccctcagctgc**AG**gtctgttggaataccctgcc
684	644	0.97	agtctgtctgcccgttctc**AG**atctccagctgcgtgctggg
642	602	0.36	gaaccactgctctcttcaa**AG**ctgtcagacagggacactta
571	531	0.41	ttgtctgtgccctgcccc**AG**aggtggagcctacagaggca
483	443	0.31	cggctgctttgtttacctaa**AG**caagcctgggcaatggcggg
430	390	0.70	agcctcgttgccgccttgc**AG**tttgatctcagactgctgtg
419	379	0.29	cgccttgcagtttgatctc**AG**actgctgtgctagcaatcag
280	240	0.45	gggagtgacccgattttcc**AG**gtgcgaccgtcacccctttc
250	210	0.86	tcacccctttctttgactc**AG**aaagggaactccctgacccc
217	177	0.95	ctgacccctttgcgcttccc**AG**gtgaggcaatgcctcgccct
135	95	0.31	ccactgtctggcactccct**AG**tgagatgaacccggtacctc

Table 3.2: Elements in the HGR With the Same Splice Junction as #6-113

Table 3.2 depicts the 96 L1s from the HGR that contain the same splice variant seen in #6-113 (the location of #6-113 is listed in row 1 of the table). Column 1 lists the chromosome with the L1 insertion. Column 2 indicates the approximate nucleotide position of the beginning of the L1 in the hg19 version of the human genome reference sequence. Column 3 is the family to which the L1 belongs. L1P1 contains L1 PA2-PA3 elements and L1P2 contains L1 PA4-PA6 elements. Column 4 is the strand on which the insertion occurred. Column 5 is the length of the L1, from the first nucleotide after the 5′ target site duplication to the beginning of the poly(A) tail. Column 6 is the target site duplication sequence, if one could be discerned. Column 7 indicates whether an L1 resides in an intron of a RefSeq gene. Column 8 is the orientation of the L1 with respect to the transcript of the gene. ***** indicates three L1s with the same TSDs that are duplicated in a segmental duplication. **#** indicates an L1 element that is prematurely polyadenylated within the poly purine track of the 3′UTR.

Table 3.2: Elements in the HGR With the Same Splice Junction as #6-113

Chr.	Loc. Start	L1	Strand	Length	TSD	Gene	Orient.
4	98103775	L1Hs	-	5489	AAAGATTGTGTGTCCGG	-	-
X	100795911	PA3	+	5508	T/AGGAGATATACCTAATGTA	-	-
X	68683760	PA3	-	(1) 5488	T/AAAGACATTGC	-	-
X	140536457	PA5	+	5652	None Discernable	*SPANX*	Opp.
X	112689296	PA3	+	5491	T/AAAAGCAGTATTCCC	-	-
X	151493771	PA5	+	5612	T/AAAAATATAATTTATG	*C1orf146*	Opp.
*1	206506538	PA2	-	5490	T/AAATGATTCAGTGTAG	-	-
*1	144043832	PA2	-	5490	T/AAATGATTCAGTGTAG	*SRGAP2P*	Opp.
1	92703162	PA3	-	5607	T/AAAAATATAATTTATG	*C1orf146*	Opp.
1	104464687	L1Hs	+	3398	N/A (3' Truncation/Interrupted)	-	-
1	101147989	PA3	-	5634	C/AAAATTATGTTATA	-	-
1	228843130	PA3	+	(12) 5493	T/AAAATACCCAAAGC	*RHOU*	Same
1	71681575	PA3	+	5488	T/AAGAAACAATGTAA	*BC054887*	Same
1	37673603	PA3	+	5496	N/A (3' Truncation/Interrupted)	-	-
1	197798854	PA4	+	5614	None Discernable	-	-
2	193902611	PA3	-	5485	C/AGAATGATACATGTATT	-	-
2	162392756	PA3	+	5468	C/AAGAAAATGA	-	-
2	132131447	PA3	+	(1) 5622	T/AAAGTTATAAGG	-	-
2	49309217	PA4	+	3494	N/A (3' Truncation/Interrupted)	*FSHR*	Opp.
2	48285530	PA5	-	(8) 5620	C/A_{20} (?)	-	-
2	77132072	PA3	+	5615	T/GAAAAGGCATCATTCTT	*LRRTM4*	Opp.
2	77363207	PA5	+	(12) 5604	T/AAAAGGTTTAAC	*LRRTM4*	Opp.
2	97106917	PA3	+	5494	T/AAGAAAATGTGACAC	-	-
2	21071902	L1P1	+	238	N/A (3' Truncation/Interrupted)	-	-
3	158808095	PA3	-	5505	C/AGAACTTATG	*IQCJ/SCH*	Opp.
3#	12074759	PA2	-	5413	C/AAGAATTACTTATAACC	*SYN2*	Opp.
3	23343273	PA4	-	5578	None Discernable	*UBE2E2*	Opp.
3	135541406	PA3	+	(8) 5621	C/AAAAATACAAAAATT	-	-
3	57592577	PA5	+	(14) 5609	T/AAAGATGTATATA	-	-
3	24321653	PA3	+	5486	T/AAGCTATAGAACC	*THRB*	Opp.
4	99241647	PA2	+	5489	None Discernable	*RAP1GDS*	Same
4	78440449	PA3	-	(1) 5600	T/GAGATACCATCTCAT	*CXCL13*	Opp.
4	53371019	PA4	-	5594	None Discernable	-	-
4	167187141	L1Hs	+	5651	None Discernable	-	-
4	116093594	PA3	+	(6) 5499	G/AAATAGTGCTT (?)	-	-
4	128781276	PA3	+	5985	None Discernable	-	-
5	132521520	PA3	-	5501	C/AATATAGGAATC	-	-
5	139895099	PA3	+	(8) 5487	T/AAAGACTATCCAG	*ANKHD1*	Same
5	8159239	PA3	+	5493	None Discernable	-	-
5	147989815	PA3	+	5493	G/AAAGCTCCTAAAGGT (?)	*HTR4*	Opp.
5	108962814	PA3	-	5434	$T/AAAAGA_{10}$	-	-
5	125422248	PA6	+	5593	None Discernable	-	-
6	30215666	PA5	+	5483	None Discernable	-	-
6	27911128	L1Hs	+	859	N/A (3' Truncation/Interrupted)	-	-
6	77460735	PA3	-	5480	None Discernable	-	-
6	115330203	PA3	+	(8) 5497	T/AGAAACATAGCTATCATT	-	-
6	108712068	L1P1	-	2759	N/A (3' Truncation/Interrupted)	*LACE1*	Opp.
6	83342545	PA3	+	5425	T/AAATAAAATAA (?)	-	-
6	81530533	PA3	+	(3) 5377	T/AAAATTTTGACTT	-	-
7	86868039	PA3	+	5525	G/AAAGGCC (?)	-	-

Chr.	Loc. Start	L1	Strand	Length	TSD	Gene	Orient.
7	62521183	L1Hs	+	3822	N/A (3' Truncation/Interrupted)	-	-
7	89531589	PA3	-	5484	T/AAGAAAATGTGGCA	-	-
7	14744169	PA2	+	5488	C/A17GA6GA7 (??)	*DGKB*	Opp.
7	148086130	L1P1	-	254	N/A (3' Truncation/Interrupted)	*CNTNAP2*	Opp.
7	83378570	PA3	-	5621	None Discernable	-	-
7	140413905	PA4	-	5602	None Discernable	-	-
8	69274711	PA2	+	(20) 5546	T/GAAAAAAGTACATC	*C8orf34*	Same
8	3133822	PA3	-	(4) 5485	T/AAGATGTGACTTGC	*CSMD1*	Same
8	129724704	PA2	-	5545	None Discernable	-	-
8	130336608	PA3	+	5485	T/AAAATGTAAAAATCA	-	-
9	129755912	PA4	+	5568	None Discernable	*RALGPS1*	Same
10	7105374	PA2	-	5532	G/AAAGAAGGCAGAGATC	-	-
10	55557481	PA3	-	5487	C/GAAAGAAAGG	-	-
10	111365269	PA3	+	5526	T/AAGATATTAAA	-	-
10	7179477	PA2	+	(23) 5515	T/AAAATGATCACTGAATC	-	-
*10	44983643	PA2	-	5490	T/AAATGATTCAGTGTAG	-	-
10	36302543	PA4	-	5604	None Discernable	-	-
10	15712011	PA4	-	(5) 5608	T/AAGAGTAAATG	*ITGA8*	Same
10	50454262	PA3	+	(1) 5491	T/AAAAATACCACAAAG	-	-
11	13820761	PA3	-	5493	T/AAAAATAGCTTTC	-	-
11	100523814	PA3	+	(3) 5496	T/GAAAGCATA	-	-
11	29437736	PA3	+	5495	T/AAAATTTTT (?)	-	-
11	48588630	PA4	-	(1) 5612	C/AATAGCAGAAC	-	-
11	26314480	L1P2	-	3236	N/A (3' Truncation/Interrupted)	*ANO3*	Opp.
12	61160617	PA3	-	5481	T/AAAATTTTT (?)	-	-
12	10502325	L1Hs	-	(1) 5490	G/AAAGAAATACTATACAGC	-	-
12	77658270	L1P1	-	4120	N/A (3' Truncation/Interrupted)	-	-
12	55939663	PA5	-	2609	N/A (3' Truncation/Interrupted)	-	-
12	23674088	PA4	-	5612	T/AAAAATGATTCTTTGA	-	-
12	108116495	PA5	+	(5) 5612	T/AAATGTCATTTCCT	-	-
13	47952847	PA3	+	(11) 5621	T/CAAAAGTA (?)	-	-
13	82039211	PA4	-	5580	None Discernable	-	-
14	66327054	L1Hs	+	191	N/A (3' Truncation/Interrupted)	-	-
14	50539629	L1P1	+	245	N/A (3' Truncation/Interrupted)	-	-
14	44945865	PA4	-	5565	None Discernable	-	-
14	35649420	PA5	-	(6) 5570	T/AAGAATGAACTG	*KIAA0391*	Opp.
15	71022057	L1Hs	+	5494	None Discernable	*UACA*	Opp.
15	51465762	PA2	+	5530	T/AATATTT (?)	-	-
16	63181637	PA3	+	5566	None Discernable	-	-
17	31391710	PA5	-	(5) 5647	C/AAAAGACATACTGAGTTT	*ACCN1*	Same
17	15258595	PA5	+	5591	None Discernable	-	-
18	47302813	PA3	-	5482	None Discernable	-	-
18	44508546	PA4	-	5632	None Discernable	*DKF7 ?*	Opp.
19	35348866	L1P2	+	2269	N/A (3' Truncation/Interrupted)	-	-
20	24538447	PA3	-	5630	T/AAAGCCCTCTTCC	*TMEM90B*	Opp.
20	12643382	PA3	+	(1) 5563	G/AAAAAAGGGTAAACT	-	-
21	18535869	PA3	+	5477	None Discernable	-	-

Table 3.3: Mutations in Inactive L1 Elements Often Occur in Conserved Residues

The Table depicts the 15 L1 elements from a previous study that contain two intact ORFs, and yet are either inactive or display low-level activities in HeLa cells (Beck et al., 2010). Column 1 depicts the element number as it appears in Beck et al., 2010. Column 2 lists mutations in ORF1p that are not seen in any of the highly active L1s. Column 3 lists mutations in ORF2p that are not seen in any of the highly active L1s. Columns 4 and 5 are the amino acids from 2 and 3 that are conserved through at least *canine lupus familiaris* in evolutionary comparisons (Moran and Gilbert, 2002) (Figure 3.4). Column 6 is efficiency of the L1 in the ENi retrotransposition assay when compared to L1.3 (Figure 3.5).

Table 3.3: Mutations in Inactive L1 Elements Often Occur in Conserved Residues

Element ID	All ORF1p	All ORF2p	Conserved ORF1p	Conserved ORF2p	ENi (% L1.3)
#2-6	E 41 K	I 31 V D 289 N V 582 I V 813 M T 1034 I M 1110 V	-	I 31 V D 289 N V 813 M	32
#2-25	C 111 Y Q 277 H Y 282 F	T 525 A D 635 G T 744 N A 1205 P	**Q 277 H**	T 744 N A 1205 P	90
#2-59	-	T 44 I K 240 E	-	**T 44 I**	257
#3-6	S 254 L R 328 W	I 10 V T 224 R E 323 D I 517 T	S 254 L	I 10 V **I 517 T**	21
#3-7	K 3 I M 91 I I 181 N P 278 T S 290 N	V 208 L E 310 K **P 611 R** A 650 V S 698 T E 783 D L 1141 S	**I 181 N** P 278 T	E 310 K P 611 R S 698 T L 1141 S	0
#3-24	-	S 228 P	-	**S 228 P**	66
#3-36	T 222 N	R 49 T R 134 K N 147 S Y 181 C T 220 A L 257 P D 403 N D 434 N I 679 M T 1139 K L 1140 I C 1193 R	-	**N 147 S** Y 181 C D 403 N I 679 M T 1139 K L 1140 I C 1193 R	3
#3-38	-	T 192 P T 1109 A I 1268 V	-	T 192 P T 1109 A	250
#3-2-1	T 95 I N 327 K	N 279 S A 290 T I 340 V T 392 R R 440 T R 855 C L 1114 F N 1128 Y Y 1232 H	N 327 K	N 279 S I 340 V R 440 T R 855 C Y 1232 H	339

Element ID	All ORF1p	All ORF2p	Conserved ORF1p	Conserved ORF2p	ENi (% L1.3)
#4-5	E 199 K	P 197 S S 231 T R 311 G P 376 L T 431 K K 764 R K 790 N **M 908 T** A 1100 V V 1150 I T 1228 S	**E 199 K**	P 376 L T 1228 S	328
#4-17	T 9 A N 68 I	Y 305 H R 375 G **P 611 R** R 775 W K 876 R C 936 W I 943 L M 1002 I A 1234 T	-	R 375 G P 611 R R 775 W C 936 W A 1234 T	33
#4-19	A 17 T	K 280 N **M 908 T** A/D/T1006V A 1199 V P 1211 L	-	A 1199 V P 1211 L	289
#4-20	F 268 L T 308 A	H 45 Y I 98 L L 347 W K 397 E F 476 L L 495 R N 544 S **P 611 R** K 727 E A 762 T K 1190 E	-	**H 45 Y** I 98 L **F 476 L** L 495 R N 544 S P 611 R	0
#5-36	R 24 H R 49 L I 197 T Y 282 C R 325 S Q 333 K	S 498 T I 761 T	I 197 T **Y 282 C**	I 761 T	86
#6-100	L 90 R	H 331 Q N 669 I	-	N 669 I	9

References

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 Dev *20*, 210-224.

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 Res *32*, 3846-3855.

Babushok, D.V., and Kazazian, H.H., Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat *28*, 527-539.

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

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. Annu Rev Genomics Hum Genet *12*, 187-215.

Belancio, V.P., Hedges, D.J., and Deininger, P. (2006). LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res *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.

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

Boissinot, S., Entezam, A., and Furano, A.V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol *18*, 926-935.

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

Boissinot, S., and Furano, A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol *18*, 2186-2194.

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. Am J Hum Genet *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. Proc Natl Acad Sci U S A *100*, 5280-5285.

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 I1. Genomics *80*, 402-406.

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. Embo J *21*, 5899-5910.

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

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.

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci U S A *90*, 6513-6517.

Dong, C., Poulter, R.T., and Han, J.S. (2009). LINE-like retrotransposition in Saccharomyces cerevisiae. Genetics *181*, 301-311.

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 Genet *6*.

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. J Biol Chem *279*, 27753-27763.

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

Ewing, A.D., and Kazazian, H.H. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

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 Res *15*, 2251-2260.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution *39*, 783-791.

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.

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques *17*, 46, 48-49, 52.

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

Goodier, J.L., and Kazazian, H.H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23-35.

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

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. Hum Mol Genet *20*, 3386-3400.

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

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. Nat Genet *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. J Biol Chem *267*, 19765-19768.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

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.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

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.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

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 Res *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. Nat Struct Mol Biol *18*, 1006-1014.

Khazina, E., and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A *106*, 731-736.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F*., et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

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.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet *8*, 1557-1560.

Kopera, H.C., Moldovan, J.B., Morrish, T.A., Garcia-Perez, J.L., and Moran, J.V. (2011). Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. Proc Natl Acad Sci U S A.

Kubo, S., Seleme Mdel, 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. Proc Natl Acad Sci U S A *103*, 8036-8041.

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

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *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.

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. Proc Natl Acad Sci U S A *87*, 6990-6994.

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.

Marchani, E.E., Xing, J., Witherspoon, D.J., Jorde, L.B., and Rogers, A.R. (2009). Estimating the age of retrotransposon subfamilies using maximum likelihood. Genomics *94*, 78-82.

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. J Mol Biol *247*, 49-59.

Martin, S.L. (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol *11*, 4804-4807.

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.

McMillan, J.P., and Singer, M.F. (1993). Translation of the human LINE-1 element, L1Hs. Proc Natl Acad Sci U S A *90*, 11533-11537.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K.*, et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

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

Moran, J.V., and Gilbert, N. (2002). Mammalian LINE-1 retrotransposons and related elements. Mobile DNA II.

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. Nat Genet *31*, 159-165.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V.*, et al.* (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet *71*, 312-326.

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

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res *11*, 2050-2058.

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

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J Comput Biol *4*, 311-323.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

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.

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

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

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

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

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

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol *24*, 1596-1599.

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

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.*, et al.* (2005). Fine-scale structural variation of the human genome. Nat Genet *37*, 727-732.

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. Mol Cell Biol *21*, 1429-1439.

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem *284*, 435-438.

Weichenrieder, O., Repanas, K., and Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure (Camb) *12*, 975-986.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res *19*, 1516-1526.

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 Res *31*, 4929-4940.

Chapter 4

Conclusion

Overview

Clearly, LINE-1 elements continue to impact human genomes. Though the interspersed, repetitive nature of L1s has inhibited studies on a genome-wide scale, the extent of the effects of these mobile elements on human genomes are beginning to be understood. L1s can mediate the retrotransposition of other elements, act as substrates for ectopic recombination, disrupt genes, and impact genome structure and function in a number of other ways. In this concluding Chapter, I will discuss how findings presented in this thesis elucidate the role of L1 in inter-individual human variation. Results from Chapter 2 provide new insights regarding the model of L1 subfamily succession. The use of L1 sequences from this study to investigate splicing, amino acid changes, and the discovery of active L1 elements through transduction sequences is explored in Chapter 3. I will also propose future experiments to analyze the impact of L1 retrotransposition in humans, the importance of specific amino acid or RNA changes to retrotransposition, and the prevalence of splicing mutations in the 5'UTR.

LINE-1 in Human Variation

How Prevalent is L1 Retrotransposition in Humans?

Mutagenic L1 and Alu insertions indicate that retrotransposon activity continues to impact modern humans (see Chapter 1) (Beck et al., 2011; Kazazian et al., 1988). Additionally, genotyping and PCR-based methodologies have illustrated that there are hundreds if not thousands of polymorphic L1s resident in human genomes (Badge et al., 2003; Boissinot et al., 2004; Sheen et al., 2000). In Chapter 2, we identified 68 full-length L1s in the genomes of 6 geographically diverse humans. This publication (Beck et al., 2010) and other recent profiles of L1 and Alu sequences from genome-wide studies have begun to characterize the retrotransposon diversity present in humans (Ewing and Kazazian, 2011, 2010; Huang et al., 2010; Iskow et al., 2010; Mills et al., 2011; Witherspoon et al., 2010; Xing et al., 2009). These findings have underscored the importance of L1s and Alus in human variation, and have added to the single nucleotide polymorphisms and copy number differences that populate our genomes.

In Chapter 2, we specifically identified full-length L1 polymorphisms. These ~6 kb elements represent only ~1/3 of the human-specific L1 Ta subfamily yet are responsible for L1 amplification in humans (Boissinot et al., 2001). The fosmid paired-end sequencing strategy utilized in Chapter 2 also provided us with ~40 kb segments of an individual's genome that contained the L1. Thus, instead of previous methods that had relied upon PCR to amplify L1 sequences, we were able to clone the L1s from their fosmid genomic context. Overall, the strategy we

undertook in this study allowed a highly accurate and thorough view of the ability of each L1 to retrotranspose in cultured cells and a careful documentation of the genomic location of each L1 insertion.

We additionally illustrated that many of the 68 full-length polymorphic L1s are highly active in a cultured cell assay for retrotransposition efficiency, and that some elements are quite rare when examined in a diverse panel of individuals. Approximately 55% of the L1s in Chapter 2 retrotransposed at >10% of our reference element, L1.3, and 4 of the 26 genotyped L1s were either African specific (3) or potentially private elements (1) when genotyped on the H952 subset of the human genome diversity panel (HGDP) (Cann et al., 2002; Rosenberg, 2006). Therefore, L1 retrotransposition continues to diversify human genomes, and many full-length L1 polymorphisms are retrotransposition-competent. Indeed, it is interesting to speculate that most humans likely contain different cohorts of L1s. It is also thought provoking to consider the genomic environments that are permissive to L1 expression from their endogenous genomic location, and not solely from a plasmid expression context.

Though it is clear that inter-individual human variation in L1 exists, the rate of L1 retrotransposition and the number of retrotransposition-competent elements in an average individual genome remain poorly understood. In Chapter 1, studies estimating *de novo* retrotransposon rates were discussed, and they range from 1 in every 20 to 1 in every 200 live births for L1 (Ewing and Kazazian, 2010; Huang et al., 2010; Li et al., 2001; Xing et al., 2009). Additional studies have estimated the number of active L1 and Alu elements in an average human genome. These

approximations began with the estimate of ~30-60 active L1s in a diploid human genome (Sassaman et al., 1997). This estimate was revised to ~80-100 active elements per genome with a survey of the activity of 83 of the 90 full-length L1s from the human genome reference sequence (HGR) that contained two intact ORFs (Brouha et al., 2003; Lander et al., 2001). In Chapter 2, we focused on the highly active subset of elements, as these L1s comprise the vast majority of the activity present in an individual genome (Brouha et al., 2003). Though only 6 of the 40 active L1s in the HGR were highly active or "hot" (Brouha et al., 2003), in this study, we determined that the library of an African female, ABC13, contained at least 17 L1 alleles that are potentially highly active (Figure 2.4). This study represents the most comprehensive estimate to date of the highly active L1 alleles present in an individual. Highly active L1s are important, as they demonstrate retrotransposition efficiencies in cell culture similar to those of most progenitors of mutagenic insertions or the full-length disease-causing insertions themselves (Brouha et al., 2003; Moran et al., 1996). Therefore, the activity of L1 in humans was likely underestimated in previous studies of reference genomic sequences, and our study furthers the understanding of retrotransposition activity in extant human genomes (Beck et al., 2010; Beck et al., 2011).

Can Individuals or Populations Vary With Respect to Retrotransposition Rate?

In addition to individual variation in retrotransposon absence vs. presence, it may be possible that individual humans or populations can vary with respect to rate of retrotransposition. An *APOBEC3B* (*A3B*) gene deletion that is present at an allele frequency of ~22% in humans, yet is nearly fixed in Oceanic

populations, has been hypothesized to contribute to individual differences in the ability to repress retrotransposition (Kidd et al., 2007). A3B is expressed in developmentally relevant cell types, where L1 mobilization would need to be restricted in order to prevent accumulation of new insertions in the population. Moreover, the knockdown of A3B increases L1 retrotransposition efficiency ~2-3.7 fold in human embryonic stem cell lines (Wissing et al., 2011). Additionally, two *APOBEC3H* alleles present in many individuals are unable to restrict viruses and transposable elements due to a lack of protein stability, yet an allele of this gene present at high frequency in Africans inhibits L1 retrotransposition (OhAinle et al., 2008). Although they are not correlated with one another, population stratification of the *APOBEC3B* and *3H* active alleles suggests that there may be a higher rate of retrotransposition in some populations or individuals than in others. Other L1 restriction factors likely exist, and may also vary with respect to individuals or populations (Stetson et al., 2008). Interestingly, the high frequency of inactive *APOBEC3* alleles in some populations suggests that these proteins may have deleterious effects on the host genome.

Most of the individuals who have been characterized for genome-wide retrotransposon polymorphisms to date have been from a handful of populations (primarily CEPH Northern European, Yoruban African, Japanese, or Han Chinese) due to the ease of obtaining HapMap DNA samples (Consortium, 2005). Broadening the base of ethnicities for these studies is an important goal in completing the picture of L1 variation in the human population.

Allelic Heterogeneity Affects Individual Variation

In addition to individual and population-level variation with respect to retrotransposon insertions, a given allele of an L1 can also have variable activity. Chapter 2 discusses an element (#5-70) present in the HGR that had a nonsense mutation resulting in premature termination of ORF2p. When the L1 was cloned from the library of ABC12, a Northern European CEPH individual, it contained 2 intact open reading frames and was active in the cultured-cell retrotransposition assay (~8% the efficiency of L1.3). Though allelic heterogeneity in L1 activity has been described previously (Lutz et al., 2003; Seleme et al., 2006), our findings further document the fact that dismissing an L1 as "inactive" after testing one allele may be premature.

Location and Effects of L1 Insertions in Human Genomes

Insertion Sites of De Novo and Inherited L1s and Alus

Though both L1 and Alu insertions likely occur through target-site primed reverse transcription (TPRT), where the degenerate target endonuclease cleavage site is 5'- TTTT/A -3' (the "/" indicating the scissile phosphate) (Dewannieux et al., 2003; Luan et al., 1993), these elements have different patterns of residence in the genome (Smit, 1996). Alu sequences tend to accumulate in GC-rich regions of the genome, while L1s remain relegated to AT-rich regions. The similarity in insertion mechanisms of the two elements contrasts with their eventual genomic locations, and indicates that selective pressures play a large role in the pattern of transposable element insertion sites present in the human genome (Lander et al., 2001; Smit, 1996). Interestingly, L1s in introns appear in the antisense orientation over the sense orientation to the transcript of

the gene in an ~2:1 ratio (Smit et al., 1995). The 2:1 preference for the antisense orientation is even detectable in polymorphic L1s (Chapter 2 Table 2.4). Therefore, selection against certain insertion loci or orientations within genes might occur rapidly at the level of a whole organism.

An L1 insertion site is inherently limited by the preference of the endonuclease domain for certain cleavage sites, but the recognized sequence is short, and degenerate sites are tolerated (Cost and Boeke, 1998; Feng et al., 1996; Morrish et al., 2002). Therefore, L1 insertion sites would be assumed to be fairly random in mammalian genomes, with perhaps a slight preference for AT-rich sequences. However, some studies have proposed that in addition to the changes in element location over time, there may be preferred genomic loci for L1 insertions as well as transposon-intolerant regions of the genome (Levin and Moran, 2011). Though no one has found a true genomic insertion site preference for L1, two L1-mediated insertions of an Alu and an SVA element occurred in the same location in exon 9 of the *BTK* gene, resulting in independent cases of X-linked agammaglobulinemia (Conley et al., 2005). Additionally, an abundance of Ta subfamily L1s on chromosomes 4 and X and the clustering of insertion locations throughout the genome also suggest the existence of potential 'hotspots' for insertions (Beck et al., 2011; Boissinot et al., 2004). Conversely, there are also ~1000 regions of the human genome greater than 10 kb that lack transposon insertions throughout metazoan evolution, most notably near the *Hox* gene clusters (Lander et al., 2001; Simons et al., 2006).

Characterization of both polymorphic and *de novo* insertion sites in humans and in cell culture models is needed to increase our understanding of L1 endonuclease targeting to either random or specific sites of retrotransposition. The ENCODE project has expanded our understanding of non-coding elements in the human genome, and these features have even been cataloged for human embryonic stem cells. This resource may be exploited to investigate the overlap of DNAse hypersensitivity sites, chromatin state, and other features of global genomic architecture that coincide with L1 integration locations (ENCODE Consortium, 2004; Myers et al., 2011). The investigation of epigenetic marks in conjunction with examination of insertion sites in different cell lines could potentially shed light whether chromatin state may affect L1 endonuclease cleavage. Indeed, the use of *Drosophila* modENCODE data (Roy et al., 2010) in conjunction with analysis of ~20,000 insertion locations was recently used to determine that *P* element insertions show a preference for replication origins (Spradling et al., 2011).

Effects of L1 Insertions on Gene Expression

De novo L1 insertions can disrupt genes by interrupting coding sequences, affecting splicing or transcription, acting as substrates for non-allelic homologous recombination, and may attenuate transcription or affect the local epigenetic state (see Chapter 1 for a review) (Beck et al., 2011). Consistent with this observation, the 2 to 1 preference for intronic L1s in the antisense orientation with respect to genes may indicate that L1 insertions are more deleterious to genes when inserted in the sense orientation (Smit et al., 1995). Additionally,

though approximately 30% of L1Hs elements are full-length, successively older L1 subfamilies consist of fewer and fewer 6kb elements. This pattern indicates that the full-length L1 insertions may be more deleterious than truncated L1s (Boissinot et al., 2001; Boissinot et al., 2004). Although L1Hs length and prevalence within introns of genes are similar to those in cell-culture studies (Beck et al., 2010; Gilbert et al., 2005; Iskow et al., 2010; Ovchinnikov et al., 2001), polymorphic elements still display the antisense bias within genes (Table 2.4).

The above data suggest that L1s are subject to selective pressures. Given that full-length L1s are less tolerated in human genomes than 5' truncated elements, it is reasonable to hypothesize that the L1-encoded RNA polymerase II (pol II) promoter or the anti-sense promoter may interfere with the natural expression of genes. This interference could occur through improper expression from either of the L1-encoded promoters, or through the phenomena of “gene breaking”, where an intronic L1 insertion may result in two separate transcriptional units from both forward and antisense promoters (Wheelan et al., 2005). However, in Chapter 3, we showed that the 96 L1s that contain a 524 bp 5'UTR deletion appear to have the same preference for the antisense orientation in introns of genes as L1s that have intact 5'UTRs (Table 3.3) (Smit et al., 1995). The 524 bp deletion disrupts the minimal L1 antisense promoter and the RUNX3 and SRY binding sites of the L1 sense promoter (Athanikar et al., 2004; Speek, 2001; Swergold, 1990; Tchenio et al., 2000; Yang et al., 2003). These data

suggest that the two L1-encoded promoters may not be the primary force of selection against sense-orientation, full-length insertions in introns.

L1s may also contain sense strand RNA polymerase II pause sites or even premature polyadenylation sites that could interfere with gene expression (Chen et al., 2006; Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003). Interestingly, the A-richness of the sense strand of L1 is thought to play a role in these effects on gene expression, an the 524bp deletion in the 5'UTR of #6-113 in Chapter 3 is more G-C rich than the entire L1 (Han et al., 2004). Therefore, the 5'UTR loss in L1 splicing events (Table 3.2) affects promoter activity of the L1 (Figure 3.2) and decreases the length of the element that can act as a substrate for ectopic recombination, but it does not decrease A-rich regions of the L1. This data suggests that spliced L1 insertions may also introduce pol II pause sites or premature polyadenylation sequences within genes (Chen et al., 2006; Han et al., 2004), leading to the 2-fold preference for the antisense orientation when located in introns.

In order to discern if the transcriptional disruption of genes is important to the distribution of L1 in the genome, many insertions in cell culture would have to be recovered, and the distribution of both full-length and truncated L1s examined. If insertions within genes in cultured cells exhibit the same 2-fold preference for the antisense orientation, this would be evidence supporting non-random insertion into introns. Such a scenario could invoke a potential association of the RNA polymerase II transcriptional machinery with nascent L1 insertions, directing the element to the antisense strand of a transcribed gene (Boissinot et al., 2004).

This cultured cell experiment could also allow a more thorough examination of L1 attenuation of transcription or premature polyadenylation (Chen et al., 2006; Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003). Using a cell-culture model will alleviate other genetic factors that could vary between individuals in trios, and will allow the best chance of capturing L1-mediated interference of pol II transcription or RNA processing by full-length or truncated insertions. This experiment would be further facilitated through the use of modern high-throughput sequencing technologies, which have been used previously to determine hundreds of L1 and Alu insertion sites in human genomes (Beck et al., 2011).

L1 Poly(A) Tails as Substrates for Microsatellite Repeats

Poly(A) tails of Alu insertions are substrates for microsatellite formation and vary in length with time spent resident in the genome (Arcot et al., 1995). The poly(A) tails of L1 sequences are generally shorter with increasing age of the element, and both disease-causing *de novo* insertions and cell culture insertions generally have much longer A tails than polymorphic L1s (Gilbert et al., 2002; Ovchinnikov et al., 2001). Therefore the poly(A) tail lengths of L1s are likely variable between individuals, may be highly common polymorphisms, and are also likely substrates for microsatellite formation (Ovchinnikov et al., 2001). Discerning the poly(A) length of an L1 in multiple genomes is difficult due to the repetitive nature of L1 elements and the challenges in accurately characterizing a homopolymeric nucleotide stretch in the genomes of many individuals. In particular, some approaches for characterizing polymorphic L1s use PCR, which

can be error prone, or are directly constrained with regard to accurate sequencing of homopolymeric stretches of DNA (Wheeler et al., 2008).

One way to investigate the length polymorphism of L1 poly(A) tails would be through the examination of different alleles of the same element using standard capillary sequencing. With the fosmid libraries, sequencing of the different alleles can be accomplished with high fidelity, allowing determination of changes in the L1 that may contribute to allelic heterogeneity in activity and characterization of differences in poly(A) tail length (see Figure 2.4 and Table 2.5 for *in silico* genotyping of the L1s in 9 individuals) (Kidd et al., 2008). An additional method to approach both poly(A) length polymorphism and the genesis of microsatellite repeats associated with L1 would be through the examination of poly(A) sequences from older L1s in both our genomes and the orthologous loci in chimpanzees. A similar method was used to identify Alu-associated microsatellite repeats (Arcot et al., 1995). The use of both methods would give us both an estimation of the rapidity with which the L1 poly(A) tail degenerated (through examination of human alleles), and of the prevalence of changes when measured over millions of years.

Sequencing Elucidates LINE-1 Biology

Transduction Subfamilies in Active L1 Identification and Subfamily Succession

Transductions are generated when L1s read through their endogenous poly(A) tails and transfer 3' flanking genomic DNA to new insertion locations. This transfer tags the new insertion in a consistent manner, where one target-site duplication (TSD) directly flanks the 5'UTR, but the poly(A) tail following the L1 is

flanked by the transduced sequence and a second poly(A) tail that precedes the 3' TSD (Goodier et al., 2000; Holmes et al., 1994; Moran et al., 1999; Pickeral et al., 2000). These sequences are prevalent in human genomes, with ~21% of full-length L1s in the HGR containing 3' transductions (Lander et al., 2001).

In addition to 3' transductions being a common feature of many genomic L1 sequences, they are also present in a large number of polymorphic L1s (Beck et al., 2010; Kidd et al., 2010). Indeed, some of the 3' transductions have been repeated in numerous genomic L1s, generating transduction “subfamilies”, and some of the transductions appear to tag L1s that are currently amplifying in humans (Beck et al., 2010; Kidd et al., 2010). Moreover, the percent of highly active L1s that contain transductions in Chapter 2 (17/37 “hot” L1s or ~46%) vs. the percent of full-length L1s with transductions in the HGR (~21%) additionally suggests that these sequences may tag highly active L1s. In support of this data, many highly active, full-length, mutagenic L1 insertions contain 3' transductions (Brouha et al., 2002; Holmes et al., 1994; Kimberland et al., 1999).

In Chapter 3, I discussed a transduction-specific mapping technique (TS-ATLAS) that was used to identify an additional four elements from the same subfamily as $L1_{RP}$. This method identified an element with the genomic structure consistent with being a progenitor of the $L1_{RP}$ family. The L1 had a 3' poly(A) tail directly flanked by the 3' TSD, and the TSD included the $L1_{RP}$ transduction sequence. The putative $L1_{RP}$ progenitor element was highly active in a cultured cell retrotransposition assay, providing a proof of principle that TS-ATLAS is effective at identifying highly active transduction alleles and progenitors.

The examination of multiple human genomes using additional transduction sequences for the TS-ATLAS protocol may identify further highly active L1s. In addition to the currently amplifying L1 transcribed subset A (L1 Ta) subfamily of elements, older pre-Ta L1s are also polymorphic in the human population (Skowronski et al., 1988). In the 68 L1s examined in Chapter 2, six elements contained sequence characteristics consistent with the pre-Ta subfamily. Three of these elements (#3-5, 3-24, and 5-55) had the same 3' transduction that distinguished their poly(A) tails from other L1s (Figure 2.5). Interestingly, two of the three were highly active in a cultured-cell assay and two were African-specific when examined in the HGDP (Cann et al., 2002) (Chapter 2). One of the genotyped pre-Ta L1s (#3-24) was absent from the HGDP and contained an inherited serine 228 to proline mutation in the EN domain of ORF2p responsible for its inactivity in cell culture (Chapters 2 and 3). Thus, the examination of additional transduction subfamilies may help elucidate the prevalence of active pre-Ta L1s, and may be used to identify highly active, potentially rare L1 lineages in humans that are also informative to L1 evolution and biology.

Related L1 sequences within transduction families may also be exploited to determine amino acid or nucleotide changes that modulate the activity of L1. Comparing sequences within transduction families can pinpoint changes responsible for more subtle differences in retrotransposition competence, as opposed to the differences between highly active and inactive L1s, and may lead to insights into the amino acid changes that make some L1s more highly active than others. A transduction cluster (roman numeral IV) that contains elements

#4-27 and 4-35 is closely related in sequence to #4-8, and all three of these L1s are highly active (Figure 2.5). However, #4-27 retrotransposes at ~79% the efficiency of L1.3, whereas #4-35 and 4-8 both retrotranspose at >160% of L1.3 (Table 2.2). An examination of the three sequences shows that there is one ORF1p change (R130Q) and two ORF2p changes (T316I and I515T) in #4-27 when compared to #4-35 and 4-8. Of the three amino acid residues with changes, none are conserved past L1PA5 in the consensus alignments (Figure 3.4). Additionally, the 5' and 3' UTR sequences of the three L1s lack insertion or deletion sequences with respect to one another. Therefore, a nucleotide change in the UTRs of L1 #4-27 or a change in an amino acid residue that is not well conserved (*e.g.,* R130Q in ORF1p) may have modulated the retrotransposition activity of this element. Further examination of transduction families, especially with the identification and characterization of additional family members, may be able to pinpoint the causes of these more subtle L1 activity differences.

In addition to the ability of L1 3' transductions to tag highly active elements, in Chapter 2 we used related L1 sequences to inform the model of L1 subfamily succession. Others had proposed that a very limited number of L1 “master” genes may give rise to many progeny elements, and in this way promote the success of L1s with specific sequence features (Deininger et al., 1992). This succession pattern would propagate a single lineage of L1s, as has existed for the last ~40 million years of primate genome evolution (Khan et al., 2006). In Chapter 2, the presence of multiple active elements within 6 transduction families gave rise to a modified hypothesis for subfamily succession, where numerous

progenitors may exist at one time, including active L1s from both Ta and pre-Ta subfamilies (Figures 2.5 and 2.6) (Cordaux et al., 2004). Moreover, we suggest that L1 progenitor elements would have "life-spans" limited by mutation, and would have to generate retrotransposition-competent progeny in permissive genomic contexts prior to their inactivation by cellular processes to continue the specific L1 lineage.

Splicing of L1 Transcripts

Splice sites within the L1 sequence have been characterized previously (Belancio et al., 2006; Belancio et al., 2008). In Chapter 3 I document the widespread presence of a novel 5'UTR deletion present in ~100 copies in the HGR (Table 3.2). This deletion was first seen in element #6-113, and is presumably generated through splicing (Figure 3.2). The presence and ubiquity of this 5'UTR deletion suggests that other L1 splicing events may have populated the genome. Many additional splice donor/acceptor combinations can be predicted through the examination of Table 3.1. Though the splicing event shown in Figure 3.2 is common, it remains to be seen whether this feature is a *de novo* event, or if a spliced L1 in a permissive expression context can give rise to new retrotransposon insertions. This possibility is suggested through the retrotransposition efficiency of #6-113 (~20% of L1.3- Figure 3.2) with an exogenous CMV promoter upstream. Overall, these data suggest that both spliced and unspliced L1 transcripts can undergo retrotransposition (Garcia-Perez et al., 2007).

The Role of Non-Conserved Domains in L1 Retrotransposition

In Chapter 3, we developed resources to identify amino acid changes in L1s with two intact ORFs that lacked high levels of activity in the cultured cell assay. This natural mutagenesis experiment identified putative sequence changes responsible for the low retrotransposition efficiency of 16 L1s from Chapter 2. Additionally, we identified the splicing mutation described above in one L1 and an endonuclease domain mutation (S228P in ORF2p) in element #3-24. Identification of causative amino acid mutations in ORF1p and ORF2p of the 14 remaining L1s will require further study, but the alignments and analysis developed in Chapter 3 will be invaluable. To determine causative amino acid changes in the 14 remaining L1s, I propose exchanging ORF1p and ORF2p from the retrotransposition defective elements with the protein coding regions of a known, highly active element (L1.3 (Dombroski et al., 1993)). This domain swap would create chimeric L1s that could pinpoint whether mutations in ORF1p, ORF2p, or both of the L1-encoded proteins were responsible for the inability of an L1 to retrotranspose with high efficiency in HeLa cell culture (Figure 4.1).

It is also important to note that the data set being explored in this study provides us with an unbiased view of mutations throughout the body of the L1. When paired with information from both the literature and our lab, these mutations may provide important data on poorly understood regions of ORF1p and ORF2p. Mutations in the amino terminus of ORF1p and the carboxyl terminus of ORF2p were present in some of the inactive elements, and these domains may prove interesting in spite of (or perhaps because of) their low

evolutionary conservation (Table 3.3). In fact, the amino terminus of ORF1p is a potential site for host-L1 interactions, as it has been found to be under positive selection in primate L1 lineages (Boissinot and Furano, 2001; Khan et al., 2006). Interestingly, recently published immunofluorescence and western blotting experiments (Doucet et al., 2010) have shown that mutations in the cysteine-rich domain affect the amount of ORF2p localization to L1 RNPs. The carboxyl-terminus of ORF2p should be explored to determine if other mutations in this region interfere with the localization of ORF2p to RNPs. Such studies would elucidate the region of ORF2p that is potentially involved in efficient binding to the L1 RNA or in the stability of the protein itself. Interestingly, mutations in the cysteine-rich domain of ORF2p render the L1 unable to retrotranspose in HeLa cells, but allow mobilization in Chinese Hamster Ovary (CHO) cells deficient for nonhomologous end joining (unpublished data, thesis of Tammy Morrish). Therefore, the cysteine-rich domain or the carboxyl-terminus of ORF2p may participate in folding ORF2p to allow the endonuclease to function. Such a scenario would additionally lead to the idea that the global conformation of ORF2p is involved in protein stability, as indicated by reduced ORF2p in RNPs derived from L1 cysteine-rich domain mutants. Alternately, the cysteine-rich domain may be involved in an upstream step prior to the endonuclease cleavage of DNA, such as binding to potential genomic DNA target sites or even altering local chromatin states. The mutations identified in Chapter 3 expand the scope of amino acid changes throughout the L1-encoded proteins that can be similarly explored through functional studies and analyses.

Altering and/or Deletion of the 3'UTR

Alterations of the L1 3'UTR are consistent with retrotransposition. Deletion of the 3'UTR or insertion of the neomycin retrotransposition indicator cassette within the region does not disrupt an element's ability to retrotranspose (Moran et al., 1996). In Chapter 3, Table 3.2 details splicing mutations in the 5'UTR, and also depicts the presence of a premature poly(A) tail, immediately downstream of the poly purine tract in an L1 on chromosome 3. I identified 5 additional elements with similar poly(A) sites in BLAT searches (data not shown). Moreover, LEAP results indicate that the reverse transcription of L1 RNA can occur at sites internal to the L1 (Kulpa and Moran, 2006). However, though the 3'UTR may not be strictly necessary for retrotransposition of L1, it may be involved in other aspects of L1 mobility or regulation that are not readily assayed in HeLa cells.

Although 3'UTR deletions and alterations are tolerated in retrotransposition, the 3'UTR may be important in competition for the L1-encoded ORF2 protein. Therefore, conserved RNA changes present in human L1s may be advantageous. Indeed, though older L1s shared between human and primate genomes contain a 3'UTR GAG trinucleotide at base pairs 5929-5931, L1 Ta elements contain an ACA at this location (positions relative to L1.3) (Scott et al., 1987; Skowronski et al., 1988). To address the importance of this difference, L1.3 with a GAG trinucleotide in the 3'UTR was cloned and tested with respect to the activity of the wild type L1.3 ACA. This experiment showed that a conserved RNA change in humans (exhibited by L1.3 ACA) is potentially dispensable for

retrotransposition (Chapter 3). Equivalent retrotransposition efficiencies of the two variations of the 3'UTR indicate that the ACA difference is potentially a passenger mutation that was carried along with other beneficial changes in the human specific L1 sequence. However, the lack of a difference in the efficiency of retrotransposition in *cis* does not preclude a competitive advantage conferred by the 3'UTR ACA character in Ta subfamily elements. This advantage could be elucidated through examination of the relative efficiencies of GAG and ACA-containing alleles at competing for host proteins. In order to study this dynamic, an experiment would need to be conducted testing the retrotransposition of one allele in the presence of the other element. One experimental methodology would be to co-transfect an *mneoI* tagged ACA L1.3 with a no-neo version of GAG L1.3, and then switch the tag to the other L1 for a converse retrotransposition assay. The efficiencies of the different tagged L1 constructs should be the same if the two alleles are similar in their ability to co-opt host proteins for retrotransposition. If one allele of L1.3 has significantly higher retrotransposition efficiency than the other in this experiment, it would imply that the 3'UTR difference present in the more “active” allele imparted a competitive advantage to the L1 (Figure 4.2). Another way to analyze the importance of 3'UTR changes would be to make use of fluorescent primer-extension constructs of JM101 L1.3 ACA (developed previously in the Moran Lab) and L1.3 GAG to discern the ratio of integrants (Garcia-Perez et al., 2007).

Summary

In this thesis, I addressed the role of full-length L1s in inter-individual human variation. A majority of these elements are retrotransposition competent, and therefore may affect the genomes of future human generations through their mobility. In addition, highly active L1s were shown to be more prevalent than previously appreciated, and some of these elements were rare when genotyped in diverse individuals. The examination of the sequences of these polymorphic L1s has led to an expanded understanding of L1 subfamily succession and the cohort of amino acid or nucleotide changes that can affect retrotransposition. Overall, these results have illustrated the importance of L1 in genomic variation, and to the continuing evolution of the human genome. Clearly, advances in high-throughput sequencing technologies will further expand upon the results presented here, and will likely be used to address some of the currently open questions in L1 biology.

Acknowledgements

I thank members of the Moran Lab and my thesis committee for helpful discussions regarding experiments presented in this concluding chapter.

Figure 4.1: Domain Swap Constructs to Determine Location of Deleterious Amino Acid Changes

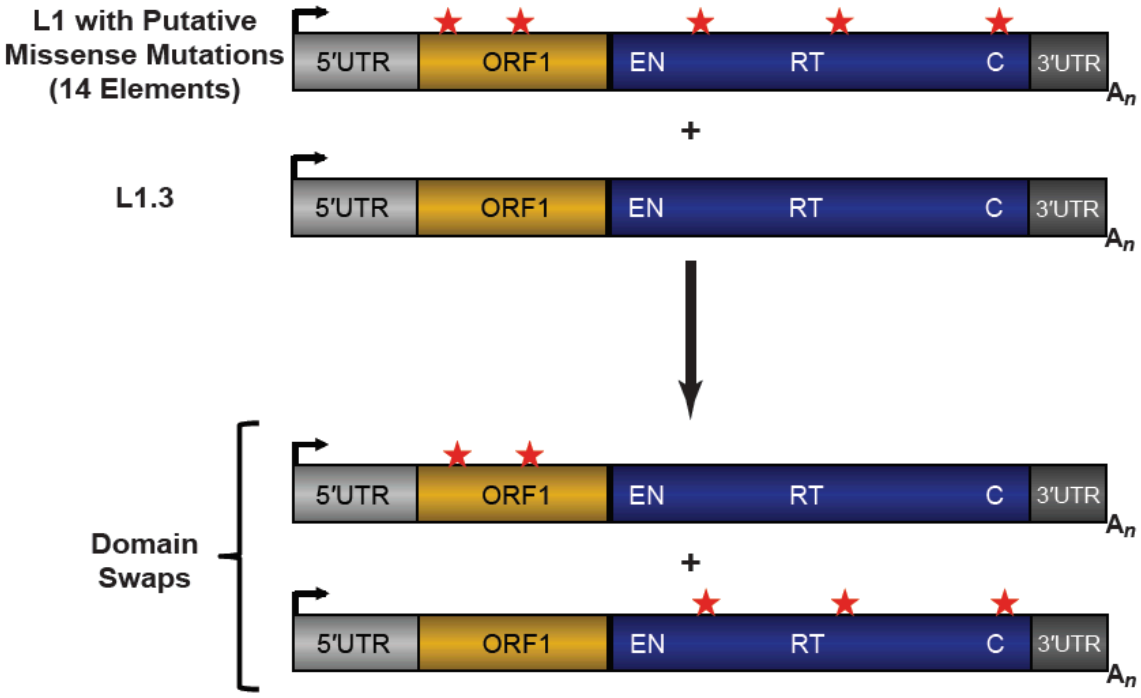


Figure 4.1: Domain Swap Constructs to Determine Location of Deleterious Amino Acid Changes

The Figure depicts the construction of chimeric L1 elements from the 14 putative missense mutation-containing elements in Chapter 3 and the “hot” reference element, L1.3. In this scheme, elements that are inactive in the retrotransposition assay that contain amino acid changes in both ORF1p and ORF2p (red stars) will be digested, and the 5'UTR through ORF1p fragment will be exchanged with that region of L1.3, and vice-versa. This generates two new L1 domain swap constructs (bottom of figure) that can be tested in the retrotransposition assay. If one of these chimeric L1s retrotransposes with high efficiency in HeLa cells, it indicates that the other ORF contains a putative missense mutation.

Figure 4.2: A LINE-1 Competition Experiment

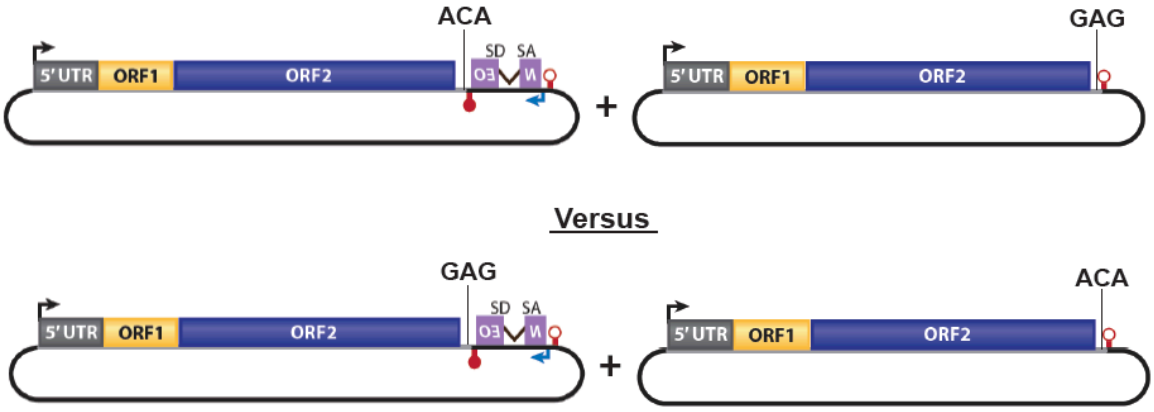


Figure 4.2: A LINE-1 Competition Experiment

Though L1.3 ACA and GAG alleles retrotranspose with similar efficiencies in HeLa cells, it may be possible that the 3'UTR change present in human L1s (ACA) can confer a competitive advantage to the element when tested against a GAG-containing allele. Testing one allele (ACA at top) for efficiency when transfected with an equal amount of a GAG allele without an indicator cassette (0.5 µg each when transfecting a well of a 6-well plate) will allow comparison with the opposite competitive transfection, as shown in the bottom of the Figure. If one of the *mneoI* containing constructs retrotransposes with a higher efficiency in this competition assay, it would imply that the 3'UTR conferred an advantage to the element when transcribed at the same time as a different (*i.e.* older) L1.

References

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. Genomics *29*, 136-144.

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 Res *32*, 3846-3855.

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

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. Annu Rev Genomics Hum Genet *12*, 187-215.

Belancio, V.P., Hedges, D.J., and Deininger, P. (2006). LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res *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.

Boissinot, S., Entezam, A., and Furano, A.V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol *18*, 926-935.

Boissinot, S., and Furano, A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol *18*, 2186-2194.

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

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. Am J Hum Genet *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. Proc Natl Acad Sci U S A *100*, 5280-5285.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A.*, et al.* (2002). A human genome diversity cell line panel. Science *296*, 261-262.

Chen, J., Rattner, A., and Nathans, J. (2006). 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. Hum Mol Genet *15*, 2146-2156.

Conley, M.E., Partain, J.D., Norland, S.M., Shurtleff, S.A., and Kazazian, H.H., Jr. (2005). Two independent retrotransposon insertions at the same site within the coding region of BTK. Hum Mutat *25*, 324-325.

Consortium, I.H. (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Cordaux, R., Hedges, D.J., and Batzer, M.A. (2004). Retrotransposition of Alu elements: how many sources? Trends Genet *20*, 464-467.

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.

Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd, and Edgell, M.H. (1992). Master genes in mammalian repetitive DNA amplification. Trends Genet *8*, 307-311.

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

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci U S A *90*, 6513-6517.

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 Genet *6,* e1001150.

ENCODE Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

Ewing, A.D., and Kazazian, H.H. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

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.

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

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of l1 retrotransposition intermediates in cultured human cells. Mol Cell Biol *25*, 7780-7795.

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

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

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.

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. Nat Genet *7*, 143-148.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

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.

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.

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 Res *16*, 78-87.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F.*, et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

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.

Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R., and Eichler, E.E. (2007). Population stratification of a common APOBEC gene deletion polymorphism. PLoS Genet *3*, e63.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet *8*, 1557-1560.

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *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.

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

Li, X., Scaringe, W.A., Hill, K.A., Roberts, S., Mengos, A., Careri, D., Pinto, M.T., Kasper, C.K., and Sommer, S.S. (2001). Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat *17*, 511-519.

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.

Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A., and Moran, J.V. (2003). Allelic heterogeneity in LINE-1 retrotransposition activity. Am J Hum Genet *73*, 1431-1437.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K.*, et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

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. Nat Genet *31*, 159-165.

Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., Birney, E., Wold, B.*, et al.* (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol *9*, e1001046.

OhAinle, M., Kerns, J.A., Li, M.M., Malik, H.S., and Emerman, M. (2008). Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. Cell Host Microbe *4*, 249-259.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res *11*, 2050-2058.

Perepelitsa-Belancio, V., and Deininger, P. (2003). RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet *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 Res *10*, 411-415.

Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet *70*, 841-847.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F*., et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.

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. Nat Genet *16*, 37-43.

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.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A *103*, 6611-6616.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., and Swergold, G.D. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res *10*, 1496-1508.

Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. (2006). Transposon-free regions in mammalian genomes. Genome Res *16*, 164-172.

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

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. Curr Opin Genet Dev *6*, 743-748.

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

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

Spradling, A.C., Bellen, H.J., and Hoskins, R.A. (2011). Drosophila P elements preferentially transpose to replication origins. Proc Natl Acad Sci U S A *108*, 15948-15953.

Stetson, D.B., Ko, J.S., Heidmann, T., and Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. Cell *134*, 587-598.

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

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

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res *15*, 1073-1078.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T.*, et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872-876.

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. J Biol Chem.

Witherspoon, D.J., Xing, J., Zhang, Y., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2010). Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics *11*, 410.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res *19*, 1516-1526.

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 Res *31*, 4929-4940.