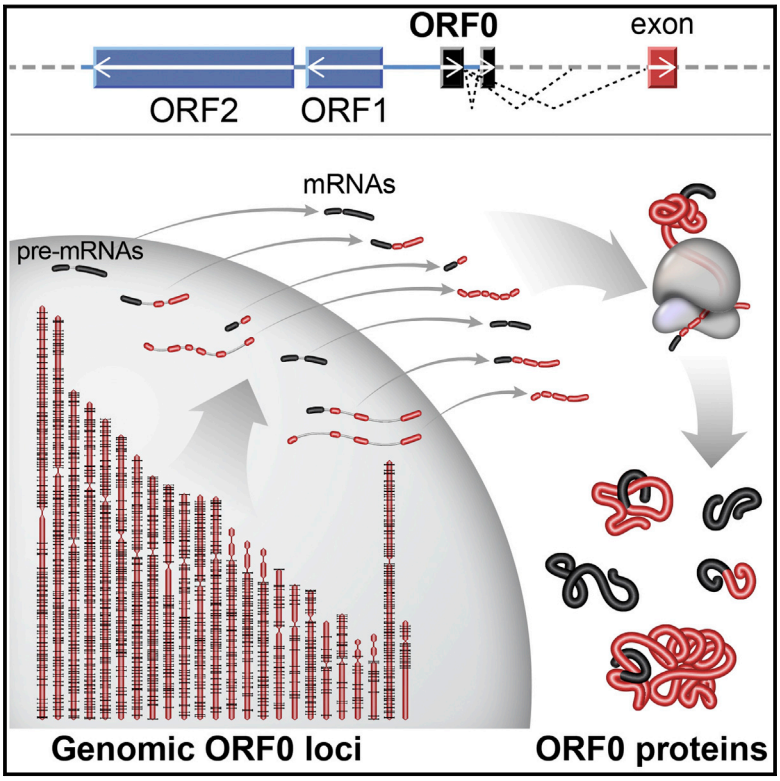
Article

Cell

Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity

Graphical Abstract



Authors

Ahmet M. Denli, Iñigo Narvaiza, Bilal E. Kerman, ..., Tony Hunter, Alan Saghatelian, Fred H. Gage

Correspondence

gage@salk.edu

In Brief

A primate-specific open reading frame, ORF0, is found in the LINE-1 retrotransposons. It not only enhances LINE-1 mobility but also leads to the generation of ORF0-proximal exon fusion proteins, contributing to retrotransposon-mediated diversity.

Highlights

- ORF0 is a primate-specific open reading frame in LINE-1 retrotransposons
- Over 3,000 potential ORF0 loci exist in human and chimpanzee genomes
- ORF0-proximal exon fusion proteins are generated through splicing
- ORF0 influences LINE-1 mobility

Denli et al., 2015, Cell *163*, 583–593
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.025

CellPress

Article

Cell

Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity

Ahmet M. Denli,[1] Iñigo Narvaiza,[1] Bilal E. Kerman,[1] Monique Pena,[1] Christopher Benner,[1] Maria C.N. Marchetto,[1] Jolene K. Diedrich,[5] Aaron Aslanian,[2] Jiao Ma,[3,6] James J. Moresco,[5] Lynne Moore,[1] Tony Hunter,[2,4] Alan Saghatelian,[3] and Fred H. Gage[1,7,8,*]

[1]Laboratory of Genetics
[2]Molecular and Cell Biology Laboratory
[3]Clayton Foundation Laboratories for Peptide Biology
[4]Cancer Center
[5]Mass Spectrometry Center
The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA
[6]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA
[7]Center for Academic Research and Training in Anthropogeny (CARTA)
[8]Kavli Institute for Brain and Mind (KIBM)
University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
*Correspondence: gage@salk.edu
http://dx.doi.org/10.1016/j.cell.2015.09.025

SUMMARY

LINE-1 retrotransposons are fast-evolving mobile genetic entities that play roles in gene regulation, pathological conditions, and evolution. Here, we show that the primate LINE-1 5′UTR contains a primate-specific open reading frame (ORF) in the antisense orientation that we named ORF0. The gene product of this ORF localizes to promyelocytic leukemia-adjacent nuclear bodies. ORF0 is present in more than 3,000 loci across human and chimpanzee genomes and has a promoter and a conserved strong Kozak sequence that supports translation. By virtue of containing two splice donor sites, ORF0 can also form fusion proteins with proximal exons. ORF0 transcripts are readily detected in induced pluripotent stem (iPS) cells from both primate species. Capped and polyadenylated ORF0 mRNAs are present in the cytoplasm, and endogenous ORF0 peptides are identified upon proteomic analysis. Finally, ORF0 enhances LINE-1 mobility. Taken together, these results suggest a role for ORF0 in retrotransposon-mediated diversity.

INTRODUCTION

Transposable elements (TEs) are mobile genetic elements that can alter their chromosomal locations in the host genomes. TEs, first discovered by Barbara McClintock in maize (McClintock, 1950), are abundantly present in nearly all genomes studied to date; they influence gene expression and shape the genomes over evolutionary time (Huang et al., 2012). There are two classes of TEs based on their transposition mechanisms: DNA transposons and retrotransposons. DNA transposons mobilize with a cut-and-paste mechanism, whereas retrotransposons move by copy-and-paste via an RNA intermediate (Kleckner, 1990; Luan et al., 1993). Autonomous elements from both classes are defined as TEs that encode the proteins required for transposition, whereas non-autonomous elements depend on such proteins to be provided in *trans*. In primate genomes, most active TEs belong to the retrotransposon families. Of these, LINE-1 (L1) elements are the only autonomous elements that are currently active (Dewannieux et al., 2003; Hancks et al., 2011) and thus have directly and indirectly contributed to ~30% of the human genome (Lander et al., 2001). At present, the majority of L1 elements are inactive, due to accumulated mutations as well as 5′ truncations that are common during the integration process, thus reducing the number of estimated active elements to ~80 per genome (Brouha et al., 2003). The first active L1 element was isolated through analysis of mutagenic L1 insertions into the factor VIII gene in hemophilia A patients (Dombroski et al., 1991). Since then, retrotransposon germline insertions have been linked to ~100 human diseases (Hancks and Kazazian, 2012).

Intact, active L1s are ~6 kb long and contain a 5′UTR, two open reading frames (ORF1 and ORF2) and a short 3′UTR (Scott et al., 1987). The L1 5′UTR has promoter activity in both the sense and antisense (ASP) directions (Speek, 2001; Swergold, 1990). ORF1 encodes an ~40 kDa RNA-binding protein that is required for L1 transposition (Kolosha and Martin, 1997; Moran et al., 1996). However, ORF1 does not have any significant sequence similarity to known proteins (Goodier et al., 2007). ORF2 is a large protein at ~150 kDa with endonuclease and reverse transcriptase activities (Mathias et al., 1991). These activities, as well as the function of a cysteine-rich region at the C terminus, are important for L1 mobility (Feng et al., 1996; Moran et al., 1996).

Regardless of their ability to mobilize, L1s contribute to transcriptome diversity and gene regulation (Cordaux and Batzer, 2009). Transcription initiated in both directions can extend beyond the L1 sequence but, due to the presence of a polyA signal at the end of the 3′UTR, most sense transcripts end within

the element. However, extensions into the genomic flank are also frequently observed and can lead to 3′ transductions (Moran et al., 1999). Analyses of cloned cDNAs provide evidence of antisense transcripts that are spliced into exons in the neighboring genomic sequences (Macia et al., 2011; Mätlik et al., 2006; Wheelan et al., 2005). Recent studies have focused on specific examples of spliced transcripts with a focus on disease, and a number of L1-driven transcripts have been shown to exist in cancer cells (Cruickshanks and Tufarelli, 2009). In addition to driving genes, antisense transcripts have been linked to chromatin modifications that influence gene expression (Cruickshanks et al., 2013).

A recent analysis of L1s in primates showed that, while ORF1 and ORF2 sequences have been relatively well conserved, acquisition of new 5′UTRs frequently occurred during primate evolution, providing the diversity that resulted in selection of the current 5′UTR (Khan et al., 2006). With the above in mind, we set out to improve our understanding of the properties of the primate L1 5′UTR. Here, we show that the currently active primate L1 5′UTR has well-conserved properties that support translation of an ORF that we have named ORF0. ORF0 is encoded by a primate-specific antisense ORF that lies downstream from the ASP and has a strong, well-conserved Kozak sequence. The gene product of this ORF is predominantly nuclear and localizes to promyelocytic leukemia (PML)-adjacent bodies. ORF0 also has two prominent splice donor (SD) sites at nucleotides 106 and 191 (amino acids 35 and 64) that can act in concert with splice acceptors (SAs) in downstream genomic sequences to generate fusion proteins. ORF0 mRNAs are capped, polyadenylated, associated with ribosomes, and upon immunoaffinity purification, peptides from endogenous ORF0 products can be detected by mass spectrometry. Lastly, overexpression of ORF0 leads to a modest but significant increase in L1 mobility. Thus, we have identified and begun to characterize a third ORF from primate L1 retrotransposons.

RESULTS

Identification of an ORF in the Human Antisense L1 5′UTR

We started by analyzing the antisense 5′UTR for the presence of ORFs that have an upstream promoter, start with ATG, and have a strong Kozak sequence determined by the presence of A/G in position −3 and G at position +4 (Kozak, 1987). Only one potential ORF exists that meets these criteria and, due to its 5′ position with respect to ORF1 and ORF2, we have called it ORF0. ORF0 lies between nucleotides 452–236 from the 5′ end of LINE-1 in the antisense orientation and contains two SD sites (red boxes) within the potential coding sequence (Figure 1A). There are ~781 loci that could encode full-length (FL) ORF0 in the human genome; the consensus sequence for the FL ORF0 protein obtained from these loci is shown in Figure 1A. The chimp ORF0 consensus sequence from ~395 FL ORF0 loci is identical to that of the human.

The previously mapped L1 ASP lies upstream of ORF0, with some overlap (Speek, 2001). This overlap prompted us to check whether the promoter activity resided upstream of the initiator methionine (1st Met) of ORF0. Results from luciferase reporter assays suggested that promoter activity was upstream but not downstream of ORF0 1st Met, and we further mapped a minimal ORF0 promoter of ~150 bp that had similar activity to the previously described L1 ASP (Figure 1B). We also cloned a number of polymorphic ORF0 promoters upstream of luciferase and GFP reporters. While variable, all the tested promoters were active (data not shown). This finding is consistent with previous observations that a high percentage of L1 5′UTRs have antisense promoter activity (Macia et al., 2011). Next, in vitro translation of HA-tagged ORF0 was tested in rabbit reticulocyte lysates and confirmed with western blot analysis (Figure 1C).

To investigate whether this potential ORF could be translated in human cells, we removed the stop codon of ORF0 and cloned it upstream of a promoterless, in-frame GFP coding sequence that lacked the first ATG. Upon transfection, western blot analysis showed that, indeed, the ORF0 promoter and the context around the 1st Met of ORF0 were sufficient to translate the ORF0-GFP fusion protein (Figure 1D).

ORF0 Protein Is Predominantly Nuclear and Present in PML-Adjacent Foci

To analyze the subcellular localization of ORF0, we generated a GFP-tagged ORF0 clone in an L1 context (GFP-ORF0-L1). Since two SD sequences that were often involved in generation of spliced antisense transcripts (Speek, 2001) fell within ORF0, to allow detection of both spliced and unspliced products, GFP was placed at the N terminus but downstream of the Kozak context of ORF0 to minimize any effects on translation initiation (Figure 1E). Western blot analysis confirmed that GFP-ORF0 fusion protein was generated (Figure 1E). Importantly, when the 1st Met of ORF0 was mutated to threonine (M1T), we observed that GFP signal was lost, showing that translation started from the 1st Met of ORF0 (Figure 1F) and ruling out any potential upstream translation initiation. Furthermore, addition of a poly A signal downstream of ORF0 at the end of the L1 did not change protein localization, suggesting that the produced ORF was contained within the L1 and was not a splicing product with the downstream flank (Figure S1A). We also fused ORF0 to mCherry (29% identity to EGFP) and observed a very similar pattern, suggesting that the sequence of the tag was not driving the localization (Figures S1B and S1C). Interestingly, ORF0, but not GFP-alone from the same plasmid backbone, was localized predominantly in nuclear foci in the majority of cells (Figures 1F and S1D–S1F). As predicted by the charge distribution of amino acid residues, the C terminus portion of ORF0 was required for nuclear localization (Figure S1G). Since a number of ORF0 variants may be encoded due to polymorphisms in L1 sequences, we cloned some of these variants and observed that, unless truncated, most localized similarly (data not shown).

Based on the numbers and distribution of foci, we hypothesized that ORF0 localization could be related to PML bodies. PML bodies are nuclear proteinaceous structures often associated with the nuclear matrix and are involved in a wide variety of processes that may influence L1 biology: stress, anti-viral and DNA damage response, transcriptional regulation, heterochromatin, and post-translational protein modifications (Bernardi and Pandolfi, 2007). Indeed, in cells transfected with PML-IV-GFP and mCherry-ORF0, high-magnification imaging

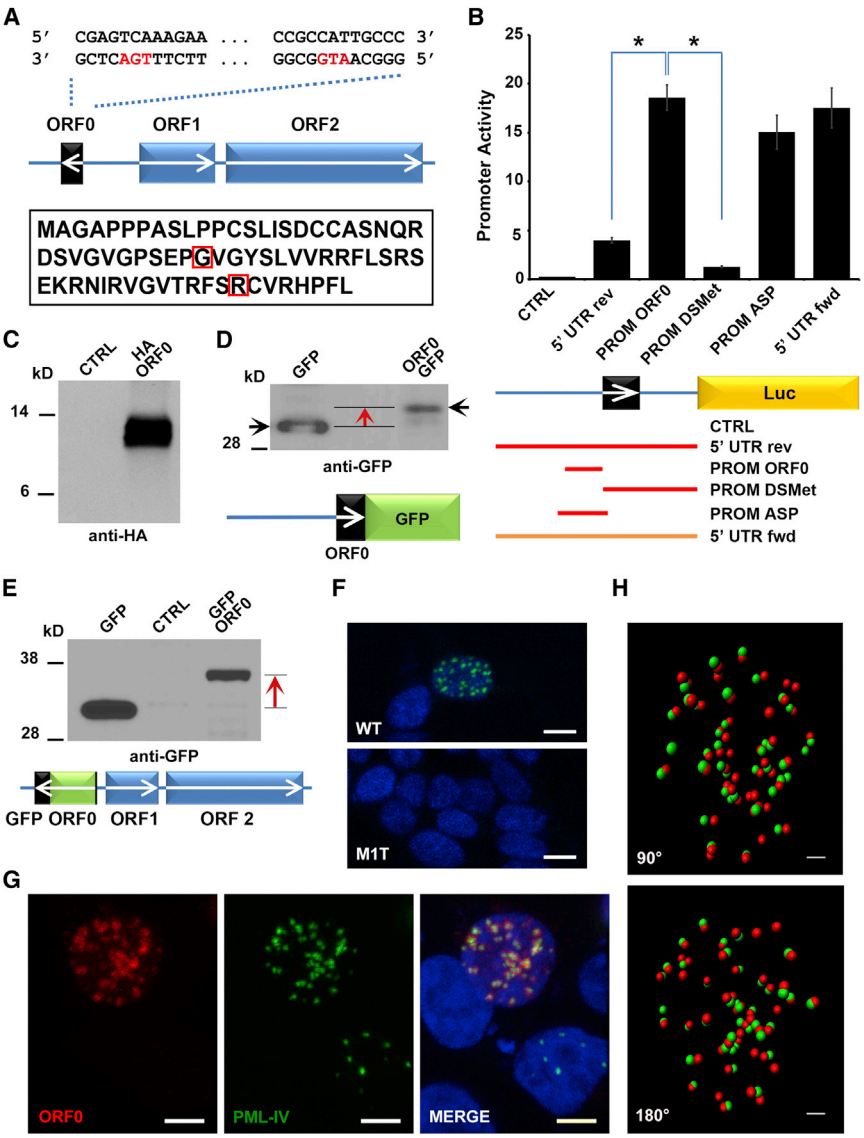


Figure 1. Identification of ORF0 in L1 5′UTR

(A) Location of ORF0 in L1. The start codon ATG and the stop codon TGA are labeled red in the antisense orientation. The positions of splice donor sites within the coding sequence are indicated with red squares. Consensus protein sequence of full-length ORF0 based on ~781 potential ORF0 loci in the human genome.

(B) Upstream ~150 bp region of ORF0 has promoter activity. Luciferase assays were performed to determine promoter activity of the L1 5′UTR regions shown in the panel below the graph. Red and orange lines represent antisense and sense strands, respectively. DSMet refers to downstream of initiator methionine. Data are presented as mean ± SEM. *Denotes $p < 0.05$ significance between indicated groups using t test. CTRL denotes control.

(C) ORF0 can be translated in vitro. HA-tagged ORF0 production was monitored by western blotting.

(D) Production of ORF0-GFP fusion protein was detected by GFP western blot. The C-terminal GFP tagged ORF0 construct driven by the upstream region of ORF0 is shown at the bottom. Black arrows indicate GFP alone and the fusion protein. Red arrow highlights the size shift.

(E) GFP-ORF0 fusion protein was detected by western blot. Design of GFP-ORF0 construct in L1 context. GFP is cloned at the N terminus of ORF0 downstream of the 1st Met and potential Kozak context. Red arrow highlights the size shift in the generated protein.

(F) Translation of GFP-ORF0 is dependent on the ORF0 initiator methionine. Fluorescent detection of ORF0 localization upon transfection of the construct depicted in E into HEK293T cells. WT, wild-type; M1T, initiator methionine to threonine mutant. Scale bar, 10 μm.

(G) Most of ORF0 protein localizes to PML-adjacent nuclear bodies. Confocal imaging of cells transfected with mCherry-ORF0- and GFP-PML-IV-encoding plasmids. Scale bar, 4 μm.

(H) Spot representation of ORF0 (red) and PML (green) foci. Images from 90° and 180° relative to Movie S1 are shown. Scale bar, 1 μm.

See also Figure S1.

showed that ORF0 was present in PML-adjacent foci (Figure 1G). Spot analysis of confocal z series confirmed this observation (Figure 1H; Movie S1).

A Large Number of ORF0 Loci with a Conserved Functional Kozak Context Exist in Primate Genomes

We sought to determine how many loci could potentially encode ORF0 in the human and chimp genomes. Taking splicing into consideration, we scanned these genomes for potential ORF0 loci that are untruncated up to the two commonly used SD sites and have an adjacent GT dinucleotide. Human and chimp genomes have ~3,528 and ~3,299 such loci (of which ~974 and ~745 are species-specific, respectively) that have the potential to splice into the genomic flanks and generate fusion proteins (Figures 2A and 2B). All FL ORF0 loci contain at least one SD and, as a result, they are present in this set. L1 family classification of ORF0 loci are shown in Table S1. Considering insertional polymorphisms within populations and somatic insertions, the number of ORF0 loci may be even larger. Analysis of human and chimp genomes for ORF0 loci revealed a conserved strong Kozak context around the first ATG (Figure 2C). To test the functionality of the consensus wild-type ORF0 Kozak (WT ORF0), we mutated it to an optimal Kozak sequence (OPT) as well as a −3/+4 mutant (MT ORF0). Expression of GFP-ORF0 was comparable between WT ORF0 and OPT, whereas the −3/+4 mutation abolished translational activity (Figures 2D and S2A).

We also extended our ORF0 analysis across mammalian genomes and found ORF0 loci with homology throughout the potential coding sequence, only in the genomes of Catarrhini. Within this parvorder of primates, Old World monkey and ape genomes contain on average ~50 and ~2,500 such ORF0 loci, respectively. Consensus Kozak sequences derived from these

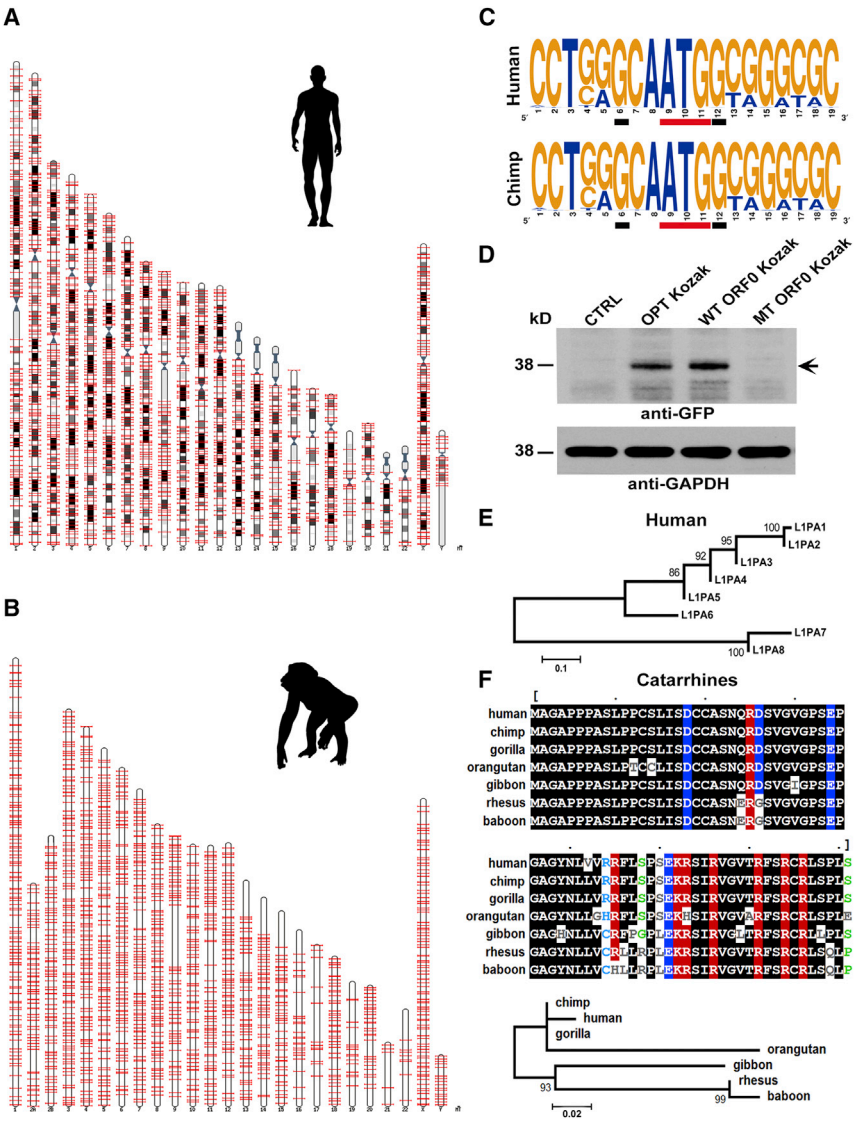


Figure 2. More than 3,000 Potential ORF0 Loci with a Conserved and Functional Kozak Sequence Exist in the Human and Chimp Genomes

(A and B) Chromosomal locations of ORF0 loci in the human and chimp reference genomes. The human and chimp genomes have ~3,528 and ~3,299 loci, respectively, that have the potential to splice into the genomic flanks and generate fusion proteins.

(C) ORF0 loci have a conserved strong Kozak context. Logo of Kozak sequences of ORF0 loci in human and chimp genomes. Start codon is underlined with red, and important nucleotides for translation initiation are underlined with black.

(D) The ORF0 Kozak sequence is functional. Western blot analysis of ORF0-GFP fusions driven by optimal (OPT), wild-type (WT ORF0), and mutant (MT ORF0) Kozak sequences from the GFP-ORF0-L1 construct. Arrow highlights the GFP-ORF0 protein.

(E) Basic phylogenetic analysis of ORF0 sequences in human L1PA families. ORF0 coding sequences were extracted from L1PA family consensus sequences and used in generating the maximum likelihood tree.

(F) Alignment of consensus ORF0 sequences derived from Catarrhini species. Charged residues are labeled in red and blue for positively and negatively charged, respectively. These consensus sequences were used in building the maximum likelihood tree for these primate species.

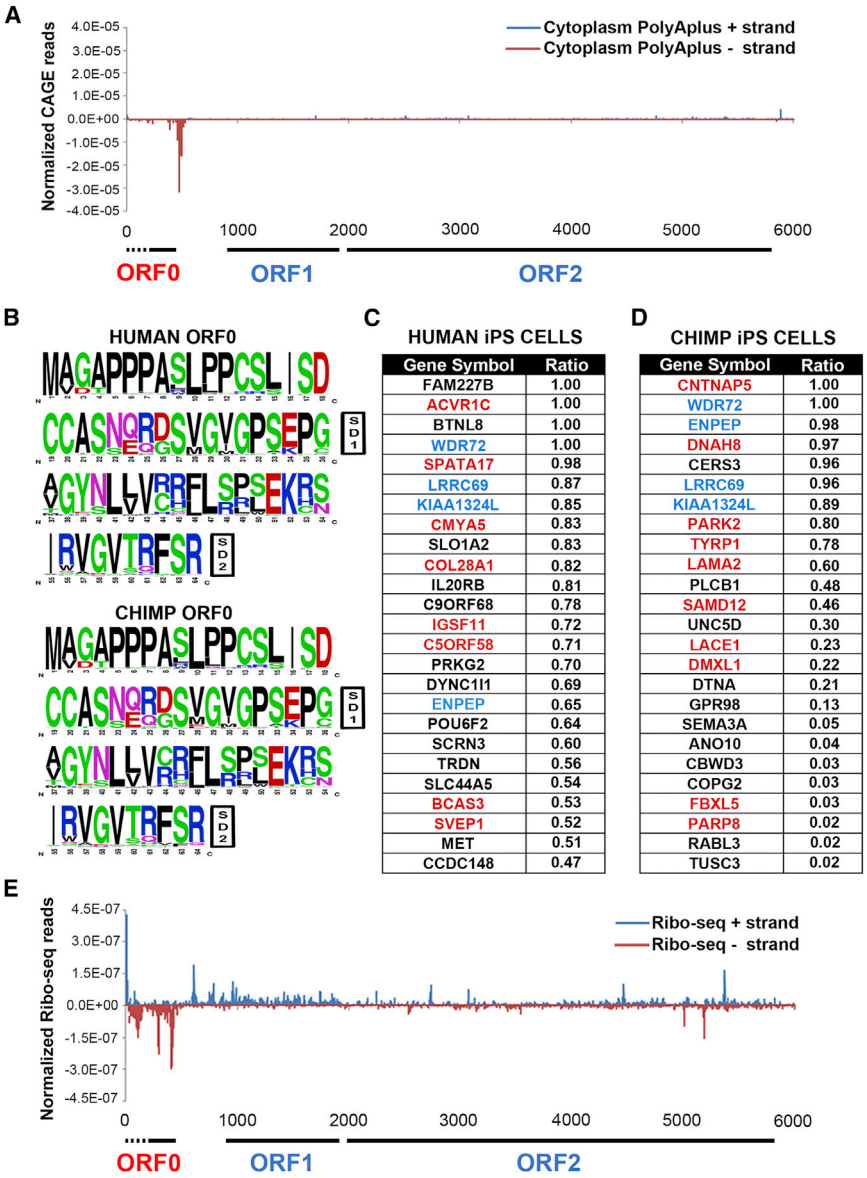
See also Figure S2 and Table S1.

loci suggest that the ORF0 Kozak context is conserved, including the G−3 and G+4 positions (red boxes) (Figure S2B). In New World monkeys, a very small number of ORF0 loci with limited N terminus homology were observed; however, due to the low number, a reliable consensus could not be built and thus these genomes were excluded from further investigation.

We next focused on the ORF0 coding sequences to get a better picture of evolutionary conservation of ORF0 within human L1 families and across primates. The alignments of ORF0 proteins from consensus L1PA1–8 sequences (Khan et al., 2006) are shown in Figure S2C. L1PA1 (that includes L1HS) and L1PA2 have intact SD1 and SD2. L1PA3-L1PA6 families contain a longer ORF0 due to a frameshift after SD2. In L1PA5 and L1PA6, SD1 is mutated but SD2 is conserved (data not shown). L1PA7 and L1PA8 have C termini that are distinct from the other L1PA families and lack SD1 and SD2. The abovementioned variation across L1PA families was recapitulated in the maximum likelihood tree (Figure 2E). Next, we generated consensus ORF0 sequences from the Catarrhines for comparison (Figure 2F). These primates have very similar consensus ORF0 proteins, except for the region between residues ~42 and 50. While all species' consensus ORF0 sequence contains SD2, rhesus and baboons lack SD1 due to a point mutation (Figure S2D). The maximum likelihood tree from the ORF0 sequences of Catarrhine genomes is shown in Figure 2F.

Capped and Polyadenylated ORF0 mRNAs Are Present in the Cytoplasm

One would expect ORF0 to be tightly regulated as a transposable element protein. In addition, short ORFs are technically challenging to uncover (Andrews and Rothnagel, 2014). To determine whether transcription from ORF0 loci could be detected, we turned to transcriptomic data. Cap analysis of gene expression (CAGE) data allow the mapping of transcription start sites (TSSs) and thus make it possible to identify the 5′ end of transcripts that originate from L1 (Faulkner et al., 2009; Shiraki et al., 2003). Our analysis of CAGE data showed that the majority of TSSs for antisense RNAs are upstream of ORF0 1st Met, suggesting that most antisense transcripts could have the capacity



Gene Symbol	Ratio
FAM227B	1.00
ACVR1C	1.00
BTNL8	1.00
WDR72	1.00
SPATA17	0.98
LRRC69	0.87
KIAA1324L	0.85
CMYA5	0.83
SLO1A2	0.83
COL28A1	0.82
IL20RB	0.81
C9ORF68	0.78
IGSF11	0.72
C5ORF58	0.71
PRKG2	0.70
DYNC1I1	0.69
ENPEP	0.65
POU6F2	0.64
SCRN3	0.60
TRDN	0.56
SLC44A5	0.54
BCAS3	0.53
SVEP1	0.52
MET	0.51
CCDC148	0.47

Gene Symbol	Ratio
CNTNAP5	1.00
WDR72	1.00
ENPEP	0.98
DNAH8	0.97
CERS3	0.96
LRRC69	0.96
KIAA1324L	0.89
PARK2	0.80
TYRP1	0.78
LAMA2	0.60
PLCB1	0.48
SAMD12	0.46
UNC5D	0.30
LACE1	0.23
DMXL1	0.22
DTNA	0.21
GPR98	0.13
SEMA3A	0.05
ANO10	0.04
CBWD3	0.03
COPG2	0.03
FBXL5	0.03
PARP8	0.02
RABL3	0.02
TUSC3	0.02

Figure 3. ORF0-Gene Fusion Transcripts Are Expressed in Human and Chimp iPS Cells

(A) Most of the antisense L1 transcription starts upstream of ORF0. Cytoplasmic polyA plus K562 CAGE (ENCODE/RIKEN) reads were mapped to L1HS consensus sequence.

(B) Protein logo of ORF0 loci that are untruncated until splice donor sites in the human and chimp genomes. Sequences from SD1 and SD2 loci are represented as protein sequence logos. Positions of SD1 and SD2 are indicated with black boxes.

(C and D) Table of top 25 protein-coding genes, for which RNA-Seq reads were detected at the splice junction with ORF0 in human and chimp iPS cells. Red-labeled genes have ORF0 fusions due to species-specific L1 insertions. Blue labeling represents genes for which ORF0 fusion transcripts were detected in both human and chimp iPS samples. Transcripts of black-labeled gene fusions are detected only in one species. The ratios of ORF0 isoforms with respect to the total (i.e., ORF0 + annotated gene isoforms) are shown in the ratio column. Table was sorted for ratio from high to low.

(E) Ribosome footprinting data from HEK293T cells were mapped to the L1HS consensus sequence.

See also Figure S3 and Table S2.

to encode ORF0 (Figure S3A). More importantly, ORF0 mRNA could be detected not only in whole cell but also in the cytoplasmic fraction; capped and polyadenylated ORF0 mRNAs were present in the cytoplasm (Figure 3A).

Most intronic L1s are in the reverse orientation with respect to their host genes (Smit, 1999), including L1s with intact ORF0: ~650 protein coding genes in human and ~450 in chimp contain ORF0 loci in the same direction as host gene transcription (data not shown), raising the possibility of a number of ORF0-host gene fusion events. The sequence logos of ORF0 loci in human and chimp that have the potential to splice, along with commonly used SD sites, are shown in Figure 3B.

To identify ORF0 fusion transcripts in human and chimp, we turned to RNA sequencing (RNA-seq) data that we had generated from iPS cells (Marchetto et al., 2013). Indeed, RNAs for a number of fusion events were observed in both species (Figures 3C, 3D, and S3B). Analysis of the contribution of ORF0 isoforms (iso-0) to the expression of these genes suggested that some genes were primarily driven by ORF0 in iPS cells, whereas for other genes, iso-0 contribution ranged from moderate to minor (Figures 3C and 3D; Table S2). We also extended our analysis to fibroblast-iPS pairs and observed that ORF0 transcript levels were dramatically upregulated in both human and chimp pluripotent stem cells compared to respective source fibroblasts (Figures S3C and S3D and data not shown).

ORF0 mRNAs Are Associated with Ribosomes

The presence of capped ORF0 mRNAs with a polyA tail in the cytoplasm as well as fusion transcripts with proximal exons of protein coding genes prompted us to investigate, by analyzing ribosome footprinting data, whether ORF0 RNAs were associated with ribosomes (ribosome footprinting [Ribo-seq]) (Ingolia et al., 2011). First, we mapped Ribo-seq reads obtained from HEK293T cell line (Shalgi et al., 2013) to L1HS consensus sequence (Figure 3E). In the sense orientation, a plateau of ribosome footprints was detected for ORF1 but ORF2 signal was much weaker, a finding that is in accordance with the known translation levels of ORF1 and ORF2 proteins (Alisch et al.,

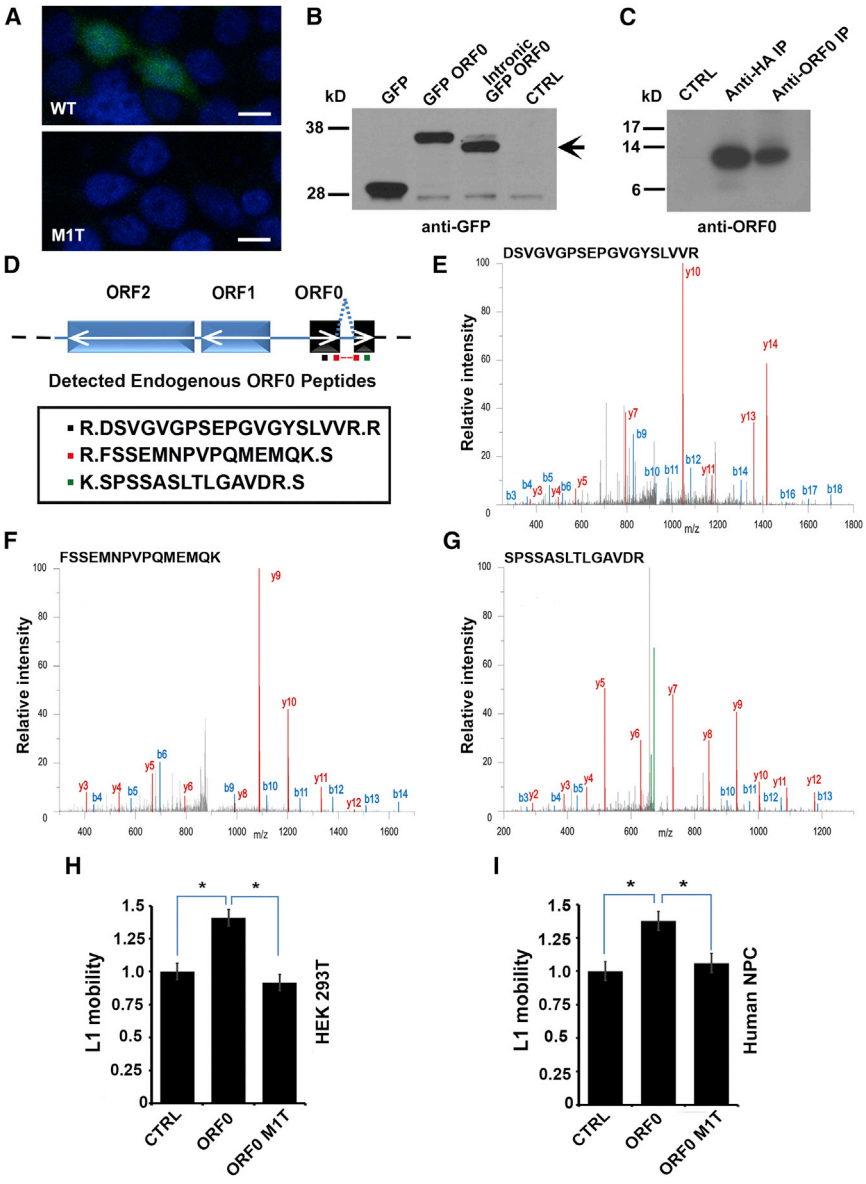


Figure 4. ORF0 Protein: Intronic Expression, Endogenous Detection, and Effect on L1 Mobility

(A) GFP-ORF0 can be expressed from an intronic position in an ORF0 initiator Met-dependent manner. The GFP-ORF0-L1 cassette (wild-type or M1T) was cloned in the antisense orientation in an intron. GFP was detected by confocal microscopy. Scale bar, 10 μm.

(B) Western blot analysis of GFP-ORF0 expression suggests that intronic ORF0 protein is produced but is not full-length. The fusion protein expressed from the intronic construct is indicated with the black arrow.

(C) Functionality of ORF0 antibody was tested using overexpressed protein, by immunoprecipitation, and subsequent western blotting.

(D) Schematic description and sequences of identified ORF0 peptides. The first peptide (black square) resides upstream of SD2. The second peptide (red square) spans the splice junction of proteins formed through splicing between SD2 and SA1. The third peptide (green square) is located downstream of SA1 within the L1 sequence.

(E–G) Spectra of peptides (#1, #2, and #3) identified by proteomic searches. Green peaks in (G) represent neutral losses.

(H and I) Overexpression of ORF0 protein, but not ORF0 RNA, increases L1 mobility based on luciferase L1 reporter in HEK293T cells and human NPCs. Potential antisense RNA effects were controlled for by using a single-nucleotide mutant ORF0 that replaces the initiator Methionine with Threonine. Data are presented as mean ± SEM. *Denotes $p < 0.05$ significance between indicated groups using t test.

See also Figure S4.

2006). In the antisense orientation, a strong signal was evident for ORF0 (Figure 3E). Interestingly, this signal also extended beyond the FL ORF0 sequence, which may be due to within-L1 splicing events (see below and data not shown) or L1s from older families, in which the encoded consensus ORF0 extends until the end of L1 (see Figure S2C). Even though reads obtained by ribosome footprinting were shorter than those gained from RNA-seq, we observed spliced ORF0 footprints of in-frame fusions to *SCAMP1*, *SLC44A5*, *GJB4*, *HTR2C*, and *RABGAP1L* (driven by a human-specific L1 insertion). Thus, the influence of ORF0 may not necessarily be limited to L1 biology.

ORF0-Downstream Exon Fusion Protein Is Expressed

To test whether ORF0 could be transcribed and translated from an intronic position, we cloned the GFP-ORF0-L1 cassette in the antisense orientation within a natural human intron. Upon transfection of this construct into cells, GFP-ORF0 was expressed. Moreover, translation started at the ORF0 1st Met, as the M1T mutation abolished expression (Figure 4A). Interestingly, GFP signal was localized throughout the cell instead of in nuclear foci. This difference in localization was explained by western blot analysis, which showed that intronic GFP-ORF0 fusion protein was different from GFP alone or GFP-FL ORF0, suggesting that a spliced product was translated (Figure 4B). Generation of a fusion protein via splicing between SD1 of ORF0 and the downstream exon was confirmed by sequencing (data not shown).

Proteomic Detection of Endogenous ORF0 Peptides

Having observed ORF0 transcripts as well as expression from reporter plasmids, we investigated endogenous ORF0 products. Proteomic identification of ORF0 requires detection of peptides within unspliced ORF0 or N terminus ORF0 fragments of fusion proteins. Therefore, due to the small size of ORF0, a limited number of possible peptides are available for detection by mass

spectrometry. In addition, the distribution of the target residues (K and R) for trypsin, the most commonly used enzyme for proteomics, leads to the generation of non-ideal peptide fragment sizes (see Figure 2F): the N terminus is poor in these residues whereas the C terminus is rich, generating a very small number of peptides optimal for mass spectrometry. In fact, only one peptide from the main body of ORF0 could be detected in our mass spectrometry analysis of overexpressed ORF0 (Figure S4A).

Nevertheless, we proceeded to attempt detection of endogenous ORF0 peptides. We started by raising polyclonal anti-ORF0 antibodies targeting the consensus L1HS FL-ORF0 protein. Upon confirmation that the ORF0 antibody worked for immunoaffinity enrichment from overexpressed HA-ORF0 extracts (Figure 4C), we turned to the cultured cell type that expressed the highest levels of ORF0 transcripts as a class: human pluripotent stem cells. In parallel, we computationally generated an RNA expression-based ORF0 proteomics database that included potential unspliced and spliced ORF0 proteins. The combined ORF0-Human Uniprot database was used in spectra searches. Next, immunoprecipitates from control and ORF0 antibody were subjected to mass spectrometry analysis. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) spectra searches did not find any ORF0 fragments in control antibody samples. However, searches of anti-ORF0 immunoprecipitates led to identification of endogenous ORF0 peptides (Figures 4D–4G). Spectra obtained from overexpressed peptides for comparison and further information on all the spectra are presented in Figures S4A–S4D. The first peptide (black square) resides upstream of SD2. The second peptide (red square) spans the splice junction of proteins formed through splicing between SD2 and SA1 (SA1: based on RNA-seq analysis, a functional splice acceptor site 336 nucleotides downstream of the ORF0 start site in the L1 5′UTR antisense). The third peptide (green square) is located downstream of SA1 within the LINE-1 sequence (Figure 4D). There are multiple loci that can encode the observed ORF0 peptides and the exact identities of source loci are currently unknown.

ORF0 Enhances L1 Mobility

Given the fact that the ORF0 coding sequence resides in the L1 5′UTR with bidirectional promoter activity, the most parsimonious function for ORF0 would be a potential effect on L1 mobility. Human L1s driven by CMV or CAG promoters are mobile (Moran et al., 1996); thus it is clear that ORF0 is not essential for L1 activity. We attempted to test potential in *cis* effects of ORF0 mutations; however, this task was hampered by the fact that the ORF0 sequence overlaps with the forward L1 promoter (data not shown). Thus, we overexpressed ORF0 in *trans* and tested for its effect on L1 mobility. To prevent any direct antisense L1 RNA effect due to transcription of ORF0, we used a CAG promoter-driven L1 reporter. In HEK293T cells, ORF0 expression led to a ~41% increase in L1 mobility (Figure 4H). To rule out any indirect effects of expressing antisense L1 RNA, we also used the single nucleotide mutant control, ORF0 M1T, that did not produce ORF0 protein. This construct had no effect on L1 mobility, strongly suggesting that ORF0 protein was responsible for the observed increase (Figure 4H). Importantly, wild-type, but not M1T mutant, ORF0 also increased L1 mobility in human embryonic stem (ES) cell-derived neural progenitors (human NPC) by ~38% (Figure 4I), bringing forth the possibility that ORF0 may contribute to somatic variation by enhancing L1 activity in pluripotent cells.

DISCUSSION

The constant competition between transposable elements and host-protective mechanisms contributes to genome evolution (Daugherty and Malik, 2012; Slotkin and Martienssen, 2007). It is currently unclear whether L1 antisense promoter activity has been a major factor in this arms race. From an L1 perspective, antisense transcription can positively influence sense expression through recruitment of transcriptional machinery, inducing open chromatin structure or via formation of a non-coding RNA. On the other hand, expression of antisense RNA can lead to dsRNA formation, which may trigger an RNAi response (Mätlik et al., 2006; Yang and Kazazian, 2006). Our results suggest that, in addition to the aforementioned roles, L1 5′UTR has the ability to initiate translation in the antisense direction.

ORF0 is present in more than ~2,500 loci in the ape genomes, whereas this number is much smaller in the Old World monkeys. While some of this difference may be related to variable genome sequence quality, we expect this difference to mostly represent L1 biology. The alignment of ORF0 sequences from human L1PA1–8 suggests that the main difference between these families is the C terminus of ORF0. We have also noticed that the sequences around the ORF0 translation start site influence forward promoter activity. It is possible that the translation activity in the antisense L1 5′UTR is coupled with the forward promoter activity, and thus the N terminus is more conserved with respect to the rest of the ORF0 sequence due to evolutionary pressure. If that indeed is the case, translation activity in rhesus and baboon may generate distant relatives of the ape ORF0. Consistent with this hypothesis, searches for ORF0 in New World monkey genomes reveal a very small number of loci that have homology to human ORF0, with similarity only at the N terminus. Considering the fact that L1 retrotransposons recruit new 5′UTRs over time, it is conceivable that distant primates such as marmosets and squirrel monkeys may have significantly different 5′UTRs. Improved primate genome sequence quality and future experimentation will allow the testing of these possibilities.

Expression of ORF0, but not an untranslated point mutant version, enhances L1 activity from L1 luciferase mobility reporter in human cells, suggesting a role for ORF0 protein in L1 activity. We currently do not know the mechanism of this effect. Similar to ORF1, ORF0 does not share any extensive homology with known genes, so it is not possible to propose a domain-based prediction. However, ORF0 is a highly positively charged protein that may act by binding to nucleic acids. The PML proximity to ORF0 is intriguing, especially given that a large number of proteins are recruited to PML bodies depending on the cellular state, with stress playing a prominent role in determining the content as well as the morphology of PML bodies. Interestingly, PML is involved in antiviral responses and protects cells from viral infections. Some viral proteins target the integrity of PML bodies and a large number of components are transcriptionally regulated by the interferon pathway (Everett and Chelbi-Alix, 2007). Whether

localization adjacent to PMLs is reflective of ORF0 function or the cell's response remains to be seen. It is possible that ORF0, analogous to some viral proteins, may interfere with the functions of PML and enhance mobility. Further studies will be required to gain insight into the mechanism of action of ORF0.

The influence of ORF0 may not necessarily be limited to L1 biology. Our transcriptomic analysis suggests that exons of host genes provide splice acceptor sites for intronic or proximal ORF0 loci. Overall, ORF0 expression levels correlate with the pluripotency of the cell types and ORF0-proximal exon fusion products are detected by proteomics. While any effects of ORF0 expression on the host or proximal gene would be context and sequence dependent, one could make certain predictions. If the downstream exon is in frame with respect to ORF0 upon splicing, the N terminus of the host protein would be replaced by an ORF0 variant, which could alter the localization and/or function. Out-of-frame ORF0 fusions would contain amino acids from an alternative frame of the gene and most would encounter a stop codon. Such transcripts, depending on the context, might be expressed or be subject to nonsense-mediated decay (NMD). By virtue of high copy numbers and sequence variants, one would expect to see varying degrees of NMD response. In addition, cell-state transitions, stress, and crosstalk with the RNAi pathway might provide opportunities for NMD targets to be translated (Kervestin and Jacobson, 2012). In cases of fusions of ORF0 located upstream of coding sequences, ORF0 might act as an upstream ORF (uORF). Since uORF function is affected by the length and sequence of the uORF as well as by the distance between the upstream and the main ORF, variations in ORF0 sequences could result in differential translation regulation (Andrews and Rothnagel, 2014).

L1s, as the sole autonomously active retrotransposons in primate genomes, continue to shape our genomes. Our data suggest that, in addition to their previously ascribed roles in gene regulation (Huang et al., 2012), L1s contain a third ORF and have the ability to generate insertion site-dependent ORFs via splicing. Considering the fact that transcription and translation start within L1 elements, these ORF0 variants could be co-regulated. Analogous to the other L1 proteins, disorders such as neoplasms (Rodić et al., 2014) may provide opportunities for higher ORF0 expression, which in turn could contribute to the pathological phenotypes. It is tempting to speculate that, over evolutionary time, the propensity of ORF0 to splice into proximal exons may have led to not only gene regulatory changes but also the emergence of new proteins. The extent to which ORF0 variants contribute to diversity, both in evolutionary terms and disease conditions, remains to be investigated.

EXPERIMENTAL PROCEDURES

Cloning and Mutagenesis

Primers from IDT and Phusion High Fidelity Polymerase (NEB) were used for PCRs. pGL4.10 (Promega) was the plasmid backbone used for promoter luciferase assays. To test the effect of ORF0 expression on L1 mobility, ORF0 promoter, coding sequence, and the downstream sequence (until the end of L1 in the antisense orientation) was cloned into pEF-BOS-EX (Mizushima and Nagata, 1990). To include any potential within-L1 splicing products and prevent contribution from the plasmid backbone, a fragment containing stop codons in all three frames as well as a polyA signal was included immediately downstream of the insert. ORF0-GFP construct was cloned into a modified (SV40 promoter and luciferase removed) pSICheck2 vector (Promega). A modified (luciferase cassette removed) pYX014 plasmid (Xie et al., 2011) was used for GFP-ORF0 and mCherry-ORF0 cloning: nucleotide 13 of ORF0 was mutated (C → G) to generate an AscI site that was used for subsequent cloning of GFP and mCherry. HA-tagged ORF0 was cloned into pCDNA3.1 for in vitro translation. GFP-ORF0-L1 cassette was cloned into pEF-BOS-EX with BglII for intronic expression. Mutagenesis was carried out using the Quick Change II XL Site Directed Mutagenesis Kit (Agilent Technologies).

RNA Extraction, Reverse Transcription, and cDNA Preparation

RNA was prepared using Trizol (Invitrogen). cDNA was synthesized using the Superscript III First Strand Synthesis System for RT-PCR (Invitrogen).

Cell Culture and Transfection

HEK293T cells (ATCC) were cultured in DMEM$^+$ GlutaMax medium (Life Technologies) supplemented with 10% fetal bovine serum (Omega Scientific) and grown at 37°C in 5% CO_2. Cells were transfected using polyethylenimine (Polysciences). HUES6 human ES cells were cultured feeder-free on Matrigel-coated dishes (BD) using mTeSRTM1 (StemCell Technologies) and passaged once every 3–4 days using Collagenase type IV enzyme.

Human NPC Derivation, Growth, and Nucleofection

NPCs were differentiated from HUES6 cells through embryoid body and rosette generation and grown as previously described (Marchetto et al., 2010). Plasmid delivery into human NPCs was performed by nucleofection (Lonza/Amaxa Nucleofector, kit VPG-1005).

In Vitro Translation

ORF0 was synthesized in vitro by employing the TNT Coupled Reticulocyte Lysate System (Promega) using T7 polymerase.

Cell Extracts and Western Blot Analysis

Cells were harvested 2 days post transfection, washed with cold DPBS, and lysates were prepared with ice cold RIPA lysis buffer (50 mM Tris-HCl [pH 7.4], 150 mM NaCl, 0.25% deoxycholic acid, 1% NP-40, 0.1% SDS, and 1 mM EDTA) containing complete protease inhibitor cocktail with EDTA (Roche) and 1 mM DTT. Lysates were incubated on ice for 15 min, spun at 14,000 × *g* for 15 min at 4°C, and the supernatants were collected. Primary antibodies: rabbit α-GFP (1:2000, Santa Cruz sc-8334), rat α-HA peroxidase high-affinity 3F10 (1:1000, Roche), and α-ORF0 (1:300). Secondary antibody: (1:5,000, GE NA934).

Fluorescence Detection

Cells were grown in poly-L-lysine (Sigma) coated 2-well LabTek chamber slides (Nunc, Fisher), fixed in 4% paraformaldehyde (Sigma) for 15 min at room temperature, and washed with TBS. The nuclei were stained with DAPI (1:1,000, Sigma) and the slides were mounted using polyvinyl alcohol with DABCO (Sigma).

Computational Analyses

Detection and visualization of ORF0 loci in human and chimp genomes: the UCSC genome browser and Ensembl databases were used to retrieve potential ORF0 coding sequences, which were subsequently in silico translated. The Ensembl databases (hg19, panTro4) were used for blastn, allowing some local mismatch but no gap to obtain ORF0 loci. An alternative method of retrieving all potential full-length ORF0 sequences from RepeatMasker was tested and led to very similar results. Custom python scripts and EMBOSS suite (Rice et al., 2000) were used for identification and characterization of ORF0 loci, full-length as well as untruncated-until-splice-donor, in the genome. Sequences that did not contain a GT dinucleotide at the splice donor site were removed. Ensembl Karyotype View tool was used for visualization of the ORF0 loci. Upon confirmation of an annotation error in the Chimp Chr 2B, the erroneous fragment was removed from the image. The removed region contained no genes or TEs. Analysis of RNA-seq datasets: RNA-seq (human and chimp iPS cells) data from GEO: GSE47626 (Marchetto et al., 2013), GEO: GSE44646 (Wang et al., 2014), GEO: GSE60996 (Gallego Romero et al., 2015), and ArrayExpress: E-MTAB-2031 (Chan et al., 2013); CAGE

(capped 5′ RNA-seq) data from GEO: GSE34448 (Djebali et al., 2012); Ribo-seq (ribosome footprinting) data from GEO: GSE32060 (Shalgi et al., 2013) were analyzed from raw FASTQ files in a consistent manner. Reads were aligned to the reference human (hg19) and chimpanzee (panTro4) genomes with STAR, which is capable of identifying novel splice junctions (Dobin et al., 2013). Spliced ORF0 reads were identified by filtering out all multimappers and only considering reads originating from an ORF0 locus (direct overlap of 5′ end for stranded RNA-seq and direct overlap of either read end for unstranded RNA-seq). Read distributions along L1 were found by aligning reads to the consensus L1HS element using STAR (Dobin et al., 2013). Read densities along the + and – strands were further normalized based on the total number of reads in each experiment that were alignable to the full genome. Ratios of isoforms (ORF0 versus total) were determined by comparing the splice junction reads (j) of ORF0, j(0c), to j(ab), j(bc), j(cd), j(de), where the order of exons are a-b-0-c-d-e: ratio = average((j(0c)/(average(j(ab),j(bc)) + j(0c))), (j(0c)/average(j(cd) + j(de)))). This allowed us to get a more reliable estimate compared to calculations that rely solely on ratio at one exon j(0c) and j(bc) and to reduce the 3′ bias that is observed in polyA-based sequencing. In the few cases where the ratio is higher than 1 (maximum being 1.2), these ratios are presented as 1 in the tables (Figures 3C, 3D, and Table S2). Genes in the tables went through further manual inspection. Proteomic database generation: RNA-seq reads from human iPS/ES cells were assembled using Cufflinks, ORF0 containing transcripts were selected and redundancies were removed. In parallel, ORF0-containing mRNAs that are either ESTs or annotated transcripts were added to the RNA-seq list. The combined list was in silico translated and appended to the current human Uniprot database for spectra searches. Determination of species that have ORF0 loci: L1HS/L1Pt consensus ORF0 sequence (identical) was used in Blat and blast searches to determine the genomes that contain ORF0 loci. The absence of ORF0 loci in non-Catarrhine primates was further confirmed by in silico translation of L1 sequences (with repeat start <1,000) and subsequent search for loci that can encode a polypeptide with > = 50% identity to ORF0 protein (FL or SD1-ORF0). Generation of consensus primate ORF0 sequences for phylogenetic analysis: ORF0 loci that can encode an untruncated protein >210 nucleotides were retrieved via blast searches and subsequent in silico translation and filtering. The sequences were trimmed to 213 nucleotides (length of FL ORF0) and used in molecular phylogenetic analysis. Basic molecular phylogenetic analysis: Clustal Omega was used to generate the alignments. The evolutionary history was inferred by using the maximum likelihood (ML) method based on the JTT matrix-based model. The tree with the highest log likelihood is shown. A total of 1,000 bootstrap replicates were used for test of phylogeny. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model and then selecting the topology with superior log likelihood value. The tree is drawn to scale, and branch lengths represent number of substitutions per site. The analysis involved amino acid sequences and all positions with <95% site coverage were eliminated. Muscle generated alignments as well as maximum parsimony analysis generated a very similar tree. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013). Analyses using RAxML and PhyML as well as neighbor joining methods resulted in very similar trees. DNA and protein logos were generated using WebLogo (Crooks et al., 2004).

L1 Mobility Assays

Luciferase-based L1 mobility reporters used were previously described (Xie et al., 2011). Cells were transfected/nucleofected with experimental constructs together with L1 mobility reporter plasmid pYX017. Luciferase activity was quantified at day 3 using the Dual Luciferase Reporter 1000 Assay System (Promega, E1980) and a Perkin Elmer Victor X Luminometer. A two-tailed t test was used for statistical analysis.

Promoter Activity Assays

Promoter activity was measured by co-transfecting ORF0 promoter constructs cloned into pGL4.10 (Promega) along with the normalization vector phRLTK (Promega). Activity was measured after 2 days, as in the L1 activity assays. A two-tailed t test was used for statistical analysis.

Antibody Generation and Immunoprecipitations

Peptides corresponding to ORF0 amino acid residues 20-34, 33-49 and 50-65 in the L1HS consensus were synthesized, conjugated to KLH and used in generation of rabbit polyclonal antibodies (Covance). For immunoprecipitations (IPs), cells were washed with DPBS, collected, and frozen. Cell pellets were thawed in mDm lysis buffer (25 mM Tris [pH 7.5], 150 mM NaCl, 1.5 mM $MgCl_2$, 1% Triton X-100, 1 mM DTT, protease inhibitors [Roche]) and supernatant from a 15,000 × *g* 15 min spin was used in IPs. Control and ORF0 antibodies were conjugated to magnetic beads (Pierce). IP duration was 4–6 hr, washes were done with the mDm buffer, and beads were heated to 95°C for 10–12 min for elution.

Proteomic Sample Prep and Analysis

Samples were precipitated by methanol/chloroform. Dried pellets were dissolved in 8 M urea/100 mM Tris, [pH 8.5]. Proteins were reduced with 5 mM tris(2-carboxyethyl) phosphine hydrochloride (TCEP, Sigma-Aldrich) and alkylated with 10 mM iodoacetamide (Sigma-Aldrich). Proteins were digested overnight at 37°C in 2 M urea/100 mM Tris, [pH 8.5], with trypsin (Promega). Digestion was quenched with formic acid, 5% final concentration and a final volume of 50 μl.

The digested samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo). Samples were analyzed with injections of 8 μl of the protein digest per LC/MS run. The digest was injected directly onto a 40-cm, 75-μm ID column packed with BEH 1.7 μm C18 resin (Waters). Samples were separated at a flow rate of 200 nl/min on a nLC 1000 (Thermo). Buffer A and B were 0.1% formic acid in water and acetonitrile, respectively. Two reverse phase gradients of 140 min and 450 min were used to maximize sampling efficiency of the digest. Ninety percent buffer B was used for 10 min final washes at the ends of gradients. Column was re-equilibrated with 20 μl of buffer A prior to the injection of sample. Peptides were eluted directly from the tip of the column and nanosprayed into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap Fusion was operated in a data-dependent mode. Full MS^1 scans were collected in the Orbitrap at 120 K resolution with a mass range of 400–1,600 m/z and an AGC target of $5e^5$. The cycle time was set to 3 s, and within this 3 s the most abundant ions per scan were selected for CID MS/MS in the ion trap with an AGC target of $1e^4$ and minimum intensity of 5,000. Maximum fill times were set to 50 ms for MS scans and 100 and 35 ms for MS/MS scans in the 140 min and 450 min methods, respectively. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled and dynamic exclusion was used with an exclusion duration of 5 s.

Protein and peptide identification were done with Integrated Proteomics Pipeline- IP2 (Integrated Proteomics Applications). Tandem mass spectra were extracted from raw files using RawConverter and searched with ProLuCID against ORF0-human UniProt database. The search space included all fully tryptic and half-tryptic peptide candidates. Carbamidomethylation on cysteine was considered as a static modification. Data were searched with 50 ppm precursor ion tolerance and 500 ppm fragment ion tolerance. Data were filtered to 10 ppm precursor ion tolerance post search. Identified proteins were filtered using DTASelect (Tabb et al., 2002) and utilizing a target-decoy database search strategy to control the false discovery rate to 1% at the protein level.

Imaging

All imaging was carried out using a Zeiss LSM 780 Confocal Microscope. Images were taken using either a 20× or a 100× oil objective. The z stack intervals were 1 μm. Image analysis was performed with ZEN (Zeiss) and Imaris (Bitplane). Both PML and ORF0 foci were identified using the Spots object on Imaris (Bitplane) using a fixed spot size of 0.5 μm (the measured average XY diameter of nuclear bodies).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, two tables, and one movie and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.025.

AUTHOR CONTRIBUTIONS

A.M.D. is the lead author, designed the study, and was involved in execution of all wet lab, computational studies, and data analysis. I.N. contributed to the concept and performed in vitro translation assays. B.E.K. contributed to the concept, performed imaging, and image processing. M.P. provided technical support in performing all wet lab experiments (including imaging) and helped with manuscript preparation. C.B. contributed to the concept, provided bioinformatics guidance, and performed the initial computational analysis of RNA-seq datasets. M.C.N.M. performed experiments involving human NPC derivation, culture, and nucleofection. J.K.D. and J.J.M. performed proteomic sample prep and analysis. J.K.D., J.J.M., A.A., and J.M. performed proteomic searches. L.M. provided technical support. A.S. contributed to the concept and provided proteomic guidance. F.H.G. is the senior author, contributed to the concept, analyzed the data, revised the manuscript, and provided financial support. A.M.D. and F.H.G. wrote the manuscript. All authors contributed comments on the manuscript.

ACKNOWLEDGMENTS

This work was supported by funds from the Leona M. and Harry B. Helmsley Charitable Trust, the JPB Foundation, and the G. Harold and Leila Y. Mathers Charitable Foundation. We thank Gökhan Şentürk, Ruth Keithley, Ana Mendes, Dilara Halim, and Iryna Gallina for technical help; Sara Linker, Christos Tzitzilonis, and Stephane Boissinot for discussions; Peter Hemmerich and Wenfeng An for reagents; James Fitzpatrick and Michael Adams for help with imaging; Mary Lynn Gage for editorial comments on the manuscript; and everybody involved in GEO and ENA database generation and data contributions.

Received: January 4, 2015
Revised: July 7, 2015
Accepted: August 25, 2015
Published: October 22, 2015

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.

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.

Bernardi, R., and Pandolfi, P.P. (2007). Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. Nat. Rev. Mol. Cell Biol. *8*, 1006–1016.

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. USA *100*, 5280–5285.

Chan, Y.S., Göke, J., Ng, J.H., Lu, X., Gonzales, K.A., Tan, C.P., Tng, W.Q., Hong, Z.Z., Lim, Y.S., and Ng, H.H. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663–675.

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

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

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.

Cruickshanks, H.A., Vafadar-Isfahani, N., Dunican, D.S., Lee, A., Sproul, D., Lund, J.N., Meehan, R.R., and Tufarelli, C. (2013). Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer. Nucleic Acids Res. *41*, 6857–6869.

Daugherty, M.D., and Malik, H.S. (2012). Rules of engagement: molecular insights from host-virus arms races. Annu. Rev. Genet. *46*, 677–700.

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

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

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.

Everett, R.D., and Chelbi-Alix, M.K. (2007). PML and PML nuclear bodies: implications in antiviral defence. Biochimie *89*, 819–830.

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.

Gallego Romero, I., Pavlovic, B.J., Hernando-Herraez, I., Zhou, X., Ward, M.C., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., Mitrano, A., et al. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. eLife *4*, e07103.

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.

Hancks, D.C., and Kazazian, H.H., Jr. (2012). Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. *22*, 191–203.

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.

Huang, C.R., Burns, K.H., and Boeke, J.D. (2012). Active transposition in genomes. Annu. Rev. Genet. *46*, 651–675.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Kervestin, S., and Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. Nat. Rev. Mol. Cell Biol. *13*, 700–712.

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.

Kleckner, N. (1990). Regulation of transposition in bacteria. Annu. Rev. Cell Biol. *6*, 297–327.

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. USA *94*, 10155–10160.

Kozak, M. (1987). An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. *15*, 8125–8148.

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.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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.

Macia, A., Muñoz-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.

Marchetto, M.C., Carromeu, C., Acab, A., Yu, D., Yeo, G.W., Mu, Y., Chen, G., Gage, F.H., and Muotri, A.R. (2010). A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. Cell *143*, 527–539.

Marchetto, M.C., Narvaiza, I., Denli, A.M., Benner, C., Lazzarini, T.A., Nathanson, J.L., Paquola, A.C., Desai, K.N., Herai, R.H., Weitzman, M.D., et al. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. Nature *503*, 525–529.

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.

Mätlik, K., Redik, K., and Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. J. Biomed. Biotechnol. *2006*, 71753.

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

Mizushima, S., and Nagata, S. (1990). pEF-BOS, a powerful mammalian expression vector. Nucleic Acids Res. *18*, 5322.

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.

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

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Rodić, N., Sharma, R., Sharma, R., Zampella, J., Dai, L., Taylor, M.S., Hruban, R.H., Iacobuzio-Donahue, C.A., Maitra, A., Torbenson, M.S., et al. (2014). Long interspersed element-1 protein expression is a hallmark of many human cancers. Am. J. Pathol. *184*, 1280–1286.

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.

Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013). Widespread regulation of translation by elongation pausing in heat shock. Mol. Cell *49*, 439–452.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA *100*, 15776–15781.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. *9*, 657–663.

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.

Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd. (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. *30*, 2725–2729.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V., et al. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature *516*, 405–409.

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.

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.

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.