Revisiting the impact of synthetic ORF sequences on engineered LINE-1 retrotransposition

Sandra R Richardson[1,5,*], Dorothy Chan[1,5], Patricia Gerdes[1], Jeffrey S Han[2], Jef D Boeke[3], Geoffrey J Faulkner[1,4,*]

1. Mater Research Institute - University of Queensland, TRI Building, Woolloongabba QLD 4102, Australia.

2. Department of Biochemistry and Molecular Biology, Tulane University School of Medicine, New Orleans, LA, 70112, USA.

3. Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY, 10016, USA; Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn NY 11201

4. Queensland Brain Institute, University of Queensland, Brisbane QLD 4072, Australia.

5. These authors contributed equally.

*Correspondence: S.R.R. (sandra.richardson@mater.uq.edu.au), G.J.F. (faulknergj@gmail.com)

Abstract: The retrotransposon Long Interspersed Element 1 (L1) contains adenosine rich ORFs, a characteristic that limits its expression in mammalian cells. A previously developed synthetic mouse L1 (smL1) with ORF adenosine content decreased from 40% to 26% showed higher mRNA expression levels and retrotransposed far more efficiently than the native parental element, L1spa. Here, we observe two nonsynonymous substitutions between the L1spa and smL1 ORF1 sequences, and note that the smL1 3'UTR lacks a conserved guanosine-rich region (GRR). We find that the combined effect of these amino acid changes and the 3'UTR deletion, rather than synthetic ORF sequences, accounts for the dramatic and reproducible increase in smL1 retrotransposition efficiency over L1spa. Furthermore, we demonstrate that the position of the GRR within the L1 reporter construct impacts retrotransposition efficiency. Our results prompt a reevaluation of synthetic L1 activity and suggest that native mouse L1 mobility is frequently underestimated in engineered retrotransposition assays.

Introduction: L1 retrotransposons are an ongoing source of mutagenesis in mammalian genomes[1,2]. Mice contain ~3000 retrotransposition-competent L1s (RC-L1s) representing three subfamilies (L1_T_F, L1_G_F, L1_A), distinguished by their promoter-harboring 5'UTRs[3-6]. RC-L1s encode two proteins required for their mobility in cis: ORF1p, a nucleic acid binding and chaperone protein, and ORF2p, which has endonuclease and reverse transcriptase activities critical for the generation of new L1 insertions by target-primed reverse transcription (TPRT)[7-13]. The L1 3'UTR contains a weak polyadenylation (poly-A) signal and a poly-purine tract which is found in L1s across species but which varies in length and nucleotide sequence[14-18]. Hallmarks of TPRT-mediated L1 integrants include insertion at the degenerate consensus 5'-TTTT/AA-3', terminal poly-A tracts, and flanking target site duplications (TSDs)[19-23]. Approximately 1 in 8 mice and 1 in 62.5 humans harbor a de novo L1 insertion[24,25]. Moreover, >100 cases of human genetic disease and numerous spontaneous mouse mutants have been linked to L1-mediated retrotransposition events, highlighting the impact of L1 mutagenesis on mammalian genomes[26,27].

The mobilization efficiency of individual L1 copies can be evaluated using a cultured cell retrotransposition assay[28,29]. A retrotransposition indicator cassette[30], consisting of a reporter gene in the opposite transcriptional orientation to the L1 and equipped with its own promoter and polyadenylation (poly-A) signal, is placed into the L1 3'UTR (Figure 1A). The reporter gene is interrupted by a backwards intron, with expression achieved only upon splicing and reverse transcription of the L1 mRNA to deliver an intact copy of the reporter cassette into genomic DNA. Expression and polyadenylation of the full-length L1 mRNA can be augmented by a strong heterologous promoter and poly-A signal flanking the L1 element and reporter cassette[28,31]. In selection-based assays, quantification of antibiotic-resistant foci provides a readout of retrotransposition efficiency[28].

The adenosine richness of L1 ORFs limits L1 expression in mammalian cells and influences expression of genes containing intronic sense oriented L1 insertions[32]. A synthetic mouse L1 element

(smL1, or ORFeus-Mm) was developed to increase the retrotransposition efficiency of engineered L1 reporter constructs, improving their utility for random mutagenesis screens of mammalian genomes[33]. smL1 was derived from the L1_T_F subfamily element L1spa[4,34] whereby the adenosine content of the L1spa ORFs was reduced from 40% to 26% via synonymous substitutions. The synthetic ORF sequences of smL1 yielded markedly increased mRNA expression levels, and smL1 retrotransposed ~200-fold more efficiently in human HeLa cells than the retrotransposition indicator construct pTN201/L1spa[4,33]. Here, we note both coding and non-coding differences between smL1 and pTN201/L1spa, and we systematically query the impact of these differences, and of synthetic ORF sequences, on mouse L1 retrotransposition efficiency.

Results: We capillary sequenced the pCEP4-based smL1 and pTN201/L1spa plasmids (Supplemental Table 1), and observed that the synthetic sequences of the smL1 ORFs largely comprise synonymous nucleotide substitutions relative to L1spa. However, we noticed that smL1 ORF1 contains two nonsynonymous substitutions (G53D and H159D) and bears an engineered Kozak consensus at its initiation codon that is absent from pTN201/L1spa. Furthermore, the terminal 159 bp of 3'UTR sequence present in pTN201/L1spa construct is absent from smL1 (Figure 1B). We refer to this 159 bp sequence as the guanine-rich region (GRR), as it encompasses the conserved poly-purine tract. We traced this deletion to the undetected loss of a NdeI restriction fragment during subcloning of the smL1 construct (Figure 1—figure supplement 1). The remaining 3'UTR sequence of smL1 contains 19 nucleotide substitutions and two single nucleotide deletions relative to the L1spa 3'UTR, which arose during synthesis of the smL1 construct (Figure 1—figure supplement 2).

The L1_T_F subfamily consensus, and previously described active mouse L1_T_F elements, contain aspartic acid at ORF1p positions 53 and 159[4,5,24,35-37]. Indeed, the substitution H159D improves L1spa ORF1p nucleic acid chaperone activity and retrotransposition efficiency[38]. To quantify the contribution of G53D and H159D to the increased retrotransposition efficiency of smL1 over L1spa, we generated L1spa_G53D, L1spa_H159D, and a construct with both changes, L1spa_Dbl (Figure 1C).

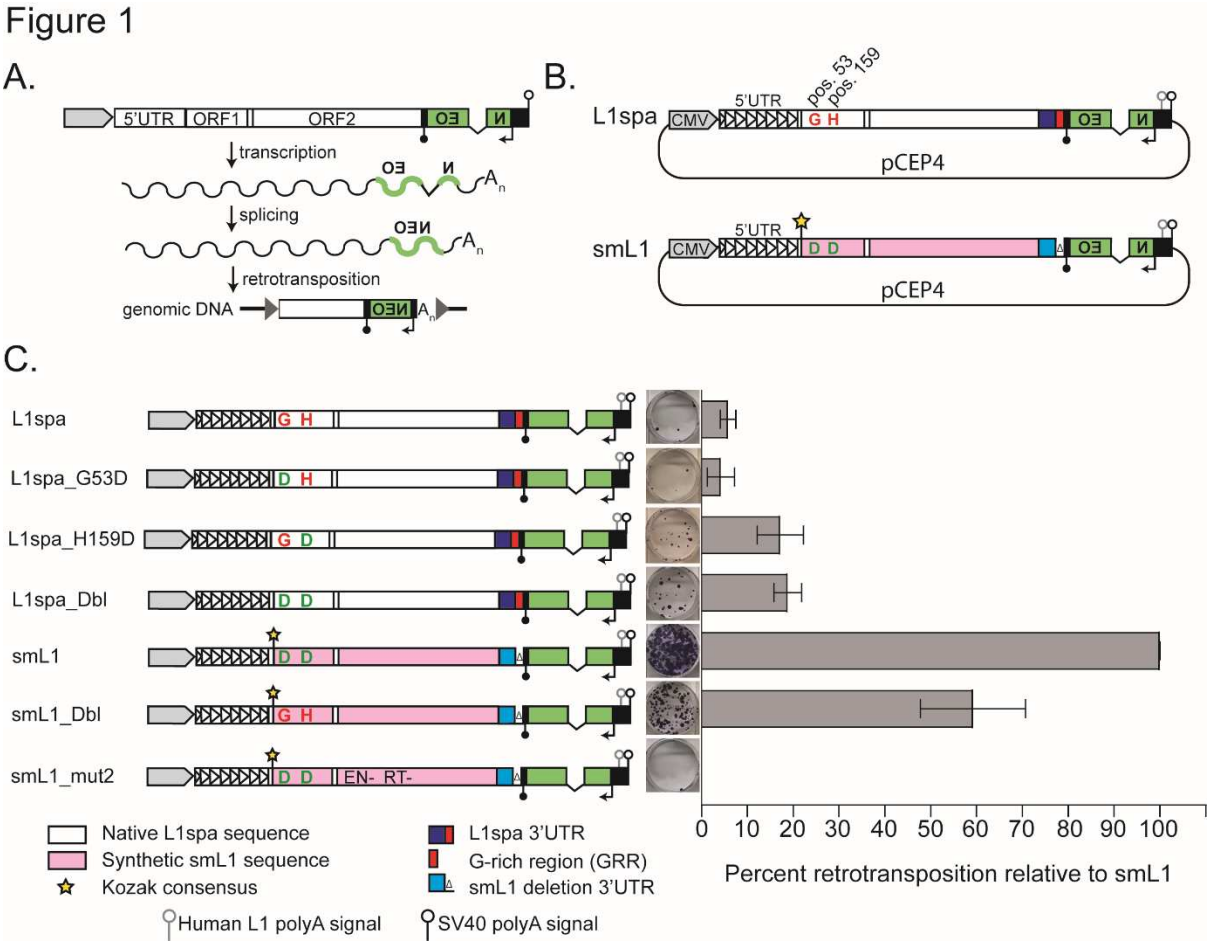


Figure 1. L1spa and smL1 differ at two ORF1p amino acid positions.

A. The cultured cell retrotransposition assay. A full-length L1 containing intact open reading frames (ORF1 and ORF2) is driven by its native 5'UTR promoter or a strong heterologous promoter (grey), and equipped with a polyadenylation signal (native and/ or engineered, open lollipop). The L1 is tagged with a retrotransposition indicator cassette (green) inserted in its 3'UTR. The reporter consists of a backwards intron-containing neomycin phosphotransferase (NEO) reporter gene in opposite transcriptional orientation to the L1 and equipped with a separate promoter and polyadenylation signal (black arrow and black lollipop, respectively). Expression of the reporter gene from genomic DNA is achieved upon retrotransposition.

B. L1spa/pTN201 and smL1 plasmids. Features are depicted as follows: CMV promoter (grey), L1spa 5'UTR (white triangles, representing monomers). NEO reporter cassette (green) with SV40 early promoter (black arrow) and HSV-Tk polyadenylation signal (filled black lollipop). Native L1spa nucleotide sequence (white); synthetic sequence (pink). The amino acids at positions 53 and 159 in ORF1p are indicated. Engineered Kozak consensus sequence (star). L1spa 3'UTR (dark blue); G-rich region (red). Mutated smL1 3'UTR (light blue), deleted GRR (Δ symbol). Human L1 polyadenylation signal (open grey lollipop) and the SV40 polyadenylation signal (open black lollipop).

C. Effects of ORF1p amino acid changes on retrotransposition. From left to right: a schematic of each construct (as described in B), a representative well from the retrotransposition assay, and its retrotransposition efficiency relative to smL1. The histogram displays the mean and standard deviation of three biological replicate assays, each comprising three technical replicates per construct.

In a HeLa cell transient transfection-based retrotransposition assay[29], L1spa mobilized at ~6% the efficiency of smL1, and, consistent with previous results[38], the substitution G53D had little effect on retrotransposition. However, L1spa_H159D and L1spa_Dbl mobilized at ~19% the efficiency of smL1. The reciprocal changes (D53G and D159H) reduced smL1 retrotransposition efficiency by ~40% (smL1_Dbl; Figure 1). Thus, the H159D amino acid substitution partially explains the increased retrotransposition efficiency of smL1 over L1spa.

To determine the contribution of non-coding changes between smL1 and L1spa_Dbl to their differing retrotransposition efficiencies, we equipped L1spa_Dbl ORF1 with a Kozak consensus and replaced the L1spa 3'UTR with the deletion-containing smL1 3'UTR to generate L1spa_Dbl_smBB. Despite entirely lacking synthetic ORF sequences, L1spa_Dbl_smBB retrotransposed equally to smL1 (Figure 2A). When tested independently, removing the Kozak consensus from this construct (L1spa_Dbl_smBB_delete_Kozak) had no impact on retrotransposition efficiency. However, restoring the GRR-containing L1spa 3'UTR (L1spa_Dbl_smBB_restore_3'UTR) reduced retrotransposition to similar levels as L1spa_Dbl (Figure 2A). We therefore concluded that the deletion-containing smL1 3'UTR largely accounts for the increased retrotransposition efficiency of smL1 over L1spa.

We next asked whether placement of the GRR relative to the NEO cassette impacts retrotransposition efficiency. We replaced the L1spa GRR into L1spa_Dbl_SMBB, inserting it downstream of NEO but upstream of the SV40 poly-A signal, to generate L1spa_Dbl_smBB_GRRAN. Strikingly, this construct retrotransposed with equal efficiency to smL1 and L1spa_Dbl_smBB (Figure 2). To verify that insertions arising from this rearranged construct represented bona-fide retrotransposition events, we characterized 10 L1spa_Dbl_smBB_GRRAN integrants by inverse PCR. All 10 insertions incorporated L1 sequence, the spliced NEO cassette, and the GRR, and all insertions terminated in poly-A tracts directed by the SV40 poly-A signal (Figure 2B, Figure 2--figure supplement 1, Supplemental Table 1). Like smL1 insertions previously characterized in HeLa cells[33], the L1spa_Dbl_smBB_GRRAN insertions were variably 5' truncated, bore TSDs (1 to 365 bp) or, in one case, a small target-site deletion, and predominantly occurred at sequences resembling the

Figure 2

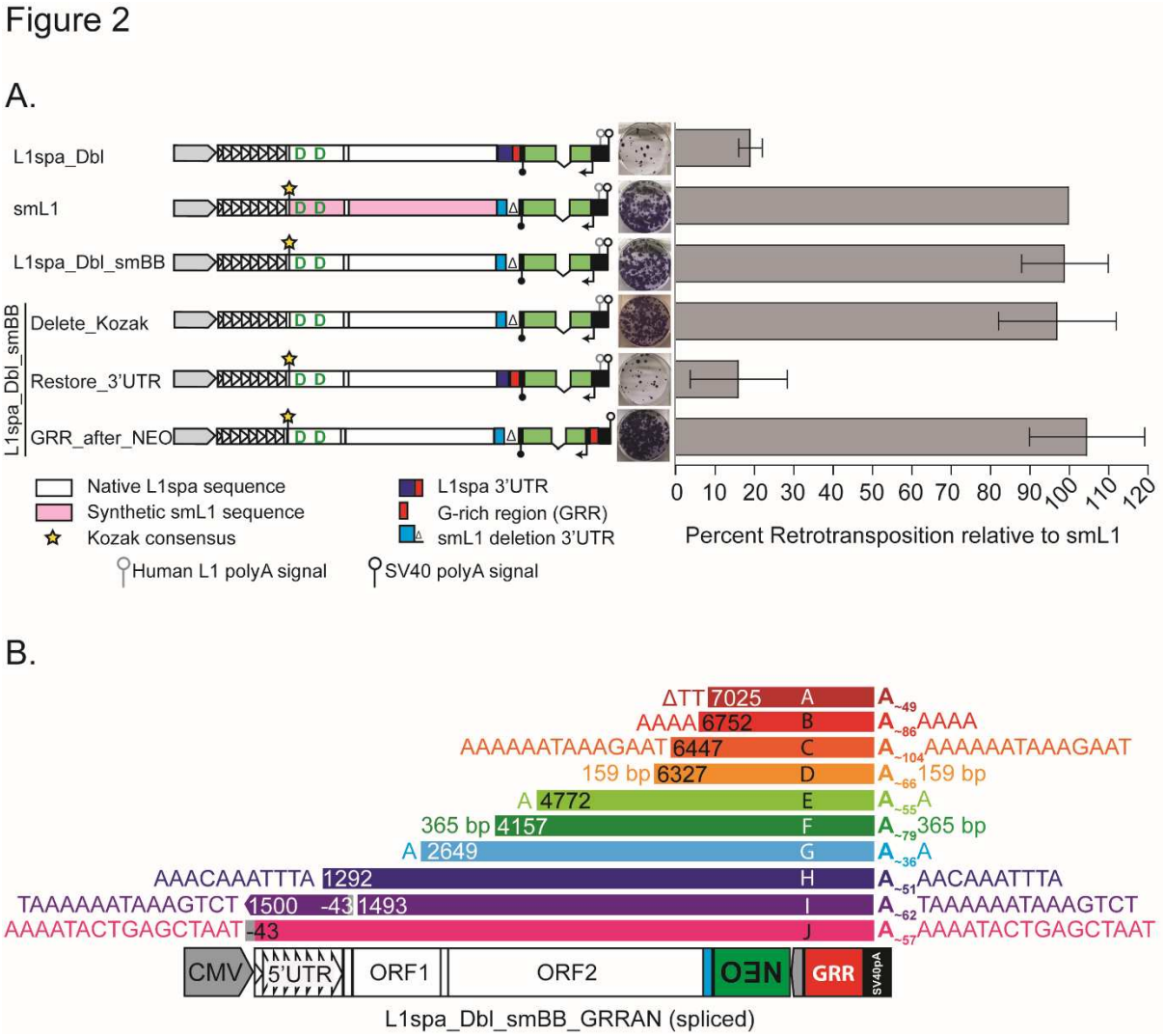


Figure 2. The presence and position of the L1spa 3'UTR G-rich region impact retrotransposition.

A. From left to right: a schematic of each construct (as described in Figure 1), a representative well from the retrotransposition assay, and its retrotransposition efficiency relative to smL1. L1spa_Dbl_smBB contains the ORF1 and ORF2 sequences of L1spa_Double with a Kozak consensus at the ORF1p initiation codon (as in smL1), and the deletion 3'UTR of the smL1 construct. Restore_3'UTR is equivalent to L1spa_Dbl_smBB but with the full L1spa 3'UTR restored upstream of the NEO cassette. GRR_after_NEO is equivalent to L1spa_Dbl_smBB but with the G-rich region of the L1spa 3'UTR placed downstream of the NEO cassette, and the human L1 polyadenylation signal removed. The histogram displays the mean and standard deviation of three biological replicate assays, each comprising three technical replicates per construct.

B. Structural hallmarks of ten insertions generated with the GRR_after_NEO construct are shown, including the 5' TSD, 5' truncation position relative to the L1spaDbl_smBB_GRRAN, polyadenylation position, poly-A tail length and 3'TSD. Below, the structure of GRR after NEO, with the spliced NEO cassette in green, the GRR in red, and the SV40 polyadenylation signal in black. Schematic is not to scale. Delta symbol indicates a target-site deletion. Insertion I contained a 5' inversion for which the break-points are indicated. Insertions I and J arose from a transcript initiating upstream of the L1 5'UTR, at the transcription start site of the CMV promoter within the pCEP4 vector[56] denoted as position -43. For two insertions bearing >100 bp TSDs, the TSD length, rather than TSD sequence, is shown.

degenerate consensus 5'-TTTT/AA-3'. Thus, situating the GRR downstream of the NEO cassette greatly improves the engineered retrotransposition efficiency of a native mouse L1 and generates structurally typical L1 insertions.

Discussion: The highly active smL1 construct and its derivatives have facilitated numerous insights into the mechanism, regulation, and consequences of L1 retrotransposition in vivo and in vitro[39-48]. Here we demonstrate that synthetic reduction in mouse L1 ORF adenosine content has little impact on retrotransposition efficiency, at least in human HeLa cells. We find instead that the elevated activity of smL1 over pTN201/L1spa arises partly from the nonsynonymous ORF1 substitution H159D[38], and to a larger extent from the deletion of the L1spa GRR from the smL1 3'UTR. Our results resolve a long-standing incongruity regarding the impact of synthetic ORF sequences on L1 retrotransposition efficiency, as the similarly recoded synthetic human L1 ORFeus_Hs retrotransposes no more than 3-fold more efficiently than the native parental element $L1_{RP}$[49,50]. Thus, while both human and mouse synthetic L1 ORFs yield massively increased mRNA levels relative to native ORF sequences, this elevated expression does not effect a proportionate increase in retrotransposition[33,50]. This result is consistent with the modest increase in retrotransposition achieved by augmenting L1 expression with a strong heterologous promoter[28], and suggests that L1 mRNA levels are not the rate-limiting determinant of retrotransposition efficiency.

The molecular mechanism by which deleting or repositioning the L1spa GRR impacts retrotransposition efficiency remains unclear. Importantly, the impaired retrotransposition efficiency observed when the L1spa GRR is placed upstream of the NEO cassette does not extend to human L1 reporter constructs. In contrast, deletions encompassing the human poly-purine tract from NEO- and EGFP-based L1 constructs, in which this motif is also positioned upstream of the reporter cassette, reduced retrotransposition by ~10% and ~30%, respectively[28,51]. Furthermore, experimental evidence suggests that the human poly-purine tract can form a G-quadruplex (G4) secondary structure[14,52] and that stabilization of this structure stimulates engineered retrotransposition[51]. The decreased

retrotransposition efficiency of constructs incorporating the L1spa GRR upstream of the NEO cassette may represent an artifact specific to the L1spa GRR sequence composition and its ability to form secondary structures in the context of the reporter construct[14], rather than evidence that this motif intrinsically inhibits retrotransposition. Indeed, the possibility that disruption of the L1 3'UTR by an indicator cassette could impact retrotransposition has long been acknowledged[53], and the conservation of a self-attenuating motif is difficult to reconcile with the evolutionary imperative of L1 as a selfish genetic element[14,15].

As we observe no difference in retrotransposition efficiency when the L1spa GRR is deleted entirely compared to when it is repositioned downstream of the NEO cassette, we speculate that a potential function for the GRR in endogenous mouse L1 retrotransposition is likely to be subtle, perhaps influenced by the stabilization of RNA or DNA G4 secondary structures and dependent on the cellular milieu. The extent to which the deletion of the GRR from the smL1 construct and its derivatives impacts the conclusions drawn using these reporters remains to be evaluated[39-48]. Engineered mouse L1 constructs reconfigured with the GRR downstream of the reporter cassette represent useful tools for investigating the function of this conserved motif, and for studies of mouse L1 biology in general, as high efficiency retrotransposition and structurally normal L1 integrants can be achieved using a construct containing the entire native mouse L1 sequence.

Materials and Methods:

Constructs used in this study:

pTN201/L1spa: Has been described previously[4]. It consists of the pCEP4 backbone with the element L1spa truncated at position 7418 in the 3'UTR (removing the native mouse L1 polyadenylation signal but leaving the G-rich region (GRR) of the 3'UTR in place). The L1spa 3'UTR is followed by the mneoI retrotransposition indicator cassette[28,30]. The mneoI cassette is followed by the human L1

polyadenylation signal which is directly upstream of the SV40 polyadenylation signal in the pCEP4 backbone.

L1spa_G53D: Identical to pTN201/L1spa but with glycine 53 of ORF1 replaced by aspartic acid.

L1spa_H159D: Identical to pTN201/L1spa but with histidine 159 of ORF1 replaced by aspartic acid.

L1spa_dbl: Identical to pTN201/L1spa but with glycine 53 of ORF1 replaced by aspartic acid and histidine 159 of ORF1 replaced by aspartic acid.

smL1: Has been described previously[33]. It consists of the pCEP4 backbone with a synthetic mouse L1 element based on L1spa. The initiation codon of ORF1 has been placed into an engineered Kozak consensus. The amino acid sequence of smL1 is identical to that of L1spa except for two residues in ORF1 (position 53 and position 159). The 3'UTR of smL1 contains several single nucleotide substitutions relative to L1spa, as well as two single-nucleotide deletions, and is truncated at position 7259 of L1spa. This truncation deletes the GRR of the L1spa 3'UTR. The smL1 3'UTR is followed by the mneoI retrotransposition indicator cassette. The mneoI cassette is followed by the human L1 polyadenylation signal which is directly upstream of the SV40 polyadenylation signal in the pCEP4 backbone.

smL1_mut2: Has been described previously[33]. It is identical to smL1 but with inactivating mutations in the endonuclease and reverse transcriptase active sites of ORF2.

smL1_Dbl: Identical to smL1 but with aspartic acid at position 53 of ORF1 replaced by glycine and aspartic acid at position 159 of ORF1 replaced by histidine.

L1spa_dbl_smBB: "L1spa double; smL1 backbone". It consists of the L1spa_dbl sequence, with a Kozak consensus at the initiation codon of ORF1 and the mutated 3'UTR of smL1.

L1spa_dbl_smBB _Del_Kozak: Identical to L1spa_dbl_smBB but with the Kozak consensus removed.

L1spa_dbl_smBB_Restore_3'UTR: Identical to L1spa_dbl_smBB but with the mutated 3'UTR of L1SM replaced by the native L1spa 3'UTR as found in pTN201/L1spa.

L1spa_Dbl_smBB_GRRAN: "L1spa double; smL1 backbone; G-rich region after NEO". Identical to L1spa_dbl_smBB but with the GRR of L1spa (nucleotides 7260-7416) and an engineered AflII restriction site placed after the mneoI cassette and before the SV40 polyadenylation signal of the pCEP4 backbone, and the human L1 polyadenylation signal deleted.

pCEP4_GFP: Consists of the pCEP4 backbone with a humanized renilla GFP reporter gene.

Construct generation: all cloning was performed using restriction enzymes from New England Biolabs (NEB) according to manufacturer's instructions. PCR fragments were generated using oligonucleotides ordered from Integrated DNA Technologies (IDT) and Q5 High Fidelity DNA Polymerase 2X master mix (NEB), according to manufacturer's instructions. Constructs were thoroughly capillary sequenced to confirm the absence of random PCR-induced mutations.

Cell culture: HeLa-JVM cells[28] were maintained in Dulbecco's Modified Eagle's Medium (DMEM) high glucose supplemented with 1% L-glutamine, 1% pen-strep and 10% fetal bovine serum. Cells were passaged every 3-5 days at 70-80% confluence using 0.25% Trypsin-EDTA (Life Technologies). Cells were grown at 37°C in 5% CO_2.

Retrotransposition assays: For transient retrotransposition assays performed in 6-well plates as previously described[29], $5x10^3$ HeLa-JVM cells were plated per well. 24 hours later, each well was transfected with 1 µg of plasmid DNA using 4 ul FuGene and 96 uL opti-MEM. Each construct was transfected in triplicate. The media was replaced 24 hours post-transfection. At 72 hours post-transfection, selection was initiated with 400 µg/ml G418 (Geneticin, Life Technologies). G418 media was replaced every other day. At 12 days post-transfection, cells were fixed with 2% formaldehyde/0.4% glutaraldehyde solution and stained with 0.1% crystal violet solution.

Assays to measure transfection efficiency for each L1 reporter plasmid were carried out as follows: $4x10^4$ HeLa cells were plated per well of a 6-well dish. 24 hours later, each well was transfected with 1 µg total plasmid DNA, consisting of 0.5 µg L1 reporter plasmid and 0.5 µg pCEP4-GFP, using 4 ul of FuGene and 96 ul opti-MEM. Each L1 reporter+GFP plasmid was transfected in duplicate. The media was replaced 24 hours post-transfection. At 72 hours post-transfection, cells were trypsinized, resuspended in PBS (Life Technologies) and assessed for GFP expression by flow cytometry using a Cytoflex flow cytometer (Translational Research Institute flow cytometry core). The average percent GFP-positive cells was determined for each L1 reporter construct, and used to normalize the results of the retrotransposition assay.

Clonal cell lines for inverse PCR characterization of L1spa_Dbl_smBB_GRRAN insertions were generated by plating $5x10^3$ HeLa cells per 10 cm dish, and 18 hours later transfecting with 1 ug plasmid DNA using 4 ul FuGene and 96 ul opti-MEM. Media was replaced 24 hours after transfection. Selection with 400 ug/ml Geneticin was initiated at 72 hours post-transfection and continued for 12 days to generate isolated colonies harboring retrotransposition events. Individual colonies were manually picked and expanded to generate clonal cell lines. Approximately $5x10^6$ cells were harvested per clonal cell line and subjected to phenol-chloroform DNA extraction. DNA concentration was measured via NanoDrop. Inverse PCR was carried out as described previously[54,55]. Briefly, 4 ug from each cell line was digested with 25 units EcoRI (New England Biolabs) in 100 ul total volume. Reactions were heat inactivated and digested DNA was ligated under dilute conditions to promote intramolecular ligation (1 ml total volume) using 3200 units T4 DNA ligase (NEB) for 24 hours. Ligations were purified by chloroform extraction and DNA was resuspended in 40 ul water. For each ligation, 4 ul of DNA was used as a template in the first-round inverse PCR reaction using the Roche Expand long-template PCR system. Each reaction consisted of 10 pmol primer NEO210as (5' GACCGCTTCCTCGTGCTTTACG 3), 10 pmol primer NEO1720s (5' TGCGCTGACAGCCGGAACACG 3'), 20 nM each dNTP, and 2.5 units of enzyme. Cycling conditions were 95°C for 2 minutes, followed by 30 cycles of 94°C for 15 seconds, 64°C for 30s, and 68°C for 15 min, followed by a 30 min extension at 68°C. For the second-round

inverse PCR reaction, 4 ul of the product from the first-round inverse PCR was used directly as template, with 10 pmol primer NEO173as (5' CATCGCCTTCTATCGCCTTCTTG 3') 10 pmol primer NEO1808s (5' GCGTGCAATCCATCTTGTTCAATG 3'), 20 nM each dNTP, and 2.5 units of enzyme. Cycling conditions were identical to first-round inverse PCR. Products were run on a 1% agarose gel and bands were excised and purified using the Qiagen Min-Elute gel extraction kit. Bands <2kb were cloned using the pGEM-T Easy kit (Promega) and bands >2kb were cloned using the Topo-XL2 PCR kit (Life Technologies) following manufacturer's instructions. Cloned PCR products were sequenced using M13 forward and reverse primers Australian Genome Research Facility, Brisbane). Validation primers (Integrated DNA Technologies) were designed based on the putative genomic location of each insertion detected by inverse PCR, and insertions were validated as empty/filled and/or by 5' and 3' junction PCRs using MyTaq HS DNA polymerase (Bioline). Validation products were run on 1% agarose gels, purified using the Qiagen Min-Elute gel extraction kit, and capillary sequenced (Australian Genome Research Facility, Brisbane) to verify the genomic location and structural hallmarks of each insertion.

Acknowledgements: HeLa-JVM cells were a gift from John V. Moran. The pTN201/L1spa construct was a gift from Haig H. Kazazian. We thank all members of the Faulkner laboratory for helpful discussion and critical reading of the manuscript. This study was funded by the Australian National Health and Medical Research Council (NHMRC) (GNT1125645, GNT1138795 and GNT1173711 to G.J.F.; GNT1173476 to S.R.R.), The Australian Research Council (ARC) (DP200102919 to G.J.F. and S.R.R.), a CSL Centenary Fellowship to G.J.F., an Advance Queensland Women's Academic Fund Maternity Funding award to S.R.R., the Mater Research Strategic Grant for Outstanding Women to S.R.R, the Mater Foundation (Equity Trustees / AE Hingeley, QFC Thomas George and KC BM Thomson Trusts), an Australian Government Research Training Program Scholarship to P.G. and a Mater Research Frank Clair Scholarship to P.G.. J.D.B. is a Founder and Director of CDI Labs, Inc., a Founder of Neochromosome, Inc, a Founder and SAB member of ReOpen Diagnostics, and serves or served on

the Scientific Advisory Board of the following: Sangamo, Inc., Modern Meadow, Inc., Sample6, Inc. and the Wyss Institute.

References:

1 Richardson, S. R. *et al.* The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiology Spectrum* **3**, doi:10.1128/microbiolspec.MDNA3-0061-2014 (2015).

2 Kazazian, H. H., Jr. & Moran, J. V. Mobile DNA in Health and Disease. *N Engl J Med* **377**, 361-370, doi:10.1056/NEJMra1510092 (2017).

3 Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H., Jr. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* **11**, 1677-1685, doi:10.1101/gr.198301 (2001).

4 Naas, T. P. *et al.* An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J* **17**, 590-597, doi:10.1093/emboj/17.2.590 (1998).

5 Sookdeo, A., Hepp, C. M., McClure, M. A. & Boissinot, S. Revisiting the evolution of mouse LINE-1 in the genomic era. *Mob DNA* **4**, 3, doi:10.1186/1759-8753-4-3 (2013).

6 DeBerardinis, R. J., Goodier, J. L., Ostertag, E. M. & Kazazian, H. H., Jr. Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat Genet* **20**, 288-290, doi:10.1038/3104 (1998).

7 Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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, doi:0092-8674(93)90078-5 [pii] (1993).

8 Feng, Q., Moran, J. V., Kazazian, H. H., Jr. & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-916, doi:S0092-8674(00)81997-2 [pii] (1996).

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

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

11 Kolosha, V. O. & Martin, S. L. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J Biol Chem* **278**, 8112-8117, doi:10.1074/jbc.M210487200 M210487200 [pii] (2003).

12 Khazina, E. & Weichenrieder, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc Natl Acad Sci U S A* **106**, 731-736, doi:0809964106 [pii] 10.1073/pnas.0809964106 (2009).

13 Khazina, E. *et al.* Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat Struct Mol Biol* **18**, 1006-1014, doi:10.1038/nsmb.2097 (2011).

14 Howell, R. & Usdin, K. The ability to form intrastrand tetraplexes is an evolutionarily conserved feature of the 3' end of L1 retrotransposons. *Mol Biol Evol* **14**, 144-155, doi:10.1093/oxfordjournals.molbev.a025747 (1997).

15 Boissinot, S. & Sookdeo, A. The Evolution of LINE-1 in Vertebrates. *Genome Biol Evol* **8**, 3485-3507, doi:10.1093/gbe/evw247 (2016).

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

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

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

19 Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94**, 1872-1877 (1997).

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

21 Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-7795, doi:25/17/7780 [pii] 10.1128/MCB.25.17.7780-7795.2005 (2005).

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

23 Sultana, T. *et al.* The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol Cell* **74**, 555-570 e557, doi:10.1016/j.molcel.2019.02.036 (2019).

24 Richardson, S. R. *et al.* Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res* **27**, 1395-1405, doi:10.1101/gr.219022.116 (2017).

25 Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**, 1567-1577, doi:10.1101/gr.247965.118 (2019).

26 Hancks, D. C. & Kazazian, H. H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9, doi:10.1186/s13100-016-0065-9 (2016).

27 Gagnier, L., Belancio, V. P. & Mager, D. L. Mouse germ line mutations due to retrotransposon insertions. *Mob DNA* **10**, 15, doi:10.1186/s13100-019-0157-4 (2019).

28 Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-927, doi:S0092-8674(00)81998-4 [pii] (1996).

29 Wei, W., Morrish, T. A., Alisch, R. S. & Moran, J. V. A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. *Anal Biochem* **284**, 435-438, doi:10.1006/abio.2000.4675 S0003-2697(00)94675-X [pii] (2000).

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

31 Rangwala, S. H. & Kazazian, H. H., Jr. The L1 retrotransposition assay: a retrospective and toolkit. *Methods* **49**, 219-226, doi:S1046-2023(09)00095-4 [pii]10.1016/j.ymeth.2009.04.012 (2009).

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

33 Han, J. S. & Boeke, J. D. A highly active synthetic mammalian retrotransposon. *Nature* **429**, 314-318, doi:10.1038/nature02535 nature02535 [pii] (2004).

34 Kingsmore, S. F. *et al.* Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. *Nat Genet* **7**, 136-141, doi:10.1038/ng0694-136 (1994).

35 Schauer, S. N. *et al.* L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Research* **28**, doi:10.1101/gr.226993.117 (2018).

36 Penzkofer, T. *et al.* L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res* **45**, D68-D73, doi:10.1093/nar/gkw925 (2017).

37 Zhou, M. & Smith, A. D. Subtype classification and functional annotation of L1Md retrotransposon promoters. *Mob DNA* **10**, 14, doi:10.1186/s13100-019-0156-5 (2019).

38 Martin, S. L. *et al.* A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. *Nucleic Acids Res* **36**, 5845-5854, doi:10.1093/nar/gkn554 (2008).

39 Newkirk, S. J. *et al.* Intact piRNA pathway prevents L1 mobilization in male meiosis. *Proc Natl Acad Sci U S A* **114**, E5635-E5644, doi:10.1073/pnas.1701069114 (2017).

40 Kannan, M. *et al.* Dynamic silencing of somatic L1 retrotransposon insertions reflects the developmental and cellular contexts of their genomic integration. *Mob DNA* **8**, 8, doi:10.1186/s13100-017-0091-2 (2017).

41 Richardson, S. R., Narvaiza, I., Planegger, R. A., Weitzman, M. D. & Moran, J. V. APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *eLife* **2014**, doi:10.7554/eLife.02008 (2014).

42 O'Donnell, K. A., An, W., Schrum, C. T., Wheelan, S. J. & Boeke, J. D. Controlled insertional mutagenesis using a LINE-1 (ORFeus) gene-trap mouse model. *Proc Natl Acad Sci U S A* **110**, E2706-2713, doi:10.1073/pnas.1302504110 1302504110 [pii] (2013).

43 Monot, C. *et al.* The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS Genet* **9**, e1003499, doi:10.1371/journal.pgen.1003499 PGENETICS-D-12-02666 [pii] (2013).

44 Xie, Y. *et al.* Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. *Mob DNA* **4**, 10, doi:10.1186/1759-8753-4-10 (2013).

45 Grandi, F. C., Rosser, J. M. & An, W. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol Biol Evol* **30**, 503-512, doi:10.1093/molbev/mss251 (2013).

46 An, W. *et al.* Conditional activation of a single-copy L1 transgene in mice by Cre. *Genesis* **46**, 373-383, doi:10.1002/dvg.20407 (2008).

47 Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H. & Moran, J. V. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* **20**, 210-224, doi:20/2/210 [pii] 10.1101/gad.1380406 (2006).

48 Wagstaff, B. J., Barnerssoi, M. & Roy-Engel, A. M. Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. *PLoS One* **6**, e19672, doi:10.1371/journal.pone.0019672 (2011).

49 Kimberland, M. L. *et al.* Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum Mol Genet* **8**, 1557-1560 (1999).

50 An, W. *et al.* Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs. *Mob DNA* **2**, 2, doi:1759-8753-2-2 [pii] 10.1186/1759-8753-2-2 (2011).

51 Sahakyan, A. B., Murat, P., Mayer, C. & Balasubramanian, S. G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol* **24**, 243-247, doi:10.1038/nsmb.3367 (2017).

52 Lexa, M. *et al.* Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* **15**, 1032, doi:10.1186/1471-2164-15-1032 (2014).

53 Moran, J. V. Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica* **107**, 39-51 (1999).

54 Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31**, 159-165, doi:10.1038/ng898 ng898 [pii] (2002).

55 Li, J. *et al.* Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat Genet* **23**, 348-353, doi:10.1038/15531 (1999).

56 Isomura, H. *et al.* A cis element between the TATA Box and the transcription start site of the major immediate-early promoter of human cytomegalovirus determines efficiency of viral replication. *J Virol* **82**, 849-858, doi:10.1128/JVI.01593-07 (2008).

Supplemental Figure Legends

Figure 1—Figure Supplement 1: Accidental deletion of the GRR during cloning of the smL1 reporter construct. A. The pTN201/L1spa construct [4] and an intermediate pBluescript clone containing the smL1 sequence, with features annotated as in Figure 1. The positions of known NdeI and BamHI sites are shown in black. Our map of pTN201 did not include 15 base pairs of human L1 sequence present in pTN201 immediately following the L1spa 3'UTR, which was carried over from subcloning described in [4] This 15 bp of human L1 sequence contained an unanticipated NdeI site (shown in red).

B) Deletion of the GRR-containing NdeI fragment during cloning. Left: we intended to move an NdeI/BamHI fragment containing the GRR of the L1spa element and the NEO cassette downstream of the smL1 ORFs and 5' portion of the 3'UTR in pBluescript. Right: due to the unanticipated NdeI site, a 180 bp NdeI fragment containing the terminal GRR-containing 159 bp of the L1spa 3'UTR and an engineered PacI site (not shown) was deleted from the resulting construct. Owing to its small size, the deletion was not detected when screening the constructs with diagnostic restriction digests.

Figure 1—figure supplement 2: Alignment of the 3'UTRs and flanking sequence of pTN201/L1spa and smL1. Alignment depicts the 3' end of ORF2 (L1spa ORF2, dark red; smL1 ORF2, bright red), followed by the 3'UTR (L1spa 3'UTR dark blue; smL1 3'UTR, light blue). The smL1 3'UTR deletion is highlighted in yellow, and the runs of guanine bases in the L1spa 3'UTR are highlighted in red. The positions of engineered restriction sites in each construct are shown in fuchsia, and the NdeI sites resulting in the GRR deletion are shown in green. The HSV Tk polyadenylation signal belonging to the NEO cassette (purple) and the second exon of the NEO cassette (orange) are identical between pTN201/L1spa and smL1.

Figure 2—figure supplement 1: Detailed structures of L1spa_Dbl_smBB_GRRAN integrants characterized via inverse PCR. The structure of each insertion is presented as follows: L1spa 5'UTR (triangles), ORFs (white rectangles), first part of the smL1 3'UTR (light blue), spliced NEO cassette (green), sv40 promoter (grey), GRR (red) sv40 polyadenylation signal (purple). The 5' truncation point for each insertion is shown. Insertions I and J arose from a transcript initiating upstream of the L1 5'UTR, at the transcription start site of the CMV promoter within the pCEP4 vector[56] denoted as position -43. Insertion J had a 5' inversion with breakpoint at position 1493 shown. Flanking genomic DNA is shown to the right and left, with target-site duplications in red. The empty site is shown below, with TSD sequence in red, the first-strand cleavage site indicated by a black arrow and the second-strand cleavage site indicated by a grey arrow. Insertion A had a 2 bp target-site deletion, depicted in blue. The chromosomal coordinates of the TSD sequence and the endonuclease cleavage motif for each insertion is shown.

Figure 1--Figure Supplement 1

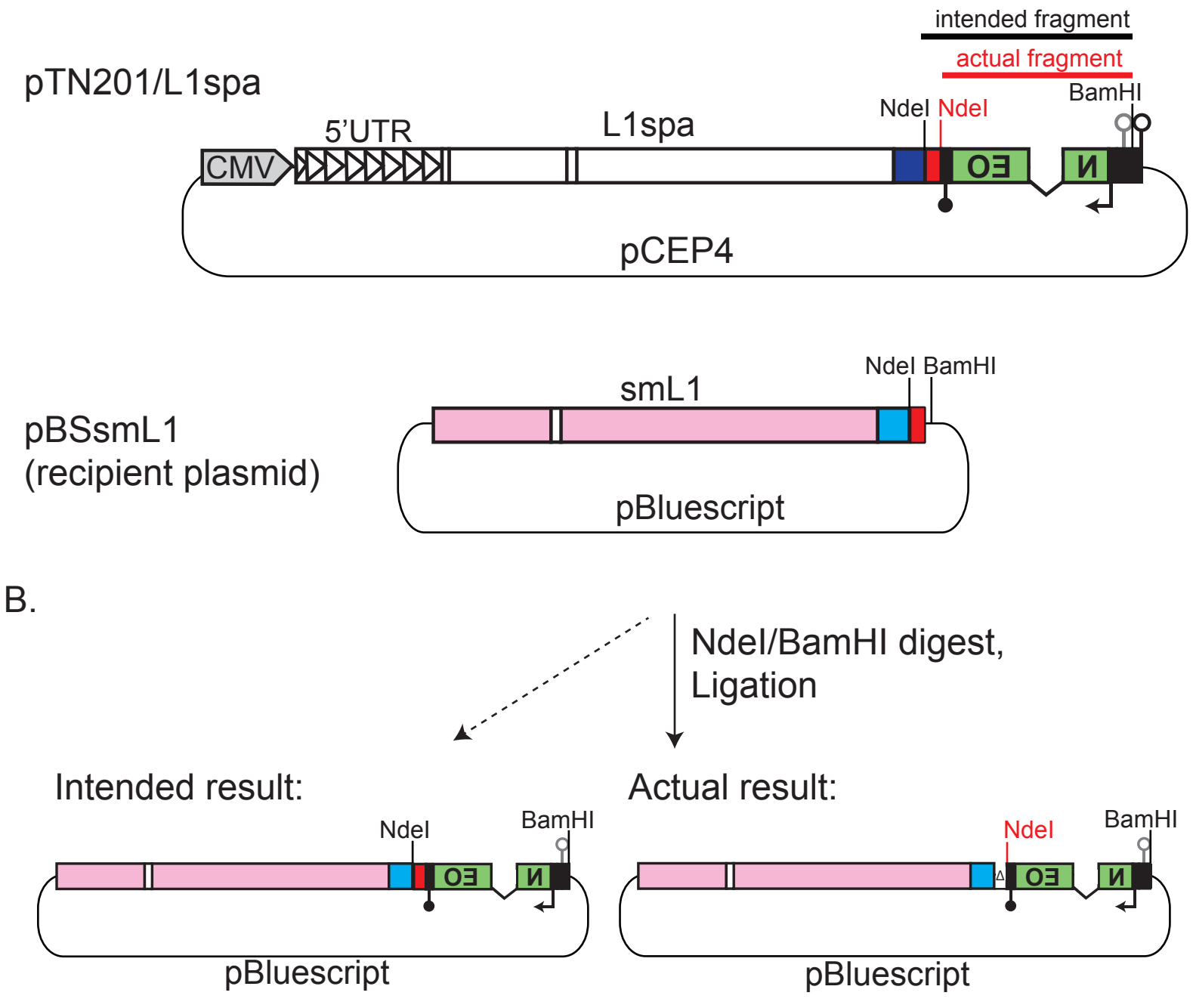


Figure 1—Figure Supplement 2

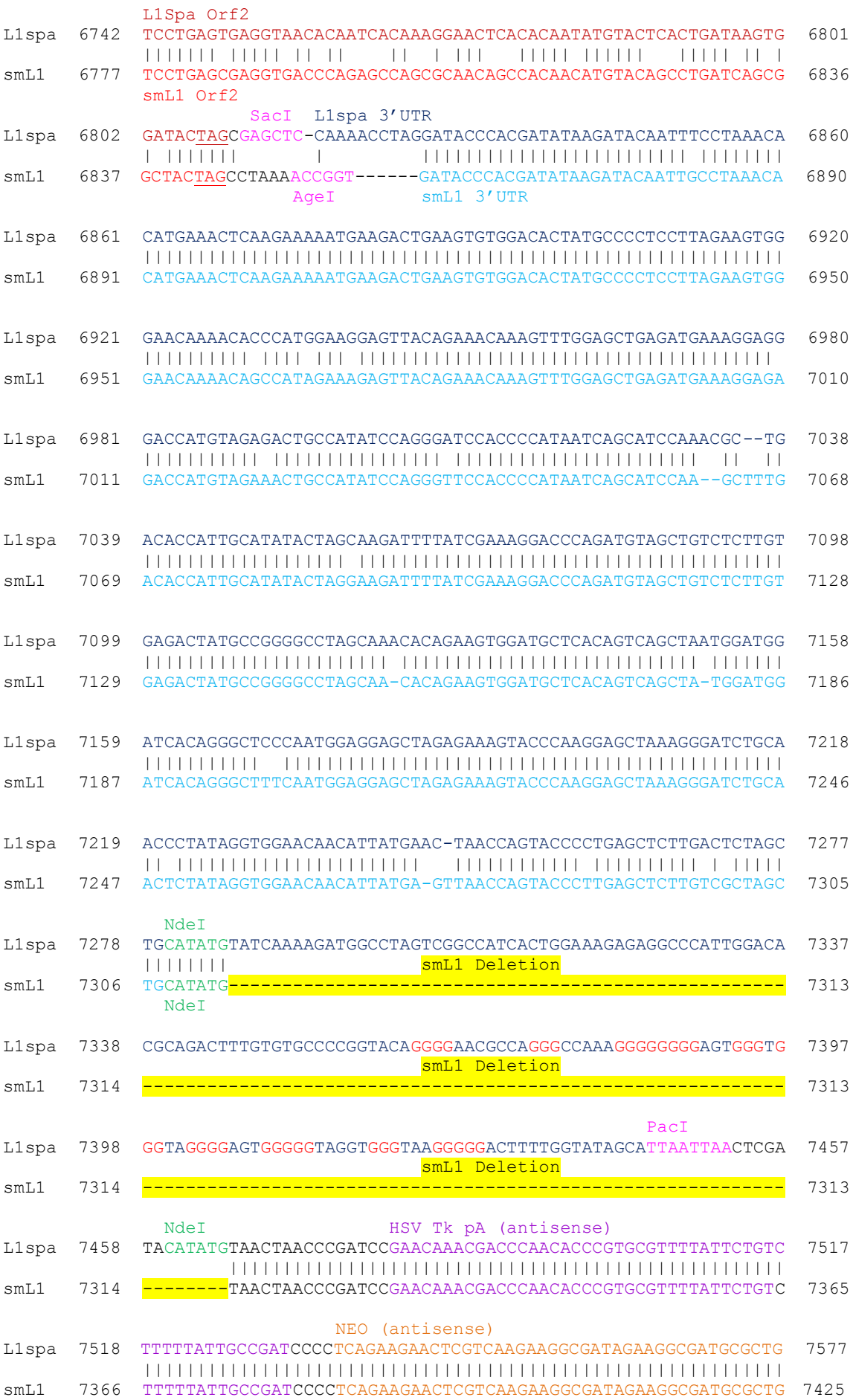
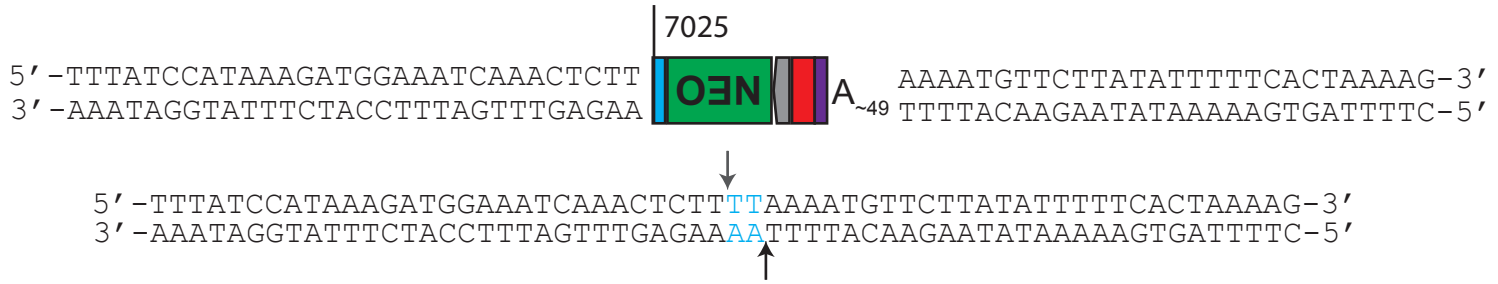


Figure 2--Figure Supplement 1

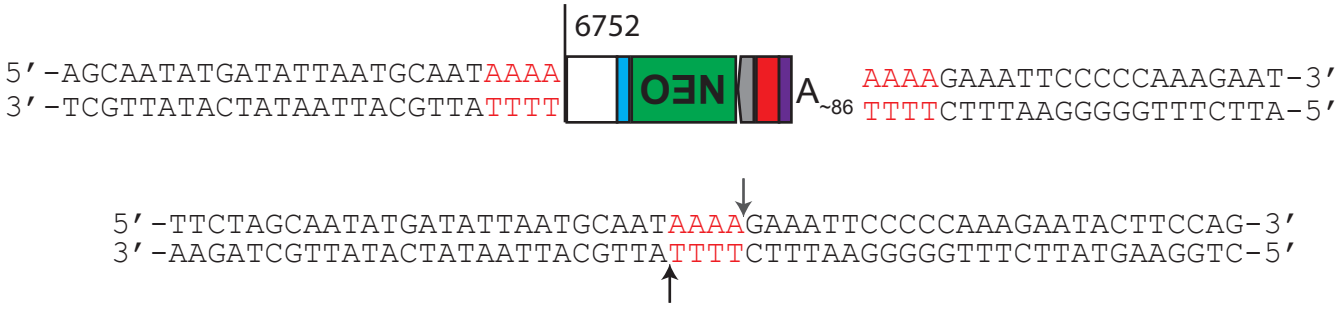
Insertion A



Target-site deletion: Chr3:130997869-130997870

EN cleavage: 5'-TTTT/AA-3'

Insertion B



TSD: Chr3:148918423-148918426

EN cleavage: 5'-TTTT/AT-3'

Insertion C

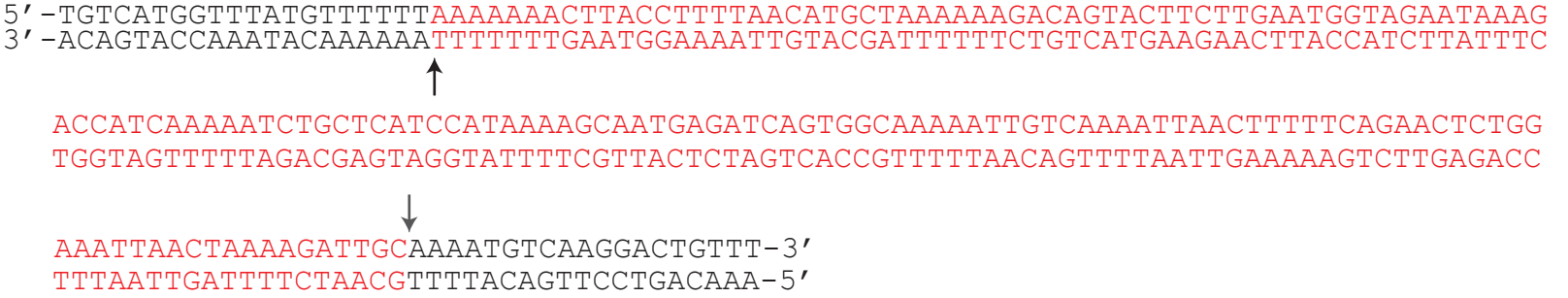
TSD: Chr5: 21539952-21539965

EN cleavage: 5'-TTTT/AT-3'

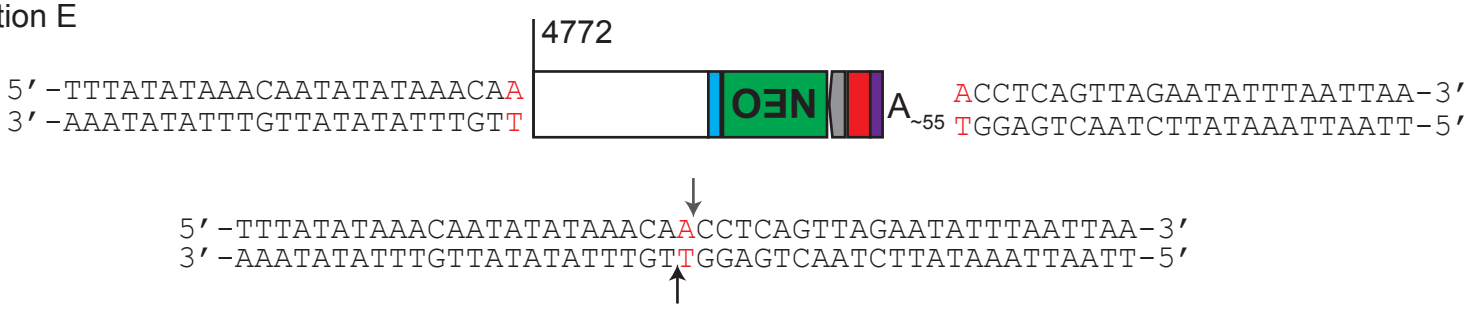
Insertion D

TSD: Chr21: 29204205-29204047

EN cleavage: 5’-TTTT/AA-3’



Insertion E



TSD: chr7: 153197463

EN cleavage: 5’-AGGT/TG-3’

Insertion F

5′-AATTCTTTGGAAAAAAGACAAAGTTGAAATAATGTCAGAGAGAATACATCAGGAAAAAGGTTGAAAATGTAAGTTCTA
3′-TTAAGAAACCTTTTTTTCTGTTTCAACTTTATTACAGTCTCTCTTATGTAGTCCTTTTTTCCAACTTTTACATTCAAGAT

TTAAACACAACATACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGAC
AATTTGTGTTGTATGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTG

TCACATTTTTAATTTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAA
AGTGTAAAAATTAAATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTT

AGAGAATAAAAGACCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATT
TCTCTTATTTTCTGGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAA

GAACTTTACAGTGGTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAAACACATAGATTTCATCGAAG
CTTGAAATGTCACCACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTC

4157

TAAAAACTTCTGTGCATTAAGTAACATTC
ATTTTTGAAGACACGTAATTCATTGTAAG

GAAAATGTAAGTTCTATTAAACACAACAT
CTTTTACATTCAAGATAATTTGTGTTGTA

ACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGACTCACATTTTTAAT
TGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTGAGTGTAAAAATTA

TTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAAAGAGAATAAAAGA
AATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTTTCTCTTATTTTCT

CCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATTGAACTTTACAGTG
GGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAACTTGAAATGTCAC

GTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAAACACATAGATTTCATCGAAGTAAAAACTTCTGT
CACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTCATTTTTGAAGACA

GCATTAAGTAACATTCTCAACAGAGTGAAAAGTCACAGAGGAAATATTTGCAAATCATATATCTGATAAGGGATTAATA-3′
CGTAATTCATTGTAAGAGTTGTCTCACTTTTCAGTGTCTCCTTTATAAACGTTTAGTATATAGACTATTCCCTAATTAT-5′

5′-AATTCTTTGGAAAAAAGACAAAGTTGAAATAATGTCAGAGAGAATACATCAGGAAAAAGGTTGAAAATGTAAGTTCTA
3′-TTAAGAAACCTTTTTTTCTGTTTCAACTTTATTACAGTCTCTCTTATGTAGTCCTTTTTTCCAACTTTTACATTCAAGAT

TTAAACACAACATACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGAC
AATTTGTGTTGTATGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTG

TCACATTTTTAATTTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAA
AGTGTAAAAATTAAATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTT

AGAGAATAAAAGACCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATT
TCTCTTATTTTCTGGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAA

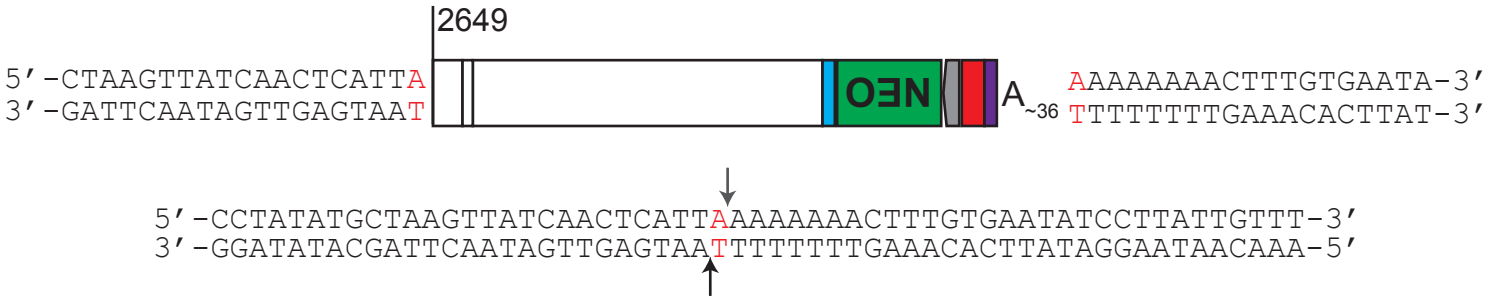
GAACTTTACAGTGGTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAAACACATAGATTTCATCGAAG
CTTGAAATGTCACCACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTC

TAAAAACTTCTGTGCATTAAGTAACATTCTCAACAGAGTGAAAAGTCACAGAGGAAATATTTGCAAATCATATATCTG-3′
ATTTTTGAAGACACGTAATTCATTGTAAGAGTTGTCTCACTTTTCAGTGTCTCCTTTATAAACGTTTAGTATATAGAC-5′

TSD: Chr12:77126920-77127284

EN cleavage: 5'-TTTC/AA3'

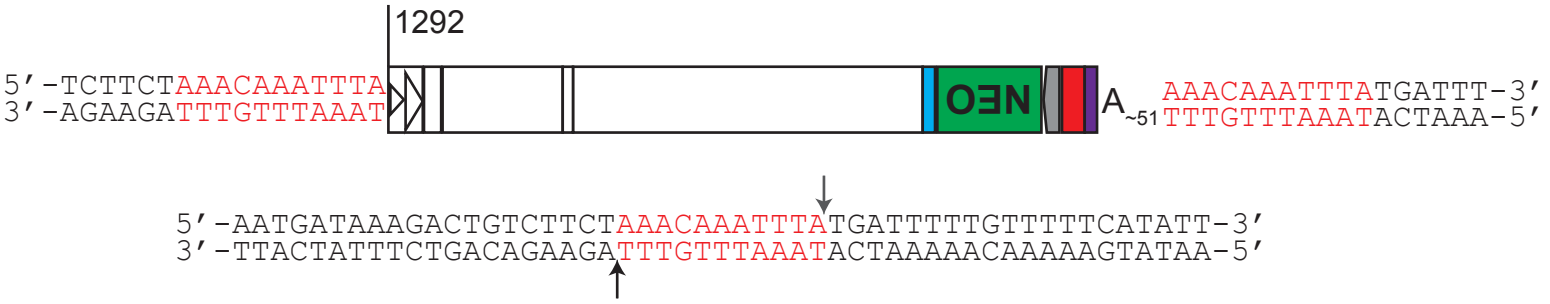
Insertion G



TSD: chr7: 108567840

EN cleavage: 5’-TTTT/AA-3’

Insertion H



TSD: chr3: 143521182-143521191

EN cleavage: 5’-GTTT/AG-3’

Insertion I

TSD: Chr5: 7111072-7111087

EN cleavage: 5’-TTTA/AA-3’

Insertion J

TSD: Chr1: 83883660-83383675

EN cleavage: 5’-TTTT/AC-3’