PNAS
www.pnas.org

Supplementary Information for

RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA

John B. Moldovan[1, *], Yifan Wang[1, 2], Stewart Shuman[4], Ryan E. Mills[1, 2], and John V. Moran[1, 3, *]

[1]Department of Human Genetics, [2]Department of Computational Medicine and Bioinformatics, and [3]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [4]Molecular Biology Program, Sloan Kettering Institute, New York, NY 10065, USA.

*Correspondence should be addressed to: jmoldova@umich.edu (J.B.M.) and moranj@umich.edu (J.V.M.)

This PDF file includes:

Supplementary text
Figs. S1 to S4
Tables S1 to S9
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1805404116

Supplementary Information Text

METHODS

Cell Culture

The following cell lines were maintained at 37°C with 7% CO_2 in humidified incubators. All tissue culture reagents were purchased from Thermo Fisher Scientific (Waltham, MA) unless stated otherwise. HeLa-JVM cells (1) were grown in high-glucose DMEM supplemented with 10% Fetal Bovine Serum (FBS), 100 U/mL penicillin-streptomycin, and 0.29 mg/mL L-glutamine. HeLa-HA cells (2) were grown in MEM supplemented with 10% FBS, 100 U/mL penicillin-streptomycin, 0.29 mg/mL L-glutamine, and 0.1 mM nonessential amino acids. PA-1 cells (3) were grown in MEM supplemented with 10% FBS, 100 U/mL penicillin-streptomycin, 0.29 mg/mL L-glutamine, and 0.1 mM nonessential amino acids. H9 human embryonic stem cells (hESCs) and H9-hESC-derived neural progenitor cells were cultured and maintained by the Garcia-Perez lab as described previously (4-6).

Plasmids

All human L1 expression plasmids contain the L1.3 (GenBank accession no. L19088) (7) DNA cloned into the pCEP4 mammalian expression vector (Thermo Fisher Scientific) unless stated otherwise. A CMV promoter augments the expression of L1 in these plasmids unless noted otherwise. The L1 expression plasmids contain a SV40 polyadenylation signal that is located downstream of the native L1 polyadenylation signal. All plasmid DNA was prepared with a Midiprep Plasmid DNA Kit (Qiagen, Germany).

pJM101/L1.3Δneo: is an engineered plasmid expression vector that expresses an active wild-type human L1 element (L1.3) (8). The L1 element has been cloned into a pCEP4 expression vector (Thermo Fisher Scientific). L1 expression is augmented by a CMV promoter located at the 5' end of the L1 and an SV40 polyadenylation sequence that flanks the 3' end of the L1.

pJM108/L1.3Δneo: is similar to pJM101/L1.3Δneo, but contains a S119X stop mutation in ORF1p (1, 8, 9)

pJM105/L1.3Δneo: is similar to pJM101/L1.3Δneo, but contains a D702A missense mutation in the ORF2p RT active site (8).

pJBM119/L1.3Δneo: is similar to pJM101/L1.3Δneo, but also contains a H230A mutation in the ORF2p EN domain and a D702A missense mutation in the ORF2p RT active site.

pCEP/GFP: is a pCEP4-based plasmid that contains the humanized *Renilla* green fluorescent protein (hrGFP) coding sequence from phrGFP-C (Agilent Technologies, Santa Clara, CA), which has been cloned downstream of the pCEP4 CMV promoter (9).

pJBM/RtcB: is a modified version of the human RtcB cDNA clone (SC319629) purchased from Origene Technologies, Rockville, MD. Site specific mutagenesis was used to make an A to C change in the SC319629 plasmid sequence upstream of the RtcB open reading frame to disrupt an upstream ATG codon.

Transfection of plasmid DNA and isolation of RNA from transfected cells

Approximately $8x10^5$ HeLa-JVM cells were seeded in 60 mm dishes (BD Biosciences, San Jose, CA) and transfected with 2.5 µg of plasmid DNA using 7.5 µL FuGENE 6 (Promega, Madison, WI) the following day according to the manufacturer's protocol. Two days after transfection, cells were collected by scraping with a cell scraper and centrifuged at 1000 X *g* at 4°C and the resultant cell pellets were frozen at -80°C. Frozen cell pellets were then thawed and total RNA was extracted using an RNeasy mini kit (Qiagen) according to manufacturer's protocol.

RT-PCR using transfected cell RNA

Synthesis of cDNA was carried out using ~2 µg of total RNA purified from transfected cells with the SuperScript First-Strand Synthesis System for RT-PCR (Thermo Fisher Scientific) according to the manufacturer's protocol. An oligo$(dT)_{12\text{-}18}$ primer supplied with the kit was used to prime cDNA synthesis reactions. Reactions were incubated at 42°C for 50 minutes followed by an incubation at 70°C for 15 minutes to inactivate the reverse transcriptase (RT). One microliter of RNase H supplied with the kit was then added to reactions followed by a 37°C incubation for 15 minutes. Reactions (20 µL) were diluted 1:5 with water to a final volume of 100 µL.

Two microliters of the diluted cDNA reaction was then subjected to nested PCR using Platinum *Taq* DNA Polymerase (Thermo Fisher Scientific) according to the manufacturer's protocol. PCR reactions (50 µL total volume) included 2 µL of the cDNA, 0.25 µL Platinum *Taq*, 0.2 µM forward and reverse primers, 1.5 mM $MgCl_2$ and 0.2 mM dNTPs. The first round of PCR used the following primers: U6s1 (sense) and SV40as (antisense). The second round of PCR used: U6s2 (sense) and 3UTRas (antisense). Thermal cycler conditions: an initial cycle of 94°C for 2 minutes, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 120 seconds at 72°C with a final cycle of 72°C for 5 minutes. The same conditions were used for each round of nested PCR. PCR products (10-20 μL) were visualized on 2% agarose gels using SYBR safe DNA gel stain (Thermo Fisher Scientific).

Cloning and Sequencing of RT-PCR products

Visible cDNA bands were excised from agarose gels and the cDNA was purified using the Wizard SV Gel and PCR Clean-Up System (Promega). For transfected HeLa cell RT-PCR experiments (Fig. 1C), RT-PCR products from untransfected HeLa cells could not be analyzed due to a lack of visible cDNA bands. Purified DNA from excised gel slices was cloned into a pCR4-TOPO TA vector (Thermo Fisher Scientific), DNA was isolated from individual clones using a Wizard Plus SV miniprep DNA purification kit (Promega) and subjected to Sanger DNA sequencing at the University of Michigan DNA sequencing core facility.

For *in vitro* reactions (Figs. 2, 3, and S3G), equivalent sized gel slices that corresponded to the expected RT-PCR product size (~305 and ~232 base pairs for U6/L1 and U6/GFP, respectively) were excised from each lane regardless of whether a band was visible under UV illumination, cloned, and sequenced as described above.

Generation of synthetic U6, L1, and GFP RNA

To generate the synthetic U6 snRNA bearing a 2',3'-cyclic phosphate (U6>P), a double-stranded DNA template (gBlock Gene Fragments, IDT technologies, Coralville, IA) was designed that consisted of a T7 promoter joined to the human U6 snRNA sequence ending in 4 thymidines followed by the sequence of a mutant form of the hepatitis delta virus (HDV) antigenomic ribozyme sequence (T7-U6-HDVr) (10, 11) (SI Appendix, Fig. S2A and Tables S2 and S9). A control template lacking the HDV ribozyme sequence (T7-U6) was used to generate synthetic U6 snRNA bearing a 3'-OH (U6-OH) (SI Appendix, Table S9).

To generate a synthetic L1 RNA fragment with a 5'-OH (OH-L1), a double-stranded DNA template (gBlock Gene Fragments, IDT technologies) was designed that consisted of a T7 promoter followed by an engineered hammerhead ribozyme (HHr) sequence and pJM101/L1.3Δneo nucleotides 5752-6087 (T7-HHr-L1) (12) (SI Appendix, Table S9). A control L1 template was also made that lacked the HHr sequence (T7-L1) in order to generate an L1 RNA fragment bearing a 5'-triphosphate (P-L1) (SI Appendix, Table S9).

To generate a synthetic GFP RNA fragment with a 5'-OH (OH-GFP), a double-stranded DNA template (gBlock Gene Fragments, IDT technologies) was designed that consisted of a T7 promoter followed by an engineered hammerhead ribozyme sequence and pCEP-GFP (9) nucleotides 471-780 (T7-HHr-GFP) (12) (SI Appendix, Table S9). The GFP RNA sequence consists of nucleotides 471-720 of the hrGFP ORF sequence followed by 60 nucleotides of the SV40 polyadenylation sequence from the pCEP4 vector (SI Appendix, Table S9).

To generate synthetic RNAs, double-stranded DNA gBlock templates were first PCR amplified using Platinum *Taq* DNA Polymerase (Thermo Fisher Scientific). PCR-amplified templates were then purified from agarose gels using the Wizard SV Gel and PCR Clean-Up System (Promega). For in vitro transcription reactions, approximately 100-300 ng of template DNA was used in 40 µL reactions using a MAXIscript T7 transcription kit (Thermo Fisher Scientific). Reactions were incubated at 37°C for 2.5 hours and then treated with 4 µL of DNase I (Thermo Fisher Scientific) and concentrated using an RNA Clean & Concentrator kit (Zymo Research, Irvine, CA). RNA was eluted from columns with water and diluted to a final concentration of ~50-100 ng/µL and stored at -80°C.

Two microliters (~100-200 ng) of RNA from T7 transcription reactions was analyzed using denaturing Urea-PAGE. Gels were stained with SYBR Green II RNA gel stain (Thermo Fisher Scientific) to visualize the RNA.

U6/L1 RNA ligation reactions using purified RtcB

In vitro transcribed U6 and L1 RNAs were first splinted with a DNA oligonucleotide (SI Appendix, Table S9) by combining U6 and L1 RNA with the DNA oligonucleotide splint (~500 nM final concentration for each RNA and DNA oligo diluted in water; 10 µL final reaction volume). The RNA/DNA oligo mixture was then incubated at 65°C for 5 minutes, 25°C for 3 minutes, and then kept at 4°C for approximately 10 minutes before being added to the ligation reaction. Next, U6/L1 ligation reactions (4 µL final volume) containing 50 mM Tris-HCl (pH 8.0), 2 mM $MnCl_2$, 100 µM GTP, 2 µM purified RtcB from *E. coli* (13), and splinted U6 and L1 RNA substrates (~250 nM final concentration for each RNA and DNA oligo) were incubated at 37°C for 1 hour. The reactions were concentrated using an RNA Clean & Concentrator kit (Zymo Research), which included an on-column DNase I (Thermo Fisher Scientific) digestion. RNA was eluted with two volumes (8 µL each) of water.

Following ligation reactions, cDNAs were prepared using 3 µL of concentrated RNA with the SuperScript First-Strand Synthesis System for RT-PCR (Thermo Fisher Scientific) using the SV40as oligonucleotide primer. Reverse transcription (RT) reactions were incubated at 42°C for 50 minutes followed by an incubation at 70°C for 15 minutes. RT reactions (20 µL) were then diluted 1:1 with water to a final volume of 40 µL.

Following the RT step, nested PCR was then carried out using Platinum *Taq* DNA

Polymerase (Thermo Fisher Scientific) according to the manufacturer's protocol in 50 µL reactions using 2 µL of template cDNA from the above RT reactions, 0.25 µL Platinum *Taq*, 0.2 µM forward and reverse primers, 1.5 mM $MgCl_2$ and 0.2 mM dNTPs. The first round of PCR used the following primers: U6s1 (sense) and SV40as (antisense). The second round of PCR used: U6s2 (sense) and 3UTRas (antisense). Thermal cycler conditions were as follows: initial cycle of 94°C for 2 minutes, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 60 seconds at 72°C with a final cycle of 72°C for 5 minutes. PCR conditions were identical for each round of nested PCR. PCR products (10-20 µL) were visualized on 2% agarose gels using SYBR safe DNA gel stain (Thermo Fisher Scientific). For all reactions, gel slices were excised from each lane and processed for Sanger sequencing as described above.

U6/L1 RNA ligation reactions using HeLa cell extracts

HeLa cell nuclear extracts were either prepared from HeLa-JVM cells (14, 15) or purchased from Protein One (Rockville, MD, P0002-02). The nuclear extracts generated in the lab and the commercially sourced HeLa nuclear extracts both performed similarly in ligation reactions. U6/L1 ligation reactions (final volume: 4 µL) containing 2 µL (~10-40 µg) of nuclear extract, 50 mM Tris-HCl (pH 8.0), 2 mM $MnCl_2$, 100 µM GTP and RNA substrates (~250-500 nM final for each RNA) were incubated at 37°C for 1 hour. The reactions were then concentrated using an RNA Clean & Concentrator kit (Zymo Research), which included an on-column DNase I (Thermo Fisher Scientific) digestion. RNA was eluted with two volumes (8 µL each) of water.

Following ligation reactions, cDNAs were prepared using 3 µL of RNA with the SuperScript First-Strand Synthesis System for RT-PCR (Thermo Fisher Scientific) using the SV40as oligonucleotide primer. Reverse transcription (RT) reactions were incubated at 42°C for 50 minutes followed by heat inactivation at 70°C for 15 minutes. RT reactions (20 µL) were then diluted 1:1 with water to a final volume of 40 µL.

Following the RT step, PCR was then carried out using Platinum *Taq* DNA Polymerase (Thermo Fisher Scientific) according to the manufacturer protocol in 50 µL reactions using 2 µL of template cDNA from the above RT reactions, 0.25 µL Platinum *Taq*, 0.2 µM forward and reverse primers, 1.5 mM $MgCl_2$ and 0.2 mM dNTPs. The first round of PCR used the following primers: U6s1 (sense) and SV40as (antisense). The second round of PCR used: U6s2 (sense) and 3UTRas (antisense). Thermal cycler conditions were as follows: initial cycle of 94°C for 2 minutes, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 60 seconds at 72°C with a final cycle of 72°C for 5 minutes. PCR conditions were identical for each round of nested PCR. PCR products (10-20 µL) were visualized on 2% agarose gels using SYBR safe

DNA gel stain (Thermo Fisher Scientific). For all reactions, gel slices were excised from each lane and processed for Sanger sequencing as described above.

CRISPR/Cas9 depletion of RtcB from HeLa cells

To deplete RtcB protein expression in HeLa-JVM cells, a single guide RNA targeting exon 2 of human RtcB (SI Appendix, Table S9) (16) was cloned into the *Bbs*I site of the pX459v2 plasmid vector (17). As a control, an sgRNA targeting GFP (SI Appendix, Table S9) was also cloned into pX459v2. Approximately $5x10^5$ HeLa-JVM cells were then transfected in 6-well plates using 6 µL of FuGENE HD and 2 µg of plasmid DNA per well. Approximately 24 hours later, transfections were stopped by the addition of fresh media to the cells. Approximately 48 hours after transfection, media was supplemented with puromycin (5 µg/mL) to select for transfected cells. Selection media was refreshed every two days thereafter. Six days after transfection, the cells were removed from puromycin selection and reseeded in 96-well plates to select for individual clones by adding 100 µL of a cell suspension (~10-20 cells/mL) to each well of a 96-well plate. Approximately 10-14 days later, 96-well plates were screened using light microscopy for wells containing a single colony. Clones were isolated from individual wells by trypsinization and transferred to 12-well plates. Clones were expanded and then screened for RtcB expression by western blotting. Sanger sequencing was used to characterize genomic RtcB edits.

Reverse transcription – quantitative real-time PCR (RT-qPCR) U6/L1 ligation assay

HeLa cell nuclear extracts were prepared as described above. U6/L1 ligation reactions (final volume: 6 µL) containing ~10 µg of nuclear extract, 25 mM Tris-HCl (pH 8.0), 1 mM $MnCl_2$, 200 µM GTP and RNA substrates (~250-500 nM final for each RNA) were incubated at 37°C for 1 hour. The reactions were concentrated using an RNA Clean & Concentrator kit (Zymo Research), which included an on-column DNase I (Thermo Fisher Scientific) digestion. RNA was eluted with two volumes (8 µL each) of water.

Following ligation reactions, cDNAs were prepared using 3 µL of RNA with the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen) with the SV40as oligonucleotide primer. Reverse transcription (RT) reactions were incubated at 42°C for 50 minutes followed by an incubation at 70°C for 15 minutes. RT reactions (20 µL) were then diluted 1:1000 with water for use in RT-qPCR reactions.

Following cDNA synthesis, RT-qPCR reactions (20 µL total volume) were carried out according to the manufacturer's protocol in triplicate for each condition using the PowerUp

SYBR Green Master Mix (Thermo Fisher Scientific) by combining 10 µL PowerUp SYBR Green Master Mix, forward and reverse primers (5 µM each, SI Appendix, Table S9), and 5 µL of diluted cDNA from above. Two sets of primers were used (SI Appendix, Table S9): the target primer pair (U6L1_qPCR_1F and U6L1_qPCR_1R) amplifies a 118 bp sequence spanning the U6/L1 junction sequence; the control primer pair (U6L1_qPCRcon_4F and U6L1_qPCRcon_4R) amplifies a 122 bp sequence at the 3' end the L1 RNA template and serves as an endogenous control to normalize total cDNA input in each reaction. RT-qPCR was carried out an ABI 7300 Real-Time PCR system (Thermo Fisher Scientific) using following thermal cycling conditions: initial cycle of 50°C for 2 minutes; 95°C for 2 minutes; followed by 40 cycles of 15 seconds at 95°C, 28 seconds at 54°C, and 60 seconds at 72°C.

The relative standard curve method was used to quantify U6/L1 ligation efficiency. Standard curves for each primer pair were generated using serial 10-fold dilutions of a 402 bp U6/L1 double stranded DNA template (U6L1_qPCR_standard) consisting of U6 snRNA nucleotides (41-106) conjoined to pJM101/L1.3Δneo nucleotides 5752-6087 (SI Appendix, Table S9; concentration range: 1×10^{-12} M – 1×10^{-16} M). U6/L1 ligation efficiency was determined by the ratio of U6/L1 junction molecules over L1 endogenous control molecules. For each experiment, untransfected HeLa-JVM nuclear extracts were used as a calibrator and each sgRNA condition were considered different treatments. Thus, U6/L1 ligation efficiency was normalized to untransfected HeLa-JVM cell extracts. The normalized ligation efficiency for each reaction condition was calculated by averaging the values from 6 independent RT-qPCR experiments. A two-tailed Student's t-test was used to determine *p* values.

Western Blotting

Standard western blotting procedures were used for protein analysis. Blots were analyzed using an Odyssey CLx (LI-COR, Lincoln, NE). Western blot quantification was performed using the Image Studio software (version 3.1.4, LI-COR). The following antibodies were used: anti-RtcB/C22orf28/FAAP (1:5000) (Bethyl Laboratories, Montgomery, TX, A305-079A), anti-eIF3 (p110) (1:2000) (Santa Cruz Biotechnology, Dallas, TX, sc-28858), anti-nucleolin (1:1000) (Cell Signaling Technology, Danvers, MA, #87792), IRDye 800CW Donkey anti-Rabbit IgG (1:10,000) (LI-COR, 925-32213) and IRDye 680RD Donkey anti-Mouse IgG (1:10,000) (LI-COR, 925-68072).

U6/GFP RNA ligation reactions using HeLa cell extracts

HeLa cell nuclear extracts were prepared as described above. U6/L1 ligation reactions (final volume: 4 µL) containing 2 µL (~10-40 µg) of nuclear extract, 50 mM Tris-HCl (pH 8.0), 2 mM $MnCl_2$, 100 µM GTP and RNA substrates (~250-500 nM final for each RNA) were incubated at 37°C for 1 hour. The reactions were concentrated using an RNA Clean & Concentrator kit (Zymo Research), which included an on-column DNase I (Thermo Fisher Scientific) digestion. RNA was eluted with two volumes (8 µL each) of water.

Following ligation reactions, cDNAs were prepared using 3 µL of RNA with the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen) with the SV40as oligonucleotide primer. Reverse transcription (RT) reactions were incubated at 42°C for 50 minutes followed by an incubation at 70°C for 15 minutes. RT reactions (20 µL) were then diluted 1:1 with water to a final volume of 40 µL.

Following the RT step, nested PCR was carried out using Platinum *Taq* DNA Polymerase (Thermo Fisher Scientific) according to the manufacturer's protocol in 50 µL reactions using 2 µL of template cDNA from the above RT reactions, 0.25 µL Platinum *Taq*, 0.2 µM forward and reverse primers, 1.5 mM $MgCl_2$ and 0.2 mM dNTPs. The first round of PCR used the following primers: U6s1 (sense) and hrGFPas1 (antisense). The second round of PCR used: U6s2 (sense) and hrGFPas2 (antisense). Thermal cycler conditions were as follows: initial cycle of 94°C for 2 minutes, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 60 seconds at 72°C with a final cycle of 72°C for 5 minutes. PCR conditions were identical for each round of nested PCR. PCR products (10-20 µL) were visualized on 2% agarose gels using SYBR safe DNA gel stain (Thermo Fisher Scientific). For all reactions, gel slices were excised from each lane and processed for Sanger sequencing as described above.

RNA-sequencing (RNA-seq)

All cDNA library preparation and sequencing was conducted at the University of Michigan sequencing core facility (Ann Arbor, MI). Briefly, total RNA was collected from HeLa-JVM, HeLa-HA, and PA-1 cells using an RNeasy mini kit (Qiagen) according to the manufacturer's instructions. Total RNA from hESC (4, 5), and hESC derived NPCs (6) was a generous gift of Dr. Jose Garcia-Perez. To generate cDNA libraries, total RNA from each cell line was first depleted of ribosomal RNA using a Ribo-Zero rRNA removal kit (Illumina, San Diego, CA), and then cDNA libraries were generated from the rRNA-depleted RNA using the TruSeq Stranded mRNA Library Prep Kit (Illumina) with random hexamers according to manufacturer protocol with the following deviations: RNA was fragmented for 1 minute to generate ~190 nucleotide fragments and 12 PCR cycles were used to enrich DNA fragments after ligating adapters. Paired-end sequencing (100 bp reads) was performed on the Illumina HiSeq 2500. RNA-seq data for PA-1, H9, and NPCs has previously been deposited to the Sequence Read Archive (SRA: PRJNA432733) (18). HeLa RNA-seq data has previously been deposited to dbGaP (dbGaP: phs00167) (18).

RNA Sequencing Analysis

Trimmomatic (19) was used to trim the sequencing adaptors from a total of ~1.1 X 10^9 RNA sequencing reads. We assessed the quality of our data using FastQC (Andrews S. [2010]. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Samtools rmdup (20) and Picard MarkDuplicates (http://broadinstitute.github.io/picard) were used to remove PCR duplicate reads. We aligned all reads that passed the quality check with BWA-MEM with default parameters (Li H. [2013] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]) to a custom built human reference genome from hg38 with all repeats masked using RepeatMasker and RepBase (21), but including a single representative copy of a human specific L1 (L1.3; GenBank accession no. L19088) (7) and human U6 snRNA (GenBank accession no. X59362). FLASH (22) then was used to reconstruct overlapping read pairs that aligned at one end to the 3' portion of U6 snRNA and the other end to L1. Merged U6/L1 sequences that contained U6 snRNA sequence at the 5' end conjoined to L1 sequence at the 3' end were then mapped back to the non-masked HGR (HGR/build Grch38) using BWA-MEM in order to differentiate events aligned to the genome from those which did not exhibit a clear mapping (Fig. 4). Our software for extracting these fusion reads from RNA-seq data can be found at https://github.com/mills-lab/U6L1. All U6/L1 reads were manually aligned to the HGR using BLAT (23) to verify BWA-MEM alignments. The L1 portion of each U6/L1 read was manually aligned to the L1.3 sequence and consensus sequences from L1 subfamilies (L1PA1-L1PA13) (24) to determine the L1 subfamily and to derive L1 sequences for L1 junction analyses (Figs. S4A and S4B; Tables S4 and S5).

U6/L1 Junction Motif Search of HeLa cells and 1000 Genomes Project High Coverage Samples

Motifs across putative U6/L1 junctions were extracted from all merged reads as described above. Each 25 base pair junction motif contains U6 snRNA nucleotides 94-102 followed by 5-8 thymidines and ~8-11 nucleotides of L1 sequence (SI Appendix, Table S6). All motifs and their reverse complements were used to interrogate HeLa cell genomic data from dbGaP (dbGaP accession number phs000640.v1.p1) (25-27) and 23 high coverage PCR-free DNA sequencing samples from the 1000 Genomes Project (SI Appendix, Table S9) (28) to look for genomic evidence of each U6/L1 junction sequence. The script for 25 base pair motif search is available at: https://github.com/mills-lab/U6L1. An exact match was required for labeling the existence of the junction from the HeLa genomic and 1000 Genomes DNA sequencing data. Two exceptions were noted in the 1000 genomes data, in which two genomes (NA20845 and HG03742) contained the same SNP within the U6 sequence for the U6/L1 chimera sequence with L1.3 junction 2052 and therefore did not initially exhibit an exact match to the genomic sequences of these samples (see Results and SI Appendix, Table S6).

HeLa cell genome sequence data

The HeLa cell genome sequence data used for analysis described in this manuscript were obtained from the database of Genotypes and Phenotypes (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000640.v5.p1). This study was reviewed by the NIH HeLa Genome Data Access Working Group.

FIGURE LEGENDS

Fig. S1. Related to Fig. 1. *A. A representative agarose gel image of control RT-PCR reactions lacking reverse transcriptase.* The transfected L1 construct is indicated above each lane of the agarose gel image. HeLa UTF: untransfected HeLa cells. Molecular weight standards (in bp) are shown in the first and last gel lanes. *B. Weblogo frequency plot of the 38 U6/L1 chimeric junction sequences obtained from RT-PCR experiments.* The X-axis indicates the L1.3 nucleotide positions upstream (negative numbers) or downstream (positive numbers) of the U6/L1 junction. The Y-axis indicates nucleotide frequency. The arrow indicates where the U6 thymidine tract (white block arrow ending in T_n) is conjoined to the L1 sequence. For weblogo analyses, 5Ts were assigned to U6 and the remaining T's at the U6/L1 junction were assigned to L1. *C. Examples of putative RT-PCR sequence artifacts.* The dashed line with arrows indicates the approximate position of a putative template-switching event from L1 RNA (white numbers in L1 correspond to the sequence of L1.3 (GenBank accession no. L19088) (7)) to U6 snRNA (GenBank accession no. X59362; black numbers below U6 arrow). The U6/L1 junction sequence is indicated below the L1. Red nucleotides align to U6. Black nucleotides align to L1. The top example depicts a putative template-switching event from L1 to U6 that is mediated by a short region of microhomology [parentheses (aa); purple text]. The bottom example depicts a putative microhomology-independent template-switching event.

Fig. S2. Related to Fig. 2. *A. Rationale of the in vitro transcription reaction to generate U6 RNA ending in a* 2',3'-cyclic phosphate. PCR was used to amplify a double-stranded DNA template that consisted of a T7 promoter (grey rectangle) upstream of the human U6 snRNA cDNA sequence (white rectangle) ending in four thymidine nucleotides that was immediately followed by the hepatitis delta virus (HDV) antigenomic ribozyme sequence (black rectangle), creating the T7-U6-HDVr transcription template. During *in vitro* transcription, the HDV ribozyme liberates itself from the transcript generating a 2',3'-cyclic phosphate at the end of 3' end of U6 RNA (>P, red circle). *In vitro* transcription reactions were analyzed using denaturing PAGE to confirm HDV ribozyme cleavage. Red arrow: U6 RNA. Blue arrow: the liberated HDV ribozyme. *B. Rationale of the in vitro transcription reaction to generate the 5'-OH-L1 RNA.* PCR was used to amplify a double-stranded DNA template that consisted of a T7 promoter (grey rectangle) followed by an engineered hammerhead ribozyme sequence (light blue rectangle) upstream of an L1 fragment (nucleotides 5752-6081) derived from pJM101/L1.3Δneo (dark blue rectangle), creating the T7-r-L1 transcription template. During *in vitro* transcription, the HHr liberates itself from the transcript generating a 5'-OH at the 5' end of the L1 RNA (OH, black circle). *In vitro* transcription reactions were analyzed using denaturing PAGE to confirm HHr cleavage. Red arrow: OH-L1 RNA. Blue arrow: the liberated HHr ribozyme. _C. Examples of putative RT-PCR sequence artifacts_. Depicted are schematics of the U6 (white block arrow ending in four thymidine nucleotides) and L1 RNA (blue rectangle) cDNA sequences. The dashed line with arrows indicates the approximate position of a putative template-switching event from L1 RNA (white numbers in L1 correspond to the sequence of L1.3) to U6 RNA (black numbers correspond to the sequence of U6 snRNA). The top example depicts the addition of an untemplated nucleotide (green capitalized A) between the U6 (red font) and L1 (black font) sequences. The middle example depicts a putative microhomology-independent template-switching event. The bottom example depicts a putative template-switching event from L1 to U6 that is mediated by a short region of microhomology (parentheses, cg, purple letters). Black numbers: U6 nucleotide junction positions. White numbers: L1 nucleotide junction positions.

Fig. S3. Related to Fig. 3. _A. Western blots demonstrate the presence of RtcB in HeLa cell nuclear extracts._ Representative images of Western blots using HeLa cell nuclear extracts that either were produced in our lab (left blot) or purchased from a commercially source (right blot). Primary antibodies indicated to the right of blot. The presence of eIF3C (red arrow) indicates that nuclear extracts may also contain some cytosolic content. The green arrow points to the band corresponding to the approximate molecular size of RtcB (~55.2 kDa). _B. Examples of putative RT-PCR sequence artifacts_. Depicted are schematics of the U6 (white block arrow ending in four thymidine residues) and L1 RNA (blue rectangle) sequences. The dashed line with arrows indicates the approximate position of a putative template switch from L1 RNA (white numbers in L1; corresponding to the sequence of L1.3 (GenBank accession no. L19088) (7)) to U6 RNA (GenBank accession no. X59362; black numbers below U6 arrow). The top example depicts the addition of untemplated nucleotides (green capitalized 5'-AAGG) between the U6 (red font) and L1 (black font) sequences. The middle example depicts a putative template-switching event from L1 to U6 that is mediated by a short region of microhomology (parentheses, atat, purple text). The bottom example depicts a putative microhomology-independent template-switching event. Black numbers: U6 nucleotide junction positions. White numbers: L1 nucleotide junction positions. _C. Weblogo frequency plot of the 53 U6/L1 chimeric junction sequences obtained from RT-PCR experiments using HeLa cell nuclear extracts._ The X-axis indicates the L1 nucleotide positions upstream (-) or downstream (+) of the U6/L1 junction. The Y-axis indicates nucleotide frequency. The arrow indicates where the U6 thymidine stretch (white block arrow ending in T_n) becomes conjoined the L1 sequence.

Sequences are based on L1.3 (GenBank accession no. L19088) (7). For weblogo analyses, 4 templated Ts were assigned to U6 and the remaining T's at the U6/L1 junction were assigned to L1. *D. Characterization of genomic RtcB edits in $RtcB^{2.1}$ and $RtcB^{2.2}$ HeLa cell lines.* Each edited HeLa cell line ($RtcB^{2.1}$ and $RtcB^{2.2}$) contained three edited RtcB alleles and no wild-type RtcB alleles. The local genomic RtcB sequence near the sgRNA target sequence within RtcB exon 2 is shown for each edited RtcB allele. The sgRNA target sequence within exon 2 is in red lettering; RtcB exon 2 sequence is in "UPPERCASE" letters; and RtcB intron 3 sequence is in "lowercase" letters. A dash (-) indicates deleted nucleotides and blue lettering indicates insertions. Sanger sequencing revealed the presence of three edited RtcB alleles in each cell line. Wild type genomic RtcB sequences were not detected in either clone. Numbers adjacent to the sequences indicate the number of nucleotides that were deleted from (Δ) or inserted into (+) the wild type RtcB sequence. Deletions or insertions of an even number of nucleotides were predicted to cause a frameshift mutation that would result in premature translation termination and a severely truncated RtcB polypeptide chain. The 21 bp deletion allele of $RtcB^{2.1}$ is predicted to result in an amino acid substitution (R51S) and a deletion of amino acid residues N52-G58 from the wild type RtcB protein sequence. The 3 bp deletion allele of $RtcB^{2.2}$ is predicted to result in the deletion of amino acid residue C54 from the wild type RtcB protein sequence. *E. HeLa cell nuclear extracts mediate the ligation of U6 and to GFP RNAs.* A synthetic human U6 RNA a 2',3'-cyclic phosphate (>P, red circle) and a synthetic GFP RNA (green rectangle) containing a 5'-OH (black circle) were generated using a ribozyme-based *in vitro* transcription reaction. The resultant RNAs were incubated with HeLa cell nuclear extracts as described in Fig. 4A and cDNAs were synthesized using the SV40as oligonucleotide primer. RT-PCR reactions using nested primers (U6s1 and GFPas1, then U6s2 and GFPas2) were used to detect U6/GFP chimeric cDNAs. *F. Schematic representations of the synthetic RNAs used in in vitro experiments.* The *in vitro* transcribed GFP sequence consists of nucleotides 471-720 of the hrGFP ORF sequence followed by 60 nucleotides of the SV40 polyadenylation sequence from the pCEP4 vector and contains a 5'-OH (OH-GFP). The *in vitro* transcribed U6 RNA ends in four uridine ribonucleotides and contains a 2',3'-cyclic phosphate (U6>P) or a 3'-OH (U6-OH). *G. Results from the ligation reactions.* The constituents of U6/GFP ligation reactions are indicated above each gel lane (+) of the representative agarose gel image. An asterisk (*) indicates that the HeLa cell nuclear extract was heat treated at 95°C for 10 minutes prior to adding it to the reaction. No RT: no RT control. H_2O: water PCR controls. DNA size markers (in bp) are shown to the left of the gel image. The predicted position of the 232 bp U6/GFP RT-PCR product is noted on the left side of the gel image (white arrow, green font).

Bands in the reactions either lacking or containing heat inactivated HeLa cell nuclear extracts are non-specific products. *H. Summary of results from product characterization experiments.* Column 1: RNAs used in the reaction. Column 2: number of RT-PCR products characterized for the reaction condition. Column 3: number of RT-PCR products that correspond to the full-length U6/GFP ligation product. Column 4: number of RT-PCR products that contain a variably 5'-truncated OH-GFP. Column 5: number of putative RT-PCR artifact products. Each experiment was repeated three times and yielded similar results. *I. Protein sequence alignments of RtcB from various species.* RtcB protein sequence alignments carried out using the align tool on the UniProt website (29).

Fig. S4. Related to Fig. 4. *A. Weblogo frequency plot of the 16 "aligned" U6/L1 chimeric junction sequences obtained from RNA-seq experiments.* The X-axis indicates the L1 nucleotide positions residing either upstream (negative numbers) or downstream (positive numbers) of the U6/L1 junction sequence. The Y-axis indicates nucleotide frequency. The arrow indicates where the U6 thymidine stretch (white block arrow ending in T_n) is conjoined to the L1 sequence. Sequences are based on L1.3. The sequences used to generate these plots are depicted in SI Appendix, Table S4. For weblogo analyses, 5Ts were assigned to U6 and the remaining T's at the U6/L1 junction were assigned to L1. *B. Weblogo frequency plot of the 33 "non-aligned" U6/L1 chimeric junction sequences obtained from RNA-seq experiments.* The X- and Y-axis are the same as indicated in panel A. The arrow indicates where the U6 thymidine tract (white block arrow ending in T_n) is conjoined to the L1 sequence. Sequences are based on L1.3. The sequences used to generate these plots are depicted in SI Appendix, Table S5. For weblogo analyses, 5Ts were assigned to U6 and the remaining T's at the U6/L1 junction were assigned to L1.

TABLE LEGENDS

Table S1. Related to Fig. 1. Analysis of U6/L1 chimeric RNA junctions from engineered human L1s. Column 1: name of the transfected L1 plasmid. Column 2: the position of the U6/L1 junction sequence; the numbers reference the sequence position in L1.3. Column 3: The L1.3 sequence 20 bp upstream of the U6/L1 junction. Please note that this sequence is not present in the U6/L1 chimeric cDNA. Column 4: The number of thymidine nucleotides at the end of the U6 snRNA cDNA sequence. The numbers in parenthesis reflect ambiguities where thymidine nucleotides also are present at the 5' end of the L1 sequence. Column 5: The L1.3 sequence conjoined to the U6 thymidine tract. Underlining highlights the ambiguous thymidine nucleotides in the downstream L1 sequence.

Table S2. Related to Fig. 2. Analysis of U6/L1 chimeras containing 5'-truncated L1s. Column 1: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 2: The L1.3 sequence 20 bp upstream of the U6/L1 junction. Please note: this sequence is not present in the U6/L1 chimeric cDNA. Some sequences will contain less than 20 bp because the junction is less than 20 bp from the 5' end of the L1 oligonucleotide. Column 3: The number of thymidine nucleotides at the end of the U6 sequence. Column 4: The L1 sequence immediately conjoined to the U6 thymidine stretch.

Table S3. Related to Fig. 3. Analysis of U6/L1 chimeras containing 5'-truncated L1s. Column 1: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 2: The sequence from the 20 bp upstream of the U6/L1 junction. Please note: this sequence is not present in the U6/L1 chimeric cDNA. Some sequences will contain less than 20 bp because the junction is less than 20 bp from the 5' end of the L1 oligo. Column 3: The number of thymidine residues at the end of the U6 sequence. Column 4: The L1 sequence immediately conjoined to the U6 thymidine stretch.

Table S4. Related to Fig. 4. Analysis of "aligned" U6/L1 sequences from RNA-seq experiments. Column 1: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 2: The subfamily designation of the "aligned" L1 sequence in the human genome reference sequence. Column 3: The L1.3 sequence 20 bp upstream of the U6/L1 junction. Please note that this sequence is not present in the U6/L1 chimeric cDNA. Column 4: The number of thymidine nucleotides at the end of the U6 cDNA sequence. The numbers in parenthesis (n) reflect ambiguities where thymidine nucleotides also are present at the 5' end of the L1 sequence. Column 5: The L1.3 sequence immediately downstream (+) of the U6 thymidine tract. Underlining highlights the ambiguous thymidine nucleotides in the downstream L1 sequence.

Table S5. Related to Fig. 4. Analysis of "non-aligned" U6/L1 sequences from RNA-seq experiments. Column 1: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 2: The L1.3 sequence 20 bp upstream of the U6/L1 junction. Please note that this sequence is not present in the U6/L1 chimeric cDNA. Column 3: The number of thymidine nucleotides at the end of the U6 sequence. The numbers in parenthesis reflect ambiguities where thymidine nucleotides also are present at the 5' end of the L1 sequence. Column 4: The L1.3 sequence immediately downstream of the U6 thymidine tract. Underlining highlights the ambiguous thymidine nucleotides in the downstream L1 sequence. The asterisk (*) indicates the L1 sequence was likely from the L1PA5 subfamily. The double asterisk (**) indicates the L1 was likely from the L1PA4 subfamily. These designations are based on the alignment of the L1 sequence to L1 subfamily consensus sequences (see Methods).

Table S6. Related to Fig. 4. Sequences features of the 25bp U6/L1 junction sequences motifs of the "aligned", "non-aligned", and putative "artifact" RNA-seq chimeras. Column 1: 25bp junction motifs are numbered sequentially from 1 to 64. Column 2: Indicates whether the 25 bp junction motif is from an "aligned", "non-aligned", or "artifact" RNA-seq chimera. Column 3: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 4: the U6/L1 merged junction sequence reads used as probes to search the 1000 Genomes Project sequencing data. Thymidine residues at the junction are underlined. Column 5: Numbers of reads that support the junction sequences. Column 6: the cell line containing the U6/L1 junction sequences, "multiple" indicates that the U6/L1 junction sequence was detected in more than one cell line. (*) Indicates that the U6/L1 junction sequence contains a SNP in HG03742 (INDIAN TELUGU) and NA20845 (GUJARATI INDIAN) 1000 Genomes project sample genomes: 5'-CATTATGTATTTTTAATTAAAAGAC (SNP is underlined). (**) Indicates that for this U6/L1 junction the L1 sequence is antisense compared to U6.

Table S7. Related to Fig. 4. Characterization of 16 genomic U6/L1 chimeric pseudogenes that served as putative source elements for the RNA-seq reads detected in Supplemental Table 4. Column 1: The nucleotide position in L1 conjoined to U6 RNA. The numbering is based upon the reference sequence of L1.3. Column 2: Genome coordinates based on HGR/Grch38. Column 3: The subfamily designation of the “aligned” L1 sequence in the human genome reference sequence. Column 4: The length (number of nucleotides) of target site duplications (TSD) that flank the U6/L1 insertion. A “-“ sign indicates instances where we could not identify a TSD, Column 5: The putative L1 EN cleavage site. The “/” indicates the site of the EN cleavage. Column 6: Remarks indicate the genomic context of U6/L1 insertion.

Table S8. Related to Fig. 4. 1000 Genomes Project sample numbers with population codes. Column 1: the 1000 Genome Project sample number. Column 2: the population code for a given sample.

Table S9. Oligonucleotides used in this study. Name of the oligonucleotide. Column 2: The oligonucleotide sequence. Underlining indicates the T7 RNA polymerase promoter sequence used to transcribe RNAs *in vitro*. Lower-case letters indicate that HDV and HHr ribozyme sequences, respectively.

REFERENCES

1. Moran JV*, et al.* (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* 87(5):917-927.
2. Hulme AE, Bogerd HP, Cullen BR, & Moran JV (2007) Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene* 390(1-2):199-205.
3. Garcia-Perez JL*, et al.* (2010) Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* 466(7307):769-773.
4. Garcia-Perez JL*, et al.* (2007) LINE-1 retrotransposition in human embryonic stem cells. *Human Molecular Genetics* 16(13):1569-1577.
5. Macia A*, et al.* (2011) Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol* 31(2):300-316.
6. Coufal NG*, et al.* (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460(7259):1127-1131.
7. Sassaman DM*, et al.* (1997) Many human L1 elements are capable of retrotransposition. *Nat Genet* 16(1):37-43.
8. Wei W*, et al.* (2001) Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* 21(4):1429-1439.
9. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, & Moran JV (2006) Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* 20(2):210-224.
10. Been MD, Perrotta AT, & Rosenstein SP (1992) Secondary structure of the self-cleaving RNA of hepatitis delta virus: applications to catalytic RNA design. *Biochemistry* 31(47):11843-11852.
11. Schurer H, Lang K, Schuster J, & Morl M (2002) A universal method to produce in vitro transcripts with homogeneous 3' ends. *Nucleic Acids Res* 30(12):e56.
12. Avis JM, Conn GL, & Walker SC (2012) Cis-acting ribozymes for the production of RNA in vitro transcripts with defined 5' and 3' ends. *Methods Mol Biol* 941:83-98.
13. Tanaka N & Shuman S (2011) RtcB is the RNA ligase component of an Escherichia coli RNA repair operon. *The Journal of biological chemistry* 286(10):7727-7731.
14. Folco EG, Lei H, Hsu JL, & Reed R (2012) Small-scale nuclear extracts for functional assays of gene-expression machineries. *J Vis Exp* (64).
15. Dignam JD, Lebovitz RM, & Roeder RG (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* 11(5):1475-1489.
16. Doench JG*, et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 34(2):184-191.
17. Ran FA*, et al.* (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8(11):2281-2308.
18. Flasch DA*, et al.* (2019) Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* 177(4):837-851 e828.
19. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.

20. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

21. Jurka J, *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462-467.

22. Magoc T & Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957-2963.

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

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

25. Mailman MD, *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 39(10):1181-1186.

26. Landry JJ, *et al.* (2013) The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* 3(8):1213-1224.

27. Adey A, *et al.* (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* 500(7461):207-211.

28. Genomes Project C, *et al.* (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

29. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46(5):2699.

Figure S1

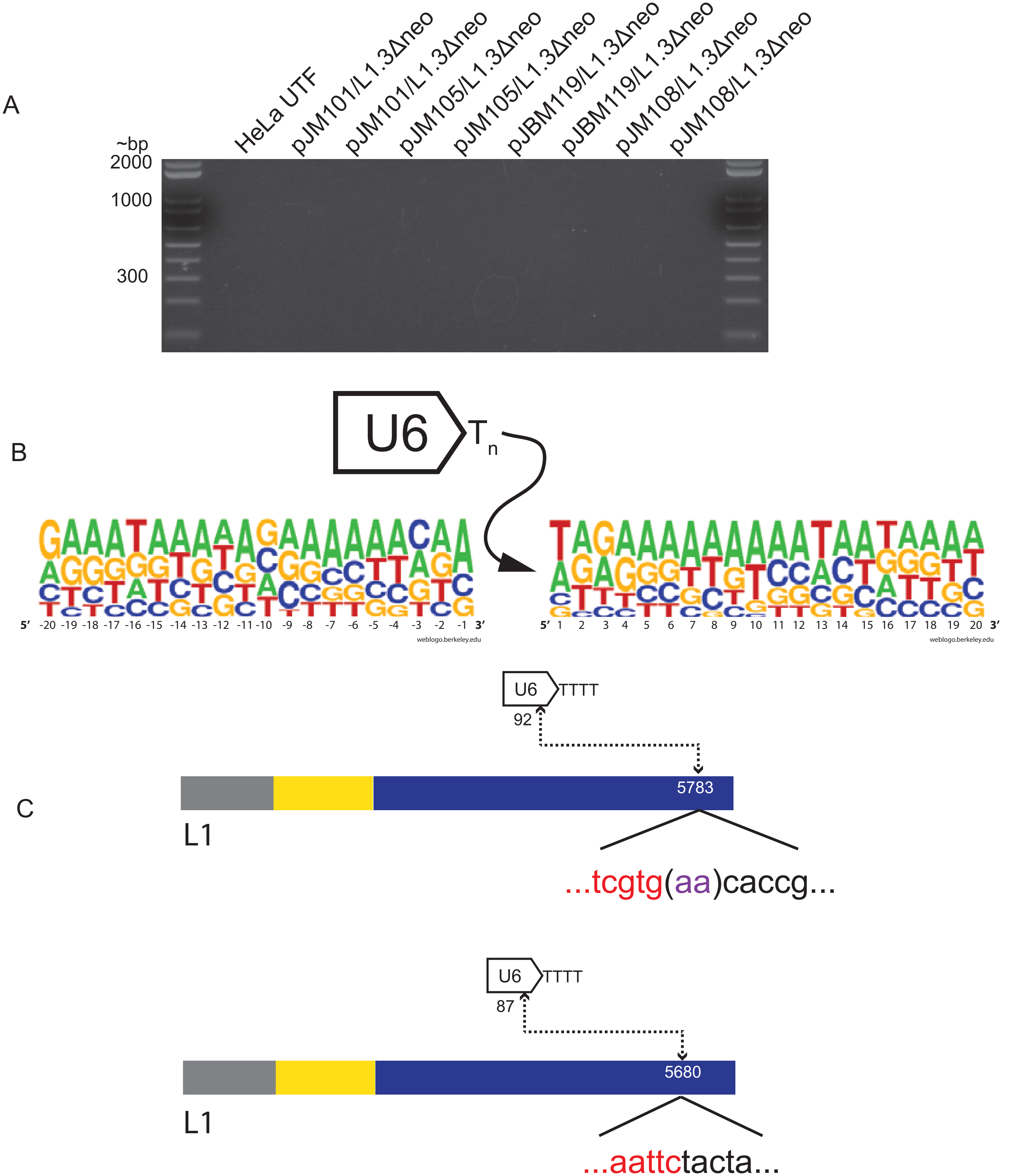
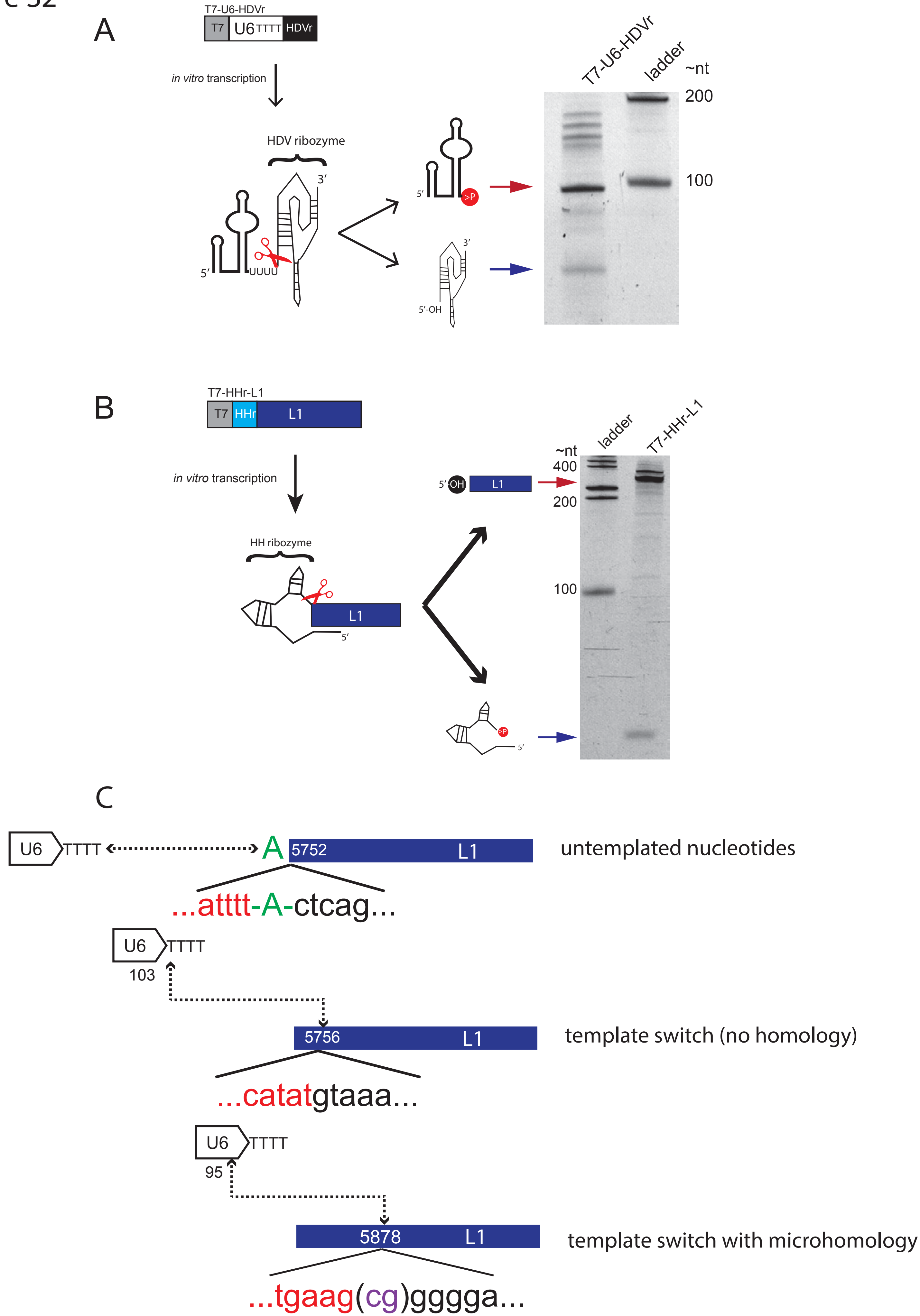
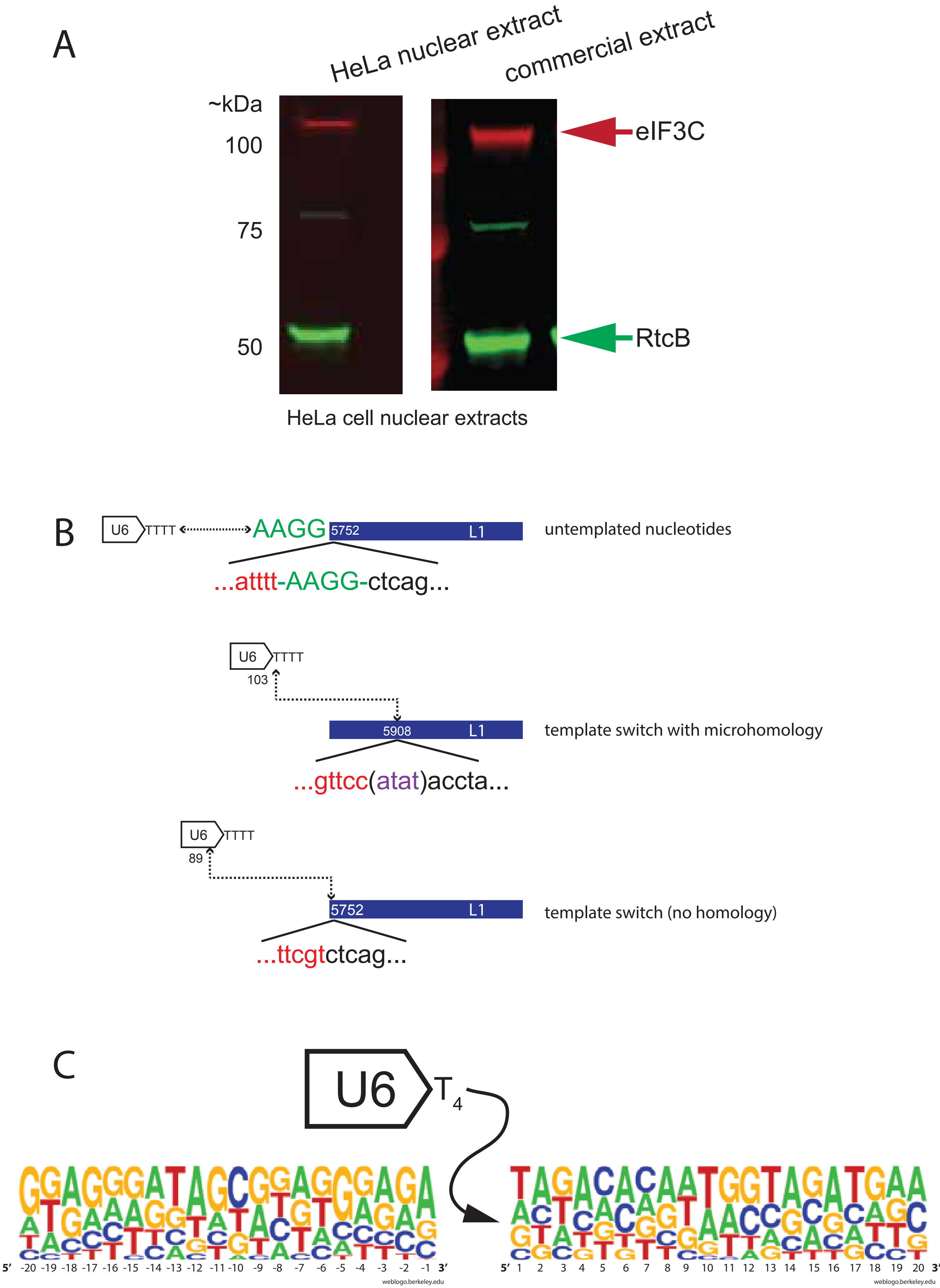

A
HeLa UTF
pJM101/L1.3Δneo
pJM101/L1.3Δneo
pJM105/L1.3Δneo
pJM105/L1.3Δneo
pJBM119/L1.3Δneo
pJBM119/L1.3Δneo
pJM108/L1.3Δneo
pJM108/L1.3Δneo
~bp
2000
1000
300
B
U6
Tn
5′ -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 3′
5′ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 3′
weblogo.berkeley.edu
C
U6
TTTT
92
5783
L1
...tcgtg(aa)caccg...
U6
TTTT
87
5680
L1
...aattctacta...

Figure S2



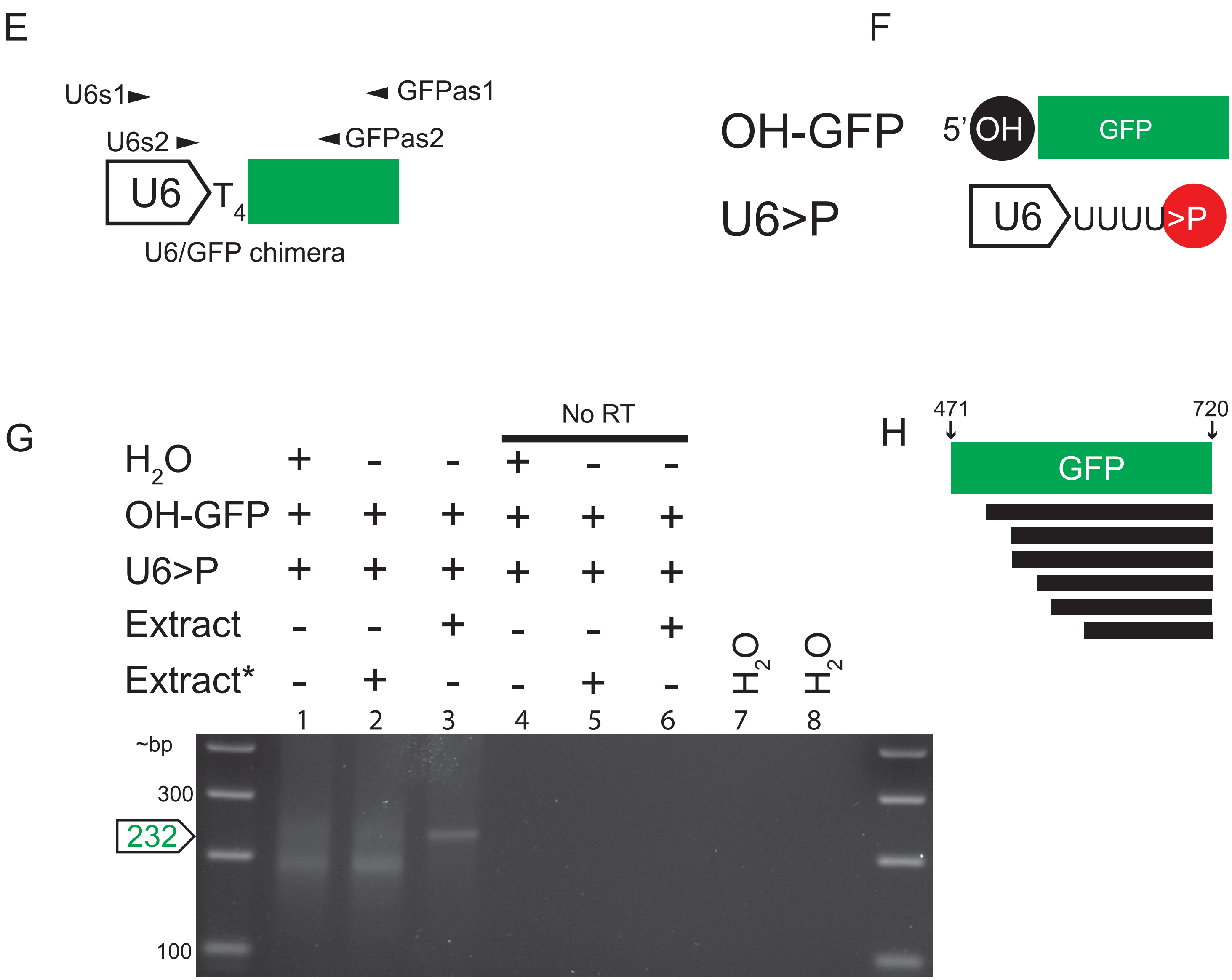
A
T7-U6-HDVr
T7
U6
TTTT
HDVr
in vitro transcription
HDV ribozyme
3′
5′
UUUU
5′
>P
3′
5′-OH
T7-U6-HDVr
ladder
~nt
200
100
B
T7-HHr-L1
T7
HHr
L1
in vitro transcription
HH ribozyme
L1
5′
5′-OH
L1
>P
5′
ladder
T7-HHr-L1
~nt
400
200
100
C
U6
TTTT
A
5752
L1
untemplated nucleotides
...atttt-A-ctcag...
U6
TTTT
103
5756
L1
template switch (no homology)
...catatgtaaa...
U6
TTTT
95
5878
L1
template switch with microhomology
...tgaag(cg)gggga...

Figure S3


A
HeLa nuclear extract
commercial extract
~kDa
100
75
50
eIF3C
RtcB
HeLa cell nuclear extracts
B
U6
TTTT
AAGG
5752
L1
untemplated nucleotides
...atttt-AAGG-ctcag...
U6
TTTT
103
5908
L1
template switch with microhomology
...gttcc(atat)accta...
U6
TTTT
89
5752
L1
template switch (no homology)
...ttcgtctcag...
C
U6
T4
5′ -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 3′
weblogo.berkeley.edu
5′ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 3′
weblogo.berkeley.edu

D

HeLa clone	genomic RtcB (exon 2) DNA sequence context:	# nucleotides deleted (Δ) or inserted (+):	Predicted change to RtcB amino acid sequence:
RtcB [2.1]	TTTGA----------------------GGTGGTGgtaagtacat	Δ22	Frameshift
	TTTGAGGAATTAAG---------------------taagtacat	Δ21	R51S, ΔN52-G58
	TTTGAGGAATTAAGG--------------TGGTGgtaagtacat	Δ14	Frameshift
RtcB [2.2]	TTTGAGGAATTAAG-------------------------tacat	Δ25	Frameshift
	TTTGAGGAATTAAGGAATGC---TCGAGGTGGTGgtaagtacat	Δ3	ΔC54
	TTTGAGGAATTAAGGAATGCCTGTGTCGAGGTGGTGgtaagtacat	+2	Frameshift



	Total	Full-length (U6 T4 GFP)	5'- truncated (U6 T4 GFP)	artifact
U6>P OH-L1	23	9	6	8

CLUSTAL O(1.2.4) multiple sequence alignment

```
SP|Q9Y3I0|RTCB_HUMAN MSRSYNDELQFLEKINKNCWRIKKGFVPNMQVEGVFYVNDALEKLMFEELRNACRGGGVG 60
SP|Q99LF4|RTCB_MOUSE MSRNYNDELQFLDKINKNCWRIKKGFVPNMQVEGVFYVNDALEKLMFEELRNACRGGGVG 60
SP|P46850|RTCB_ECOLI ----MNYEL-----------------------------------LTTENAPVKMWTKGVP 21
SP|Q4R6X4|RTCB_MACFA MSRNYNDELQFLEKISKNCWRIKKGFVPNMQVEGVFYVNDALEKLMFEELRNACRGGGVG 60
SP|Q6NZS4|RTCB_DANRE MSRSYNDELQYLDKIHKNCWRIKKGFVPNMLVEGVFYVNDPLEKLMFEELRNACRGGGFG 60
SP|Q561P3|RTCB_XENTR MSRSYNDELQYLDKIHKNCWRIRKGFVPNMQVEGVFYVNDPLEKLMFEELRNASRGGAAG 60

SP|Q9Y3I0|RTCB_HUMAN GFLPAMKQIGNVAALPGIVHRSIGLPDVHSGYGFAIGNMAAFDMNDPEAVVSPGGVGFDI 120
SP|Q99LF4|RTCB_MOUSE GFLPAMKQIGNVAALPGIVHRSIGLPDVHSGYGFAIGNMAAFDMNDPEAVVSPGGVGFDI 120
SP|P46850|RTCB_ECOLI VEADARQQLINTAKMPFIFKHIAVMPDVHLGKGSTIGSVIP-----TKGAIIPAAVGVDI 76
SP|Q4R6X4|RTCB_MACFA GFLPAMKQIGNVAALPGIVHRSIGLPDVHSGYGFAIGNMAAFDMNDSEAVVSPGGVGFDI 120
SP|Q6NZS4|RTCB_DANRE GFLPAMKQIGNVAALPGIVHRSIGLPDVHSGYGFAIGNMAAFDMENPDAVVSPGGVGFDI 120
SP|Q561P3|RTCB_XENTR GFLPAMKQIGNVAALPGIIHRSIGLPDVHSGYGFAIGNMAAFDMDNPEAVVSPGGVGFDI 120

SP|Q9Y3I0|RTCB_HUMAN NCGVRLLRTNLDESDVQPVKEQLAQAMFDHIPVGVGSKGVIPMNAKDLEEALEMGVDWS- 179
SP|Q99LF4|RTCB_MOUSE NCGVRLLRTNLDESDVQPVKEQLAQAMFDHIPVGVGSKGVIPMNAKDLEEALEMGVDWS- 179
SP|P46850|RTCB_ECOLI GCGMNALRTALTAEDLPENLAELRQAIETAVPHGRTTGRCKRDKGAWENPPVNVDAKWAE 136
SP|Q4R6X4|RTCB_MACFA NCGVRLLRTNLDESDVQPVKEQLAQAMFDHIPVGVGSKGVIPMNAKDLEEALEMGVDWS- 179
SP|Q6NZS4|RTCB_DANRE NCGVRLLRTNLDEGDVQPVKEQLAQSLFDHIPVGVGSKGVIPMGAKDLEEALEMGVDWS- 179
SP|Q561P3|RTCB_XENTR NCGVRLLRTNLDESDVQPVKEQLAQAMFDHIPVGVGSKGVIPMGAKDLEEALEMGVDWS- 179

SP|Q9Y3I0|RTCB_HUMAN LREGYAWAEDKEHCEEYGRMLQADPNKVSARAKKRGLPQLGTLGAGNHYAEIQVVDEIFN 239
SP|Q99LF4|RTCB_MOUSE LREGYAWAEDKEHCEEYGRMLQADPNKVSPRAKKRGLPQLGTLGAGNHYAEIQVVDEIFN 239
SP|P46850|RTCB_ECOLI LEAGYQWLTQK-----YPRFL-----------NTNNYKHLGTLGTGNHFIEIC------- 173
SP|Q4R6X4|RTCB_MACFA LREGYAWAEDKEHCEEYGRMLQADPNKVSARAKKRGLPQLGTLGAGNHYAEIQVVDEIFN 239
SP|Q6NZS4|RTCB_DANRE LREGYAWAEDKEHCEEYGRMLQADPNKVSSKAKKRGLPQLGTLGAGNHYAEIQVVDEIYN 239
SP|Q561P3|RTCB_XENTR LREGYAWAEDKEHCEEYGRMLQADPSKVSSKAKKRGLPQLGTLGAGNHYAEVQVVDDIYD 239

SP|Q9Y3I0|RTCB_HUMAN EYAAKKMGIDHKGQVCVMIHSGSRGLGHQVATDALVAMEKAMKRDKIIVNDRQLACARIA 299
SP|Q99LF4|RTCB_MOUSE EYAAKKMGIDHKGQVCVMIHSGSRGLGHQVATDALVAMEKAMKRDKIIVNDRQLACARIA 299
SP|P46850|RTCB_ECOLI --------LDESDQVWIMLHSGSRGIGNAIGTYFIDLAQKEMQETLETLPSRDLAYFMEG 225
SP|Q4R6X4|RTCB_MACFA EYAAKKMGIDHKGQVCVMIHSGSRGLGHQVATDALVAMEKAMKRDKIIVNDRQLACARIA 299
SP|Q6NZS4|RTCB_DANRE DYAAKKMGIDHKGQVCVMIHSGSRGLGHQVATDALVAMEKAMKRDRITVNDRQLACARIT 299
SP|Q561P3|RTCB_XENTR EYAAKKMGIDHKGQVCVMIHSGSRGLGHQVATDALVAMEKAMKRDKITVNDRQLACARIS 299

SP|Q9Y3I0|RTCB_HUMAN SPEGQDYLKGMAAAGNYAWVNRSSMTFLTRQAFAKVF---NTTPDDLDLHVIYDVSHNIA 356
SP|Q99LF4|RTCB_MOUSE SPEGQDYLKGMAAAGNYAWVNRSSMTFLTRQAFAKVF---NTTPDDLDLHVIYDVSHNIA 356
SP|P46850|RTCB_ECOLI TEYFDDYLKAVAWAQLFASLNRDAMMENVVTALQSITQKTVRQPQTLAMEEI-NCHHNYV 284
SP|Q4R6X4|RTCB_MACFA SPEGQDYLKGMAAAGNYAWVNRSSMTFLTRQAFAKVF---NTTPDDLDLHVIYDVSHNIA 356
SP|Q6NZS4|RTCB_DANRE SEEGQDYLKGMAAAGNYAWVNRSSMTFLTRQAFSKVF---STTPDDLDMHVIYDVSHNIA 356
SP|Q561P3|RTCB_XENTR SAEGQDYLKGMAAAGNYAWVNRSSMTFLTRQAFSKVF---NTTPDDLDLHVIYDVSHNIA 356

SP|Q9Y3I0|RTCB_HUMAN KVEQHVVDGKERTLLVHRKGSTRAFPPHHPLIAVDYQLTGQPVLIGGTMGTCSYVLTGTE 416
SP|Q99LF4|RTCB_MOUSE KVEQHVVDGKERTLLVHRKGSTRAFPPHHPLIAVDYQLTGQPVLIGGTMGTCSYVLTGTE 416
SP|P46850|RTCB_ECOLI QKEQHFG----EEIYVTRKGAVS-------------ARAGQYGIIPGSMGAKSFIVRGL- 326
SP|Q4R6X4|RTCB_MACFA KVEQHVVDGKERTLLVHRKGSTRAFPPHHPLIAVDYQLTGQPVLIGGTMGTCSYVLTGTE 416
SP|Q6NZS4|RTCB_DANRE KVEEHMVDGRQKTLLVHRKGSTRAFPPHHPLIPVDYQLTGQPVLIGGTMGTCSYVLTGTE 416
SP|Q561P3|RTCB_XENTR KVEQHVVDGKEKTLLVHRKGSTRAFPPHHPLIPVDYQLTGQPVLIGGTMGTCSYVLTGTD 416

SP|Q9Y3I0|RTCB_HUMAN QGMTETFGTTCHGAGRALSRAKSRRNLDFQDVLDKLADMGIAIRVASPKLVMEEAPESYK 476
SP|Q99LF4|RTCB_MOUSE QGMTETFGTTCHGAGRALSRAKSRRNLDFQDVLDKLADMGIAIRVASPKLVMEEAPESYK 476
SP|P46850|RTCB_ECOLI -GNEESFCSCSHGAGRVMSRTKAKKLFSVEDQIRATAHV----ECRKDAEVIDEIPMAYK 381
SP|Q4R6X4|RTCB_MACFA QGMTETFGTTCHGAGRALSRAKSRRNLDFQDVLDKLADMGIAIRVASPKLVMEEAPESYK 476
SP|Q6NZS4|RTCB_DANRE QGMTETFGTTCHGAGRALSRAKSRRNLDFQDVLDKLADMGIAIRVASPKLVMEEAPESYK 476
SP|Q561P3|RTCB_XENTR QGMTETFGTTCHGAGRALSRAKSRRNLDFQDVLDKLADLGIAIRVASPKLVMEEAPESYK 476

SP|Q9Y3I0|RTCB_HUMAN NVTDVVNTCHDAGISKKAIKLRPIAVIKG 505
SP|Q99LF4|RTCB_MOUSE NVTDVVNTCHDAGISKKAIKLRPIAVIKG 505
SP|P46850|RTCB_ECOLI DIDAVMAAQSD--LVEVIYTLRQVVCVKG 408
SP|Q4R6X4|RTCB_MACFA NVTDVVNTCHDAGISKKAIKLRPIAVIKG 505
SP|Q6NZS4|RTCB_DANRE NVTDVVNTCHDAGISKKAIKLRPIAVIKG 505
SP|Q561P3|RTCB_XENTR NVTDVVNTCHDAGISKKAIKLRPIAVIKG 505
```

Figure S4

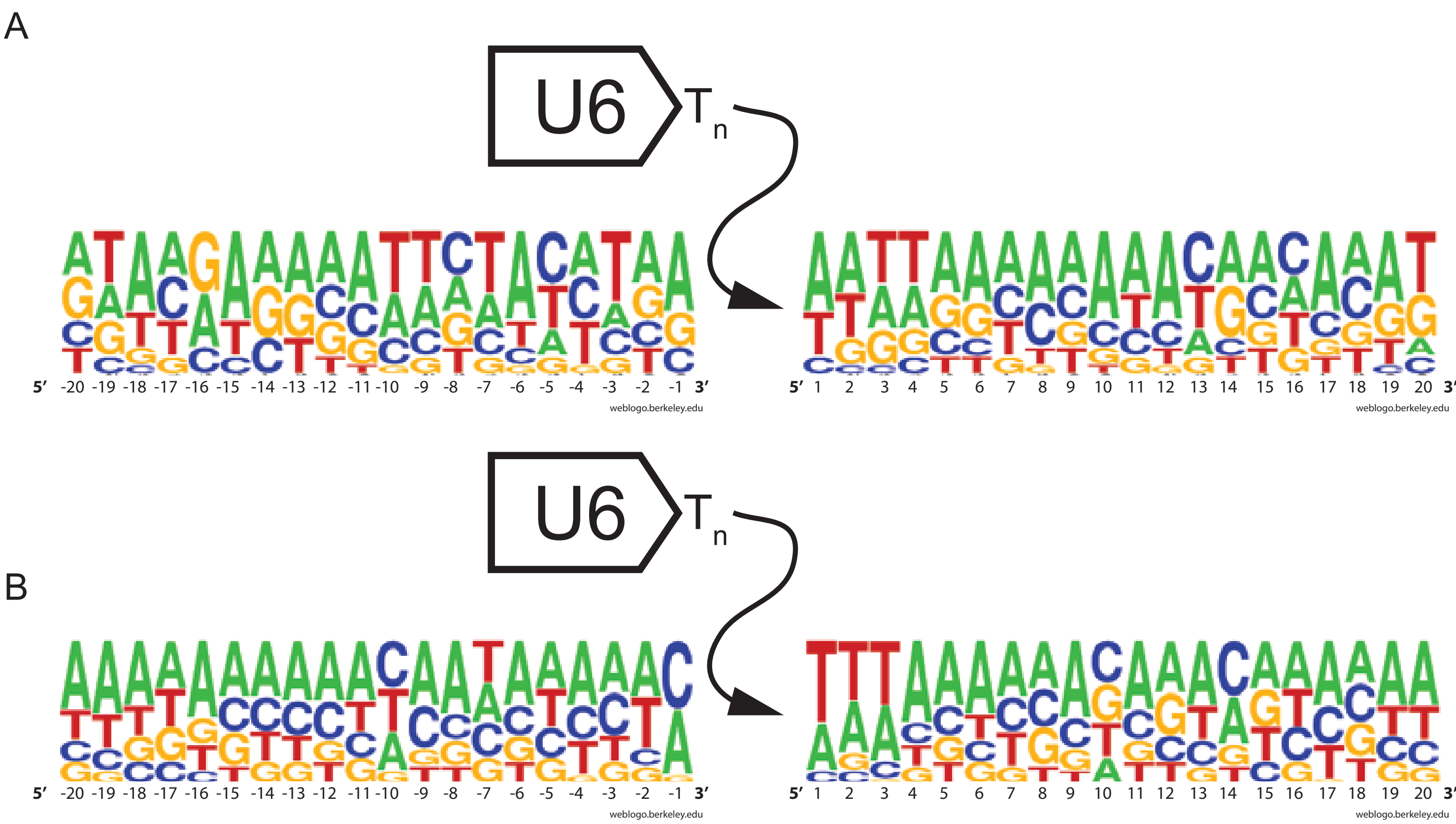

A
U6
Tn
5′ -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 3′
weblogo.berkeley.edu
5′ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 3′
weblogo.berkeley.edu
B
U6
Tn
5′ -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 3′
weblogo.berkeley.edu
5′ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 3′
weblogo.berkeley.edu

Table S1. Analysis of U6/L1 chimeric RNA junctions from engineered human L1s

L1 plasmid	L1.3 junction	U6/L1 junction -20	Junction Ts	U6/L1 junction + 20
pJM101/L1.3Δneo	4387	GACCTCTTCAAGGAGAACTA	5	AAGAACATTCCATGCTCATG
pJM101/L1.3Δneo	4819	GGGAAAACTGGCTAGCCATA	6(1)	TGTAGAAAGCTGAAACTGGA
pJM101/L1.3Δneo	4869	ACACCTTATACAAAAATCAA	7(2)	TTCAAGATGGATTAAAGATT
pJM101/L1.3Δneo	5260	ACATTTATGCAGCCAAAAAA	5	CACATGAAGAAATGCTCATC
pJM101/L1.3Δneo	5269	CAGCCAAAAAACACATGAAG	5	AAATGCTCATCATCACTGGC
pJM101/L1.3Δneo	5316	GAAATGCAAATCAAAACCAC	5(1)	TATGAGATATCATCTCACAC
pJM101/L1.3Δneo	5519	TGACCCAGCCATCCCATTAC	5(1)	TGGGTATATACCCAAATGAG
pJM101/L1.3Δneo	5539	TGGGTATATACCCAAATGAG	5(1)	TATAAATCATGCTGCTATAA
pJM101/L1.3Δneo	5625	AAGACTTGGAACCAACCCAA	5	ATGTCCAACAATGATAGACT
pJM101/L1.3Δneo	5732	ATCCTTTGTAGGGACATGGA	6(1)	TGAAATTGGAAACCATCATT
pJM101/L1.3Δneo	5734	CCTTTGTAGGGACATGGATG	5	AAATTGGAAACCATCATTCT
pJM101/L1.3Δneo	5923	GGGAGATATACCTAATGCTA	5	GATGACACATTAGTGGGTGC
pJM101/L1.3Δneo	5924	GGAGATATACCTAATGCTAG	5	ATGACACATTAGTGGGTGCA
pJM101/L1.3Δneo	5946	GACACATTAGTGGGTGCAGC	5	GCACCAGCATGGCACATGTA
pJM108/L1.3Δneo	5180	TAAACAAATTTACAAGAAAA	5	AAACAAACAACCCCATCAAA
pJM108/L1.3Δneo	5316	GAAATGCAAATCAAAACCAC	5(1)	TATGAGATATCATCTCACAC
pJM108/L1.3Δneo	5696	GGAATACTATGCAGCCATAA	5	AAAATGATGAGTTCATATCC
pJM108/L1.3Δneo	5750	GATGAAATTGGAAACCATCA	8(2)	TTCTCAGTAAACTATCGCAA
pJM108/L1.3Δneo	5843	GAGATCACATGGACACAGGA	5	AGGGGAATATCACACTCTGG
pJM108/L1.3Δneo	5906	GGGGGAGGGATAGCATTGGG	5	AGATATACCTAATGCTAGAT
pJM105/L1.3Δneo	5154	CTAATATCCAGAATCTACAA	6(1)	TGAACTTAAACAAATTTACA
pJM105/L1.3Δneo	5335	CTATGAGATATCATCTCACA	6	CCAGTTAGAATGGCAATCAT
pJM105/L1.3Δneo	5643	AAATGTCCAACAATGATAGA	5	CTGGATTAAGAAAATGTGGC
pJM105/L1.3Δneo	5674	AAATGTGGCACATATACACC	5	ATGGAATACTATGCAGCCAT
pJM105/L1.3Δneo	5685	ATATACACCATGGAATACTA	6(1)	TGCAGCCATAAAAAATGATG
pJM105/L1.3Δneo	5799	ACCAAACACCGCATATTCTC	5	ACTCATAGGTGGGAATTGAA
pJM105/L1.3Δneo	5860	GGAAGGGGAATATCACACTC	5(1)	TGGGGACTGTGGTGGGGTCG
pJM105/L1.3Δneo	5888	GTGGTGGGGTCGGGGGAGGG	5	GGGAGGGATAGCATTGGGAG
pJM105/L1.3Δneo	5889	TGGTGGGGTCGGGGGAGGGG	5	GGAGGGATAGCATTGGGAGA
pJM105/L1.3Δneo	5892	TGGGGTCGGGGGAGGGGGGA	5	GGGATAGCATTGGGAGATAT
pJM119/L1.3Δneo	4758	TTTGACAAACCTGAGAAAAA	5	CAAGCAATGGGGAAAGGATT
pJM119/L1.3Δneo	4977	GTGGGCAAGGACTTCATGTC	5	CAAAACACCAAAAGCAATGG
pJM119/L1.3Δneo	5497	ATCTAGAACTAGAAATACCA	8(3)	TTTGACCCAGCCATCCCATT
pJM119/L1.3Δneo	5582	CACATGCACACGTATGTTTA	6(2)	TTGCGGCACTATTCACAATA
pJM119/L1.3Δneo	5750	GATGAAATTGGAAACCATCA	7(2)	TTCTCAGTAAACTATCGCAA
pJM119/L1.3Δneo	5768	CATTCTCAGTAAACTATCGC	5	AAGAACAAAAAACCAAACAC
pJM119/L1.3Δneo	5808	CGCATATTCTCACTCATAGG	5(1)	TGGGAATTGAACAATGAGAT
pJM119/L1.3Δneo	5901	GGGAGGGGGGAGGGATAGCA	5(2)	TTGGGAGATATACCTAATGC

Table S2. Analysis of U6/L1 chimeras containing 5'-truncated L1s

L1.3 junction	U6/L1 junction -20	#Ts at end of U6	U6/L1 junction + 20
5755	CTC	4	AGTAAACTATCGCAAGAACA
5755	CTC	4	AGTAAACTATCGCAAGAACA
5759	CTCAGTA	4	AACTATCGCAAGAACAAAAA
5759	CTCAGTA	4	AACTATCGCAAGAACAAAAA
5817	TCACTCATAGGTGGGAATTG	4	AACAATGAGATCACATGGAC
5924	GGAGATATACCTAATGCTAG	4	ATGACACATTAGTGGGTGCA
5928	ATATACCTAATGCTAGATGA	4	CACATTAGTGGGTGCAGCGC
5942	AGATGACACATTAGTGGGTG	4	CAGCGCACCAGCATGGCACA

Table S3. Analysis of U6/L1 chimeras containing 5'-truncated L1s

L1.3 junction	U6/L1 junction -20	#Ts at end of U6	U6/L1 junction + 20
5759	CTCAGTA	4	AACTATCGCAAGAACAAAAA
5764	CTCAGTAAACTA	4	TCGCAAGAACAAAAAACCAA
5775	AGTAAACTATCGCAAGAACA	4	AAAAACCAAACACCGCATAT
5886	CTGTGGTGGGGTCGGGGGAG	4	GGGGGAGGGATAGCATTGGG
5888	GTGGTGGGGTCGGGGGAGGG	4	GGGAGGGATAGCATTGGGAG
5891	GTGGGGTCGGGGGAGGGGGG	4	AGGGATAGCATTGGGAGATA
5892	TGGGGTCGGGGGAGGGGGGA	4	GGGATAGCATTGGGAGATAT
5907	GGGGAGGGATAGCATTGGGA	4	GATATACCTAATGCTAGATG
5908	GGGAGGGATAGCATTGGGAG	4	ATATACCTAATGCTAGATGA
5908	GGGAGGGATAGCATTGGGAG	4	ATATACCTAATGCTAGATGA
5908	GGGAGGGATAGCATTGGGAG	4	ATATACCTAATGCTAGATGA
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5909	GGAGGGATAGCATTGGGAGA	4	TATACCTAATGCTAGATGAC
5924	GGAGATATACCTAATGCTAG	4	ATGACACATTAGTGGGTGCA
5924	GGAGATATACCTAATGCTAG	4	ATGACACATTAGTGGGTGCA
5924	GGAGATATACCTAATGCTAG	4	ATGACACATTAGTGGGTGCA
5924	GGAGATATACCTAATGCTAG	4	ATGACACATTAGTGGGTGCA
5925	GAGATATACCTAATGCTAGA	4	TGACACATTAGTGGGTGCAG
5925	GAGATATACCTAATGCTAGA	4	TGACACATTAGTGGGTGCAG
5928	ATATACCTAATGCTAGATGA	4	CACATTAGTGGGTGCAGCGC
5928	ATATACCTAATGCTAGATGA	4	CACATTAGTGGGTGCAGCGC
5930	ATACCTAATGCTAGATGACA	4	CATTAGTGGGTGCAGCGCAC
5930	ATACCTAATGCTAGATGACA	4	CATTAGTGGGTGCAGCGCAC
5930	ATACCTAATGCTAGATGACA	4	CATTAGTGGGTGCAGCGCAC
5931	TACCTAATGCTAGATGACAC	4	ATTAGTGGGTGCAGCGCACC
5933	CCTAATGCTAGATGACACAT	4	TAGTGGGTGCAGCGCACCAG
5944	ATGACACATTAGTGGGTGCA	4	GCGCACCAGCATGGCACATG
5946	GACACATTAGTGGGTGCAGC	4	GCACCAGCATGGCACATGTA
5946	GACACATTAGTGGGTGCAGC	4	GCACCAGCATGGCACATGTA
5946	GACACATTAGTGGGTGCAGC	4	GCACCAGCATGGCACATGTA
5946	GACACATTAGTGGGTGCAGC	4	GCACCAGCATGGCACATGTA
5946	GACACATTAGTGGGTGCAGC	4	GCACCAGCATGGCACATGTA
5948	CACATTAGTGGGTGCAGCGC	4	ACCAGCATGGCACATGTATA
5952	TTAGTGGGTGCAGCGCACCA	4	GCATGGCACATGTATACATA
5952	TTAGTGGGTGCAGCGCACCA	4	GCATGGCACATGTATACATA
5952	TTAGTGGGTGCAGCGCACCA	4	GCATGGCACATGTATACATA
5954	AGTGGGTGCAGCGCACCAGC	4	ATGGCACATGTATACATATG
5955	GTGGGTGCAGCGCACCAGCA	4	TGGCACATGTATACATATGT
5955	GTGGGTGCAGCGCACCAGCA	4	TGGCACATGTATACATATGT
5955	GTGGGTGCAGCGCACCAGCA	4	TGGCACATGTATACATATGT
5966	GCACCAGCATGGCACATGTA	4	TACATATGTAACTAACCTGC
5966	GCACCAGCATGGCACATGTA	4	TACATATGTAACTAACCTGC
5966	GCACCAGCATGGCACATGTA	4	TACATATGTAACTAACCTGC
5970	CAGCATGGCACATGTATACA	4	TATGTAACTAACCTGCACAA
5970	CAGCATGGCACATGTATACA	4	TATGTAACTAACCTGCACAA
5972	GCATGGCACATGTATACATA	4	TGTAACTAACCTGCACAATG
5974	ATGGCACATGTATACATATG	4	TAACTAACCTGCACAATGTG
5981	ATGTATACATATGTAACTAA	4	CCTGCACAATGTGCACATGT
5982	TGTATACATATGTAACTAAC	4	CTGCACAATGTGCACATGTA

Table S4. Analysis of “aligned” U6/L1 sequences from RNA-seq experiments

L1.3 Junction	L1 subfamily	U6/L1 junction -20	Juntion Ts	U6/L1 junction + 20
2052	L1PA10	GTAAATGGGCTAAATGCCCC	5	AATTAAAAGACACAGAATGG
2234	L1PA3	AATCCTAGTCTCTGATAAAA	5	CAGACTTTAAACCAACAAAG
2568	L1PA4	AATCAACAGAATATACATTC	8(4)	TTTTCAGCACCACACCACAC
2598	L1PA5	CCACATCACACTTATTCCAA	5	AATTGACCACATAGTTGGAA
3450	L1PA7	CTACCAGGAGTACAAAGAGG	5	AGCTGGTACCAATCCTTCTG
4268	L1PA5	ATGAGTGAACTCCCATTCAC	5	AATTGCTTCAAAGAGAATAA
4611	L1PA2	GGAGGCATCACACTACCTGA	5	CTTCAAACTATACTACAAGG
4683	L1PA2	CAAAACAGAGATATAGATCA	5	ATGGAACAGAACAGAGCCCT
5030	L1PA7	AATTGACAAATGGGATCTAA	6(2)	TTAAAATAAAGAGCTTCTGC
5095	L1PA7	TGAACAGACAACCTACAGAA	5(1)	TGGAAGAAAATTTTTGCAAT
5281	L1PA2	ACATGAAAAAATGCTCATCA	5(1)	TCACTGGCCATCAGAGAAAT
5358	L1PA5	GTTAGAATGGCGATCATTAA	5	AAAGTCAGGAAACAACAGGT
5558	L1PA7	TTATAAATCATTCTACTGTA	5	AAAACACATGCACACATGTT
5647	L1PA2	GTCCAACAATGATAGACTGG	5	ATTAAGAAAATGTGGCACAT
5720	L1PA5	TGATGAGTTCATGTCCTTTG	5(1)	TAGGGACATGGATGAAGCTG
5906	L1PA3	GGGGGAGGGATAGCATTAGG	5	AGATATACCTAATGCTAAAT

Table S5. Analysis of “non-aligned” U6/L1 sequences from RNA-seq experiments

L1.3 Junction	U6/L1 junction -20	Juntion Ts	U6/L1 junction + 20
474	TGAGGCTTGAGTAGGTAAAC	5	AAAGTAGCCGGGAAGCTCGA *
832	TTAGAAGGAAAACTAACAAC	5	CAGAAAGGACATCTACACCG
864	ATCTACACCGAAAACCCATC	6(1)	TGTACATCACCATCATCAAA **
1125	GTTAAAAACTTTGAAAAAAA	5	ATTAGACGAATGGCTAACTA *
1231	GTGACGAATGCACAAGCTTC	5	AGTAGCCGATTCGATCAACT
1454	AAACACTCTGCAGGATATTA	6(1)	TCCAGGAGAACTTCCCCAAT
1559	AAGAGCAACTCCAAGACACA	6(1)	TAATTGTCAGATTCACCAAA
1748	AAAGAATTTTCAACCCAGAA	8(3)	TTTCATATCCAGCCAAACTA
1752	AATTTTCAACCCAGAATTTC	5	ATATCCAGCCAAACTAAGCT
1824	GACAAGCAAATGTTGAGAGA	6(3)	TTTTGTCACCACCAGGCCTG
2025	TCACACATAACAATATTAAC	6(3)	TTTAAATATAAATGGACTAA
2395	TCAGTGACCTACAAAGAGAC	7(2)	TTAGACTCCCACACATTAAT
2657	ATGTAAAAGAACAGAAATTA	6(1)	TAACAAACTATCTCTCAGAC
2728	AGAATCTCACTCAAAGCCGC	6(1)	TCAACTACATGGAAACTGAA
2849	AGACACCACATACCAGAATC	5(1)	TCTGGGACGCATTCAAAGCA
2884	AAGCAGTGTGTAGAGGGAAA	5(3)	TTTATAGCACTAAATGCCTA
3056	AATAGAGACACAAAAAACCC	6(2)	TTCAAAAAATCAATGAATCC
3262	TCTACGCAAATAAACTAGAA	5	AATCTAGAAGAAATGGATAC
3307	ACACATACACTCTCCCAAGA	5	CTAAACCAGGAAGAAGTTGA
3311	CACATACACTCTCCCAAGAC	6(1)	TAAACCAGGAAGAAGTTGAA
3872	TATTGATGGGACGTATTTCA	5	AAATAATAAGAGCTATCTAT
3945	CAAAAACTGGAAGCATTCCC	8(3)	TTTGAAAACCGGCACAAGAC
4131	CTAGAAAACCCCATCGTCTC	6	AGCCCAAAATCTCCTTAAGC
4190	CAGGATACAAAATCAATGTA	5	CAAAAATCACAAGCATTCTT
4783	AATGGGGAAAGGATTCCCTA	8(3)	TTTAATAAATGGTGCTGGGA
4854	CCCTTCCTTACACCTTATAC	6	AAAAATCAATTCAAGATGGA
4887	AATTCAAGATGGATTAAAGA	6(3)	TTTAAACGTTAAACCTAAAA
5029	AAATTGACAAATGGGATCTA	5	ATTAAACTAAAGAGCTTCTG
5145	GACAAAGGGCTAATATCCAG	5	AATCTACAATGAACTCAAAC
5197	AAAAAACAAACAACCCCATC	7	AAAAAGTGGGCGAAGGACAT
5416	AAATAGGAACACTTTTACAC	5(1)	TGTTGGTGGGACTGTAAACT
5593	GTATGTTTATTGCGGCACTA	6(2)	TTCACAATAGCAAAGACTTG
5757	TTGGAAACCATCATTCTCAG	6(1)	TAAACTATCGCAAGAACAAA

Table S6. Sequence features of the 25bp U6/L1 junction sequences motifs of the “aligned”, “non-aligned”, and putative “artifact” RNA-seq chimeras

Number	Category	L1.3 Junction	25bp Junction Motif	# supporting reads	Cell Line
1	aligned	2052*	**5'–CATTCTGTA**TTTTTAATTAAAAGAC	1	NPC
2	aligned	2234	**5'–CGTTCTGTA**TTTTTCAGACTCTAAA	2	NPC
3	aligned	2568	**5'–CGTTCCATT**TCTTTTTCAGCACCA	4	NPC, JVM
4	aligned	2598	**5'–CGTTCCATA**TTTTTAATTGACCACA	3	H9, PA-1
5	aligned	3450	**5'–CGTTCCATA**TTTTTACTGGTACCAT	1	HA
6	aligned	4268	**5'–CGTTCCATA**TTTTTAATTGCTTCAA	6	PA-1, NPC, JVM
7	aligned	4611	**5'–CGTTCCATA**TTTTTCTTCAAACTAT	2	NPC, JVM
8	aligned	4683	**5'–CGTTCCATA**TTTTTATGGAACAGAA	5	H9, NPC, PA-1, HA
9	aligned	5030	**5'–CGTTCCGTA**TTTTTTAAACTAAAGA	1	NPC
10	aligned	5095	**5'–CATTCCATA**TTTTTGGGAGAAAATT	3	PA-1, H9
11	aligned	5281	**5'–CGTTCCATA**TTTTTCACTGGCCATC	4	JVM, NPC
12	aligned	5358	**5'–CGTTCCATA**TTTTTAAAGTCAGGAA	1	NPC
13	aligned	5558	**5'–AGTTCCGTA**TTTTTAAAACACATGC	1	NPC
14	aligned	5647	**5'–CGTTCCATA**TTTTTATTAAGAAAAT	3	NPC
15	aligned	5720	**5'–CGTTCCATA**TTTTTAGGGACATGGA	1	NPC
16	aligned	5906	**5'–CGTTCCATA**TTTTTAGATATACCTA	1	HA
17	non-aligned	474	**5'–CGTTCCATA**TTTTTAAAGCGTCCTG	1	JVM
18	non-aligned	832	**5'–CGTTCCATA**TTTTTAACAGAAAGGA	1	HA
19	non-aligned	864	**5'–CGTTCCATA**TTTTTTGTACATCACC	1	NPC
20	non-aligned	1125	**5'–CGTTCCATA**TTTTTATTGACGAATG	1	NPC
21	non-aligned	1231	**5'–CGTTCCATA**TTTTTAGTAGCTGATT	1	NPC
22	non-aligned	1454	**5'–CGTTCCATA**TTTTTTCCAGGAGAAC	1	H9
23	non-aligned	1559	**5'–CGTTCCATA**TTTTTTAATTGTCAGA	1	JVM
24	non-aligned	1748	**5'–CGTTCCATA**TTTTTTTTCATATCCA	2	HA
25	non-aligned	1752	**5'–CGTTCCATA**TTTTTATATCCAGCCA	1	NPC
26	non-aligned	1824	**5'–CGTTCCATA**TTTTTTTGTCACCACC	2	NPC
27	non-aligned	2025	**5'–CGTTCCATA**TTTTTTAAATGTAAAT	1	H9
28	non-aligned	2395	**5'–CGTTCCATA**TTTTTTTAGACTCCCA	2	NPC
29	non-aligned	2657	**5'–CGTTCCATA**TTTTTTAACAAACTGT	1	NPC
30	non-aligned	2728	**5'–CGTTCCATA**TTTTTTCAACTACATA	1	H9
31	non-aligned	2849	**5'–CGTTCCATA**TTTTTCTGGGACACAT	1	NPC
32	non-aligned	2884	**5'–CGTTCCATA**TTTTTATAGCACTAAA	2	NPC
33	non-aligned	3056	**5'–CGTTCCATA**TTTTTTCAAAAAATCA	1	NPC
34	non-aligned	3262	**5'–CGTTCCATA**TTTTTAATCTAGAAGA	1	NPC
35	non-aligned	3307	**5'–CGTTCCATA**TTTTTAGGCTAAACCA	1	NPC
36	non-aligned	3311	**5'–CGTTCCATA**TTTTTTAAACCAGGCA	1	JVM
37	non-aligned	3872	**5'–CGTTCCATA**TTTTTAAATAATAAGA	1	NPC
38	non-aligned	3945	**5'–CGTTCCATA**TTTTTTTTGAAAACTG	1	H9
39	non-aligned	4131	**5'–CGTTCCATA**TTTTTTAGCCCAAAAT	1	JVM
40	non-aligned	4190	**5'–CGTTCCATA**TTTTTATGTTCAAAAA	1	NPC
41	non-aligned	4783	**5'–CGTTCCATA**TTTTTTTTAATAAATG	3	NPC
42	non-aligned	4854	**5'–CGTTCCATA**TTTTTTATACAAAAAA	1	NPC
43	non-aligned	4887	**5'–CGTTCCATA**TTTTTTAAACGTTAGA	1	NPC
44	non-aligned	5029	**5'–CGTTCCATA**TTTTTATTAAACTAAA	1	NPC
45	non-aligned	5145	**5'–CGTTCCATA**TTTTTAATCTACAATG	1	NPC
46	non-aligned	5197	**5'–CGTTCCATA**TTTTTTTAAAAAGTGG	2	NPC
47	non-aligned	5416	**5'–CGTTCCATA**TTTTTGTTGGTGGGAC	1	H9
48	non-aligned	5593	**5'–CGTTCCATA**TTTTTTCACAATAGCA	1	HA
49	non-aligned	5757	**5'–CGTTCCATA**TTTTTTAAACTATCGC	1	HA
50	non-aligned	934**	**5'–CGTTCCATA**TTTTTTTCTGCTCTGT	1	NPC
51	non-aligned	4322**	**5'–CGTTCCATA**TTTTTTTCACATCCCT	1	JVM
52	non-aligned	5259**	**5'–CGTTCCATA**TTTTTTTGTTGGCCAC	1	NPC
53	non-aligned	5343**	**5'–CGTTCCATA**TTTTTAATGATGACGT	1	NPC
54	artifact	-	**5'–TTCGTGAAGCGT**ATACACCAATAAC	1	NPC
55	artifact	-	**5'–GACACGCAAATTC**TATTGAGGGTTT	2	H9
56	artifact	-	**5'–ACACGCAAATTC**ATCAGTGAATCCA	1	NPC
57	artifact	-	**5'–ACGCAAATTCG**ATAAAAATCCTAGA	2	H9
58	artifact	-	**5'–GACACGCAAATTC**TTTTTATGGCTG	1	NPC
59	artifact	-	**5'–CACGCAAATTC**AAAATACTGGCAAA	1	HA
60	artifact	-	**5'–ACGCAAATTCGATGAA**ATAAAGCAT	1	NPC
61	artifact	-	**5'–GACACGCAAATTCTTG**GGTTGGTTC	2	H9
62	artifact	-	5'–**CACGCAAATTCTTGAAG**ATGACATG	1	H9
63	artifact	-	**5'–CACGCAAATTCG**GTACCTGAAAGGA	1	NPC
64	artifact	-	**5'–ATGACACGCAAATTCG**ACAAAGGGC	1	NPC

Table S7. Characterization of 16 genomic U6/L1 chimeric pseudogenes that served as putative source elements for the RNA-seq reads detected in Supplemental Table 4

L1.3 Junction	Genome position (hg38)	L1 subfamily	TSD	Cleavage	Remarks
2052	chrX:102678813-102674130	L1PA10	-	-	ARMCX5-GPRASP2 Intron
2234	chr13:48987911-48988334	L1PA3	7	TTTA/T	FNDC3A intron
2568	chr1:180758722-180762284	L1PA4	7	CTTT/T	XPR1 intron
2598	chr3:98805084-98801701	L1PA5	-	-	DCBLD2 intron
3450	chr8:103384961-103387948	L1PA7	12	TGTC/T	intergenic
4268	chr13:72706123-72704270	L1PA5	19	TTTT/A	intergenic
4611	chr18:68858934-68860488	L1PA2	-	-	CCDC102B intron
4683	chr4:39296252-39297711	L1PA2	11	TCTT/A	RFC1 intron
5030	chr1:42569034-42570125	L1PA7	-	-	CCDC30 intron
5095	chr14:37434573-37433236	L1PA7	6	ATTT/A	MIPOL1 intron
5281	chr3:196784226-196785086	L1PA2	15	TTTT/A	PAK2 intron
5358	chr4:109992325-109993102	L1PA5	16	TTTT/A	EGF intron
5558	chr14:102865856-102866427	L1PA7	14	CTTT/A	TRAF3 intron
5647	chr15:65553187-65552698	L1PA2	14	TTTT/A	HACD3 intron
5720	chr4:76532327-76531908	L1PA5	10	GCTC/T	SHROOM3 intron
5906	chr2:174558072-174557836	L1PA3	16	CTTT/G	intergenic

Table S8. 1000 Genomes Project sample numbers with population codes.

Sample	Population Code
HG00096	GBR
HG00268	FIN
HG00419	CHS
HG00759	CDX
HG01051	PUR
HG01112	CLM
HG01500	IBS
HG01565	PEL
HG01583	PJL
HG01595	KHV
HG01879	ACB
HG02568	GWD
HG02922	YRI
HG03052	MSL
HG03642	STU
HG03742	ITU
NA18525	CHB
NA18939	JPT
NA19017	LWK
NA19625	ASW
NA19648	MXL
NA20502	TSI
NA20845	GIH

Table S9. Oligos used in this study.

Oligo Name	Sequence
U6s1	5'ACAGAGAAGATTAGCATGGC
SV40as	5'AAACTCATCAATGTATCTTATCATGTCTGG
U6s2	5'CCCTGCGCAAGGATGAC
3UTRas	5'GTTTTAGGGTACATGTGCACATTGC
hrGFPas1	5'TTACACCCACTCGTGCAGG
hrGFPas2	5'TCGTGCTGCTCCACGAAGC
U6as1	5'AAAATATGGAACGCTTCACG
HDVas1	5'CTTCTCCCTTAGCCTACCG
T7	5'TAATACGACTCACTATAGGG
T7_U6_HDVr	5'TAATACGACTCACTATAGGGTGCTCGCTTCGGCAGCACATATACTAAAATTGGAACGATAC AGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAATTCGTGAAGCGTTCCATATTT Tgggtcggcatggcatctccacctcctcgcggtccgacctgggctacttcggtaggctaaggg agaag
T7_U6	5'TAATACGACTCACTATAGGGTGCTCGCTTCGGCAGCACATATACTAAAATTGGAACGATAC AGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAATTCGTGAAGCGTTCCATATTT T
T7_JM101noNeo_5752-6087	5'TAATACGACTCACTATAGGGCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCAT ATTCTCACTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAATATCAC ACTCTGGGGACTGTGGTGGGGTCGGGGGAGGGGGGAGGGATAGCATTGGGAGATATACCTAAT GCTAGATGACACATTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTAAC CCGGGCAATGTGCACATGTACCCTAAAACTTAAAGTATAATAAAGACGTCAGGGTTCGAAATC GATAAGCTTGGATCCAGACATGATAAGATACATTGATGAGTTT
T7_HHr_JM101noNeo_5752-6087	5'TAATACGACTCACTATAGGgtttactgagctgatgagtccgtgaggacgaaacgtggagac acgtcCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCACTCATAGGTG GGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAATATCACACTCTGGGGACTGTGGT GGGGTCGGGGGAGGGGGGAGGGATAGCATTGGGAGATATACCTAATGCTAGATGACACATTAG TGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTAACCCGGGCAATGTGCACAT GTACCCTAAAACTTAAAGTATAATAAAGACGTCAGGGTTCGAAATCGATAAGCTTGGATCCAG ACATGATAAGATACATTGATGAGTTT
T7_HHr_GFP	5'TAATACGACTCACTATAGGgcacgccgtcgctgatgagtccgtgaggacgaaacgtggaga cacgtcCGACGGCGTGCTGGTGGCCAGGTGATCCTGGTGTACCGCCTGAACAGCGGCAAGTT CTACAGCTGCCACATGCGCACCCTGATGAAGAGCAAGGGCGTGGTGAAGGACTTCCCCGAGTA CCACTTCATCCAGCACCGCCTGGAGAAGACCTACGTGGAGGACGGCGGCTTCGTGGAGCAGCA CGAGACCGCCATCGCCCAGCTGACCAGCCTGGGCAAGCCCCTGGGCAGCCTGCACGAGTGGGT GTAATAGGTCCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAG TG
U6L1_qPCR_standard	5'ACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAATTCGTGAAGCGTTCCAT ATTTTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCACTCATAGGTG GGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAATATCACACTCTGGGGACTGTGGT GGGGTCGGGGGAGGGGGGAGGGATAGCATTGGGAGATATACCTAATGCTAGATGACACATTAG TGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTAACCCGGGCAATGTGCACAT GTACCCTAAAACTTAAAGTATAATAAAGACGTCAGGGTTCGAAATCGATAAGCTTGGATCCAG ACATGATAAGATACATTGATGAGTTT
U6L1_qPCR_1F	5 CAAATTCGTGAAGCGTTCCATA
U6L1_qPCR_1R	5'CTTCCTGTGTCCATGTGATCT
U6L1_qPCRcon_4F	5'TGACACATTAGTGGGTGCAG
U6L1_qPCRcon_4R	5'ATCGATTTCGAACCCTGACG
GFP sgRNA target	5'CACCATGGAGGGCTGCGGCA
RtcB sgRNA target	5'GAATTAAGGAATGCCTGTCG
Splint	5'GTTCTTGCGATAGTTTACATATGGAACGCTTCACGA