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

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

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 synthetic mouse L1 (smL1) with ORF adenosine content decreased from 40% to 26% showed increased mRNA expression and retrotransposed far more efficiently than the native parental element, L1spa (1). 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) which could potentially take on a G-quadruplex secondary structure. We find that the combined effect of a single amino acid change and the GRR 3'UTR deletion, rather than synthetic ORF sequences, accounts for the 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 native mouse L1 mobility has in some cases been underestimated in engineered retrotransposition assays.

Author Summary: L1 retrotransposons are mobile DNA elements or "jumping genes" that can copy-and-paste their sequences to new locations in the host genome. The jumping ability, or retrotransposition efficiency, of individual L1 elements can be evaluated using a cultured cell assay in which the L1 is tagged in its 3' untranslated region (3'UTR) with a reporter gene that becomes expressed upon successful retrotransposition. In a previous study, authors Han and Boeke reported that the retrotransposition efficiency of a mouse L1 element could be enhanced dramatically by synthetically increasing the GC content of the L1 ORFs without changing their amino acid sequence. Curiously, a similarly constructed synthetic human L1 achieved only a modest increase in retrotransposition efficiency over the native element. Here, we find that two coding changes and partial deletion of the mouse L1 3'UTR sequence which occurred during construction of the synthetic mouse L1 reporter actually are responsible for the increased jumping of this construct. We also find that changing the placement as well as the presence of this deleted 3'UTR region within the reporter construct determines its impact on engineered retrotransposition efficiency. Together, our study reconciles the disparate impacts of synthetic sequences upon human and mouse L1 retrotransposition efficiency, prompts a reconsideration of numerous studies using synthetic L1 constructs, and will inform the ongoing use of synthetic and natural mouse L1 reporter constructs in vivo and in vitro.

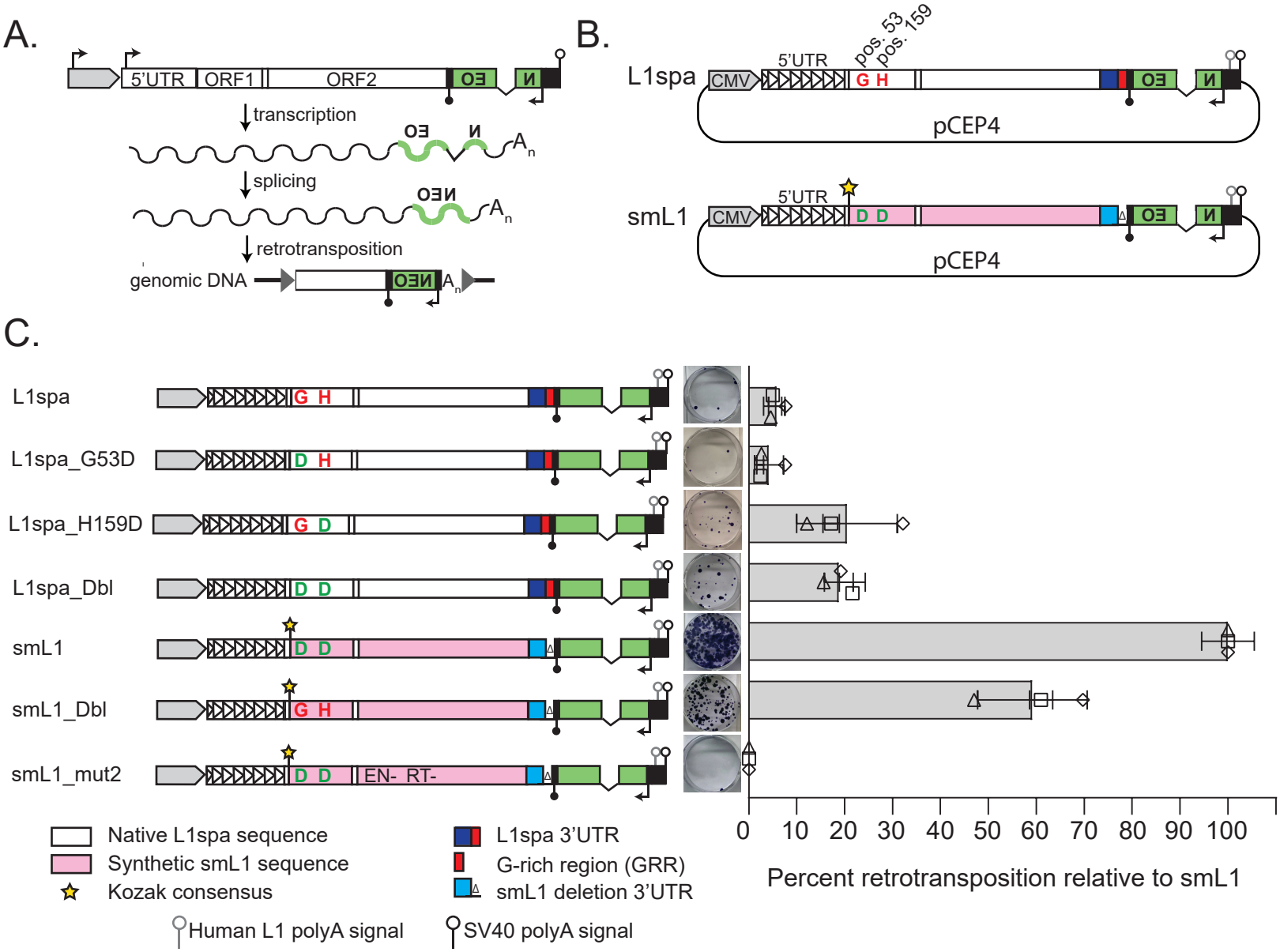
Introduction: L1 retrotransposons are an ongoing source of mutagenesis in mammalian genomes (2, 3). Mice contain ~3000 retrotransposition-competent L1s (RC-L1s) representing three subfamilies (L1_T_F, L1_G_F, L1_A), distinguished by their monomeric promoter-harboring 5'UTRs (4-9). RC-L1s encode two proteins 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) (10-18). 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 (19-23). Hallmarks of TPRT-mediated L1 integrants include insertion at the degenerate consensus sequence 5'-TTTT/AA-3', terminal poly-A tracts, and flanking target site duplications (TSDs) (24-28). Approximately 1 in 8 mice and 1 in 62.5

54 humans harbor a de novo L1 insertion (29, 30). Moreover, >100 cases of human genetic disease and
55 numerous spontaneous mouse mutants have been linked to L1-mediated retrotransposition events,
56 highlighting the impact of L1 mutagenesis on mammalian genomes (31, 32).

57 The retrotransposition efficiency of individual L1 copies can be evaluated in cultured cells and
58 in vivo (33-35). In these assays, a retrotransposition indicator cassette (36), consisting of a reporter
59 gene in the opposite transcriptional orientation to the L1 and equipped with its own promoter and
60 polyadenylation (poly-A) signal, is placed into the L1 3'UTR (Figure 1A). The reporter gene is
61 interrupted by a backwards intron, with expression achieved only upon splicing and reverse
62 transcription of the L1 mRNA to deliver an intact copy of the reporter cassette into genomic DNA.
63 Expression and polyadenylation of the full-length L1 mRNA can be augmented by a strong
64 heterologous promoter and poly-A signal flanking the L1 element and reporter cassette (33, 37). In
65 selection-based cultured cell assays, quantification of antibiotic-resistant foci provides a readout of
66 retrotransposition efficiency (33).

67 The adenosine richness of L1 ORFs limits L1 expression in mammalian cells and influences
68 expression of genes containing intronic sense oriented L1 insertions (38-40). A synthetic mouse L1
69 sequence (smL1, or ORFeus_Mm) was previously developed by Boeke and Han to increase the
70 retrotransposition efficiency of engineered L1 reporter constructs, improving their utility for random
71 mutagenesis screens of mammalian genomes (1). smL1 was derived from the L1_T_F subfamily element
72 L1spa (5, 41) whereby the adenosine content of the L1spa ORFs was reduced from 40% to 26% via
73 synonymous substitutions. The synthetic ORF sequences of smL1 yielded markedly increased mRNA
74 expression levels, and smL1 retrotransposed ~200-fold more efficiently in HeLa cells than the
75 retrotransposition indicator construct pTN201/L1spa (1, 5). Notably, similar constructs containing a
76 synthetic human L1 showed at most a 3-fold increase in retrotransposition efficiency over the native
77 element (42). Here, we note both coding and non-coding differences between smL1 and
78 pTN201/L1spa, and we systematically evaluate the impact of these differences, and of synthetic ORF

Figure 1


A.
5'UTR
ORF1
ORF2
transcription
splicing
retrotransposition
genomic DNA
B.
L1spa
smL1
CMV
5'UTR
pos. 53
pos. 159
G H
D D
pCEP4
C.
L1spa
L1spa_G53D
L1spa_H159D
L1spa_Dbl
smL1
smL1_Dbl
smL1_mut2
EN- RT-
Native L1spa sequence
Synthetic smL1 sequence
Kozak consensus
Human L1 polyA signal
L1spa 3'UTR
G-rich region (GRR)
smL1 deletion 3'UTR
SV40 polyA signal
0 10 20 30 40 50 60 70 80 90 100
Percent retrotransposition relative to smL1

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

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

510 B. L1spa/pTN201 and smL1 plasmids. Features are depicted as follows: CMV promoter (grey), L1spa
511 5'UTR (white triangles, representing monomers). NEO reporter cassette (green) with SV40 early
512 promoter (black arrow) and HSV-Tk polyadenylation signal (filled black lollipop). Native L1spa
513 nucleotide sequence (white); synthetic sequence (pink). The amino acids at positions 53 and 159 in
514 ORF1p are indicated. Engineered Kozak consensus sequence (star). L1spa 3'UTR (dark blue); G-rich
515 region (red). Mutated smL1 3'UTR (light blue), deleted GRR (Δ symbol). Human L1 polyadenylation
516 signal (open grey lollipop) and the SV40 polyadenylation signal (open black lollipop).

517 C. Effects of ORF1p amino acid changes on retrotransposition. From left to right: a schematic of each
518 construct (as described in B), a representative well from the retrotransposition assay, and its
519 retrotransposition efficiency relative to smL1. The histogram displays the mean of three biological
520 replicate assays, each comprising three technical replicates per construct. Triangle, square, and
521 diamond shapes indicate the individual values for each biological replicate, and the error bars
522 represent standard deviation among biological replicates.

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

79 sequences, on mouse L1 retrotransposition efficiency. We determine that the presence and position
80 of a conserved poly purine tract in the mouse L1 3'UTR profoundly impacts the retrotransposition
81 efficiency of engineered mouse L1 reporter constructs.

82 **Results:** We capillary sequenced the pCEP4-based smL1 and pTN201/L1spa plasmids (Supplemental
83 Table 1), and observed that the synthetic sequences of the smL1 ORFs largely comprised synonymous
84 nucleotide substitutions relative to L1spa. However, smL1 ORF1 contained two nonsynonymous
85 substitutions (G53D and H159D) and bore an engineered Kozak consensus at its initiation codon that
86 was absent from pTN201/L1spa. Furthermore, the terminal 159 bp of 3'UTR sequence present in
87 pTN201/L1spa construct was absent from smL1 (Figure 1B; Supplemental Figure 1). We refer to this
88 159 bp sequence as the guanine-rich region (GRR), as it encompasses the conserved poly-purine tract.
89 We traced this deletion to the inadvertent loss of a NdeI restriction fragment during subcloning of the
90 smL1 construct (Supplemental Figure 2). The remaining 3'UTR sequence of smL1 contained 19
91 nucleotide substitutions and two single nucleotide deletions relative to the L1spa 3'UTR, which arose
92 during synthesis of the smL1 construct (Supplemental Figure 1).

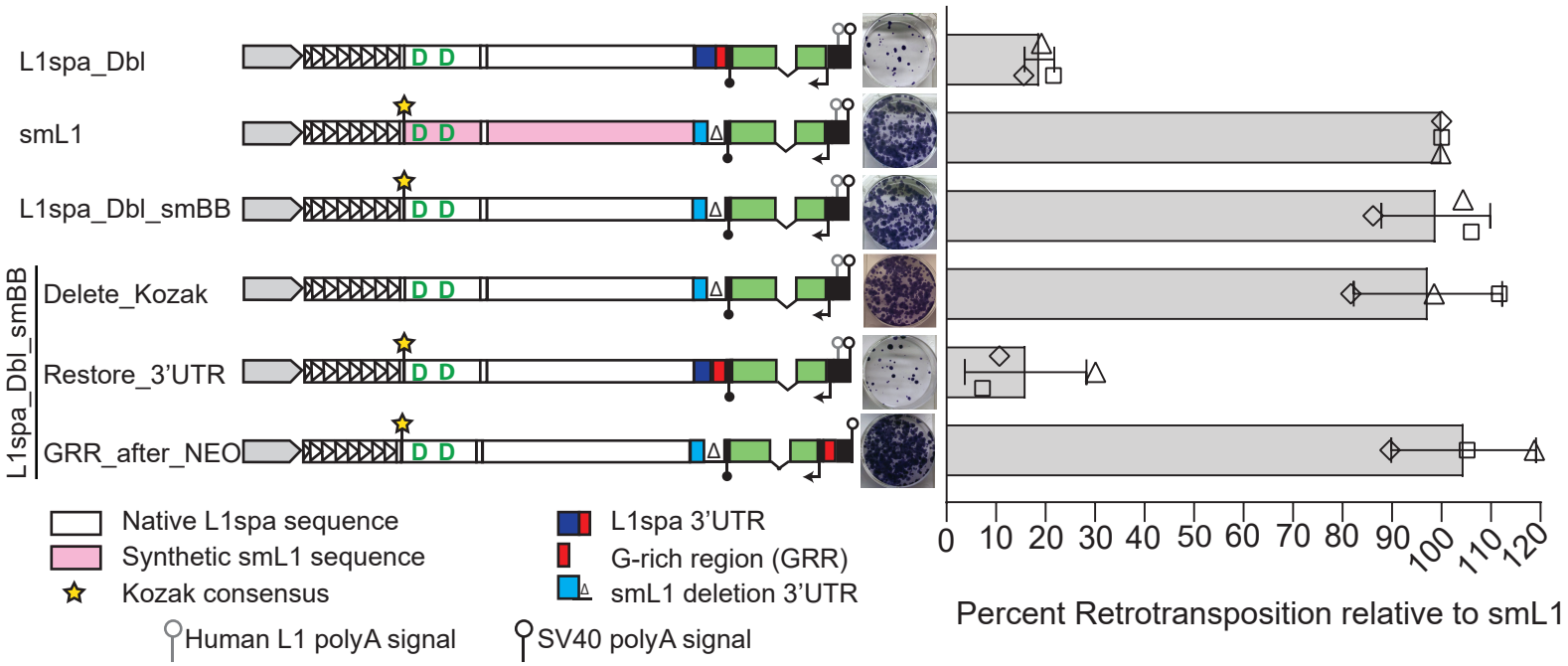
93 The L1_T_F subfamily consensus, and previously described active mouse L1_T_F elements,
94 contain aspartic acid at ORF1p positions 53 and 159 (5, 6, 29, 43). Indeed, the substitution H159D
95 improves L1spa ORF1p nucleic acid chaperone activity and retrotransposition efficiency (44). To
96 quantify the contribution of G53D and H159D to the increased retrotransposition efficiency of smL1
97 over L1spa, we generated L1spa_G53D, L1spa_H159D, and a construct with both changes, L1spa_Dbl
98 (Figure 1C). In a HeLa cell transient transfection-based retrotransposition assay (34), L1spa mobilized
99 at ~6% the efficiency of smL1, and, consistent with previous results(44), the substitution G53D had
100 little effect on retrotransposition. However, L1spa_H159D and L1spa_Dbl mobilized at ~19% the
101 efficiency of smL1. The reciprocal changes (D53G and D159H) reduced smL1 retrotransposition
102 efficiency by ~40% (smL1_Dbl; Figure 1C). Thus, the H159D amino acid substitution partially explains
103 the increased retrotransposition efficiency of smL1 over L1spa.

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

114 We speculated that the GRR might diminish observed retrotransposition by preventing
115 transcription through the NEO cassette. Indeed, G-quadruplex forming sequences have been
116 demonstrated to impede RNA polymerase procession (45-48) and implicated in the use of alternative
117 polyadenylation sites (49). To approach these possibilities, we first asked whether placement of the
118 GRR relative to the NEO cassette impacts retrotransposition efficiency. We replaced the L1spa GRR
119 into L1spa_Dbl_SMBB, inserting it downstream of NEO but upstream of the pCEP4 vector SV40
120 polyadenylation signal, to generate L1spa_Dbl_smBB_GRRAN. Unexpectedly, this construct
121 retrotransposed with equal efficiency to smL1 and L1spa_Dbl_smBB (Figure 2). We therefore
122 investigated whether formation of a G-quadruplex secondary structure within the GRR diminished
123 retrotransposition via a transcriptional block when placed upstream of the NEO cassette. To infer the
124 structure of L1spa_Dbl_smBB_GRRAN mRNA molecules from which successful retrotransposition
125 events arose, we characterized 10 L1spa_Dbl_smBB_GRRAN integrants by inverse PCR. Like smL1
126 insertions previously characterized in HeLa cells (1), all 10 L1spa_Dbl_smBB_GRRAN insertions bore
127 the hallmarks of bona fide retrotransposition events and incorporated L1 sequence, the spliced NEO
128 cassette, and the GRR. Importantly, all insertions terminated in poly-A tracts directed by the SV40
129 poly-A signal (Figure 3A, Supplemental Figure 3, Supplemental Table 1). We concluded that these

Figure 2

A.


L1spa_Dbl
smL1
L1spa_Dbl_smBB
Delete_Kozak
Restore_3'UTR
GRR_after_NEO
L1spa_Dbl_smBB
D D
Native L1spa sequence
Synthetic smL1 sequence
Kozak consensus
Human L1 polyA signal
L1spa 3'UTR
G-rich region (GRR)
smL1 deletion 3'UTR
SV40 polyA signal
0 10 20 30 40 50 60 70 80 90 100 110 120
Percent Retrotransposition relative to smL1

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

539 A. From left to right: a schematic of each construct, a representative well from the retrotransposition
540 assay, and its retrotransposition efficiency relative to smL1. L1spa_Dbl_smBB contains the ORF1 and
541 ORF2 sequences of L1spa_Double with a Kozak consensus at the ORF1p initiation codon (as in smL1),
542 and the deletion 3'UTR of the smL1 construct. Restore_3'UTR is equivalent to L1spa_Dbl_smBB but
543 with the full L1spa 3'UTR restored upstream of the NEO cassette. GRR_after_NEO is equivalent to
544 L1spa_Dbl_smBB but with the G-rich region of the L1spa 3'UTR placed downstream of the NEO
545 cassette, and the human L1 polyadenylation signal removed. The histogram displays the mean of three
546 biological replicate assays, each comprising three technical replicates per construct. Triangle, square,
547 and diamond shapes indicate the individual values for each biological replicate, and the error bars
548 represent standard deviation among biological replicates.

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

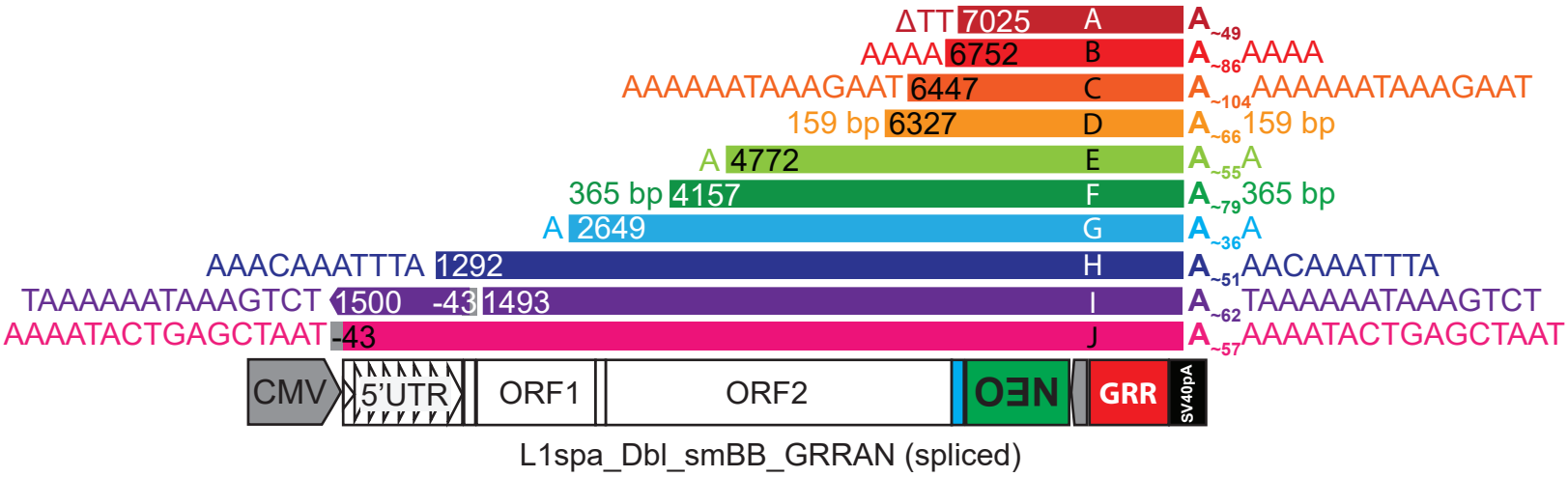
571

130 insertions occurred following complete transcription of the L1spa_Dbl_smBB_GRRAN construct,
131 including RNA polymerase procession through the GRR and utilization of the SV40 poly-A signal,
132 arguing against hinderance of RNA polymerase by the GRR.

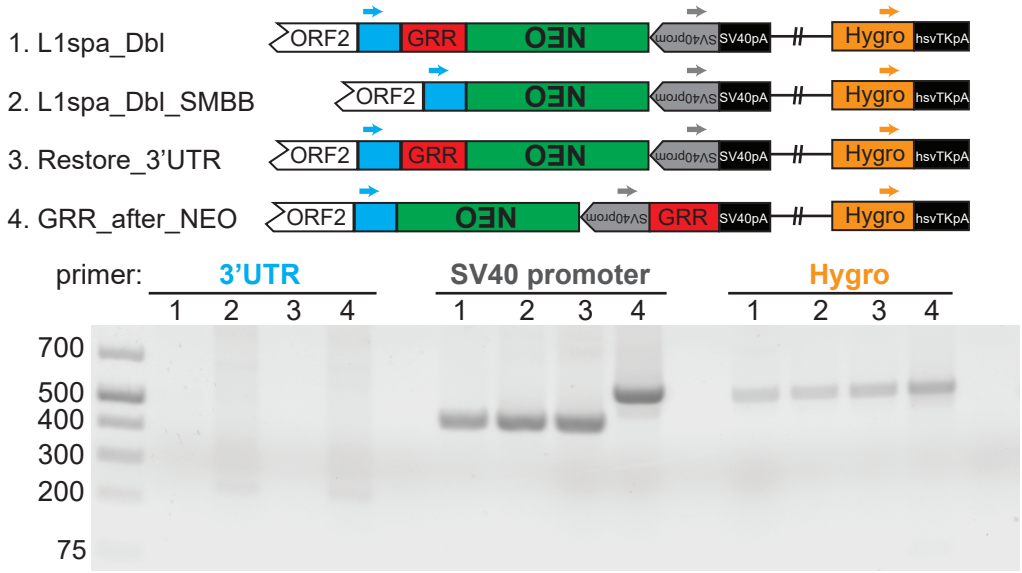
133 We next asked whether the GRR can promote premature polyadenylation. In the L1spa_Dbl
134 construct, where the GRR is upstream of the NEO cassette, polyadenylation in the vicinity of the GRR
135 would result in an L1 mRNA molecule lacking the NEO cassette, the retrotransposition of which would
136 be undetectable in a selection-based assay. To determine whether positioning the GRR upstream of
137 the NEO cassette results in premature polyadenylation, we performed 3'RACE assays on RNA
138 extracted from cells transfected with L1spa_Dbl constructs with the GRR upstream of the NEO cassette
139 (L1spa_Dbl and Restore_3'UTR), downstream of the NEO cassette (L1spa_Dbl_smBB_GRRAN), or
140 deleted entirely (L1spa_Dbl_SMBB). We used L1-specific primers located in the SV40 promoter of the
141 antisense NEO cassette to detect mRNAs correctly polyadenylated at the SV40 polyadenylation signal,
142 and in the L1spa 3'UTR to detect mRNAs potentially prematurely polyadenylated proximal to the GRR.
143 We also used a primer annealing to the 3' end of the hygromycin resistance gene in the pCEP4
144 backbone as an internal control (Figure 3B). For all constructs, we detected the expected ~520bp band
145 corresponding to polyadenylated transcripts from the hygromycin resistance gene. Using a primer to
146 the antisense SV40 promoter responsible for NEO expression, we detected the expected ~350 bp band
147 corresponding to L1 reporter transcripts polyadenylating at the SV40 polyadenylation signal for
148 L1spa_Dbl, L1spa_Dbl_SMBB, and Restore_3'UTR, and a ~520 bp band from
149 L1spa_Dbl_SMBB_GRR_after_NEO, the increased size of which is accounted for by the inclusion of the
150 GRR sequence between the antisense SV40 promoter and the SV40 polyadenylation signal. Using a
151 primer annealing to the L1spa 3'UTR, we did not detect bands in the range of 200-400 bp indicative of
152 polyadenylation within or proximal to the GRR in either the construct L1spa_Dbl or Restore_3'UTR.
153 Thus, we conclude that premature polyadenylation of sense engineered L1 transcripts is unlikely to be
154 responsible for the observed reduction in retrotransposition efficiency when the GRR is situated
155 upstream of the NEO cassette.

Figure 3

A.


ΔTT 7025 A
A~49
AAAA 6752 B
A~86 AAAA
AAAAAATAAAGAAT 6447 C
A~104 AAAAAATAAAGAAT
159 bp 6327 D
A~66 159 bp
A 4772 E
A~55 A
365 bp 4157 F
A~79 365 bp
A 2649 G
A~36 A
AAACAAATTTA 1292 H
A~51 AACAAATTTA
TAAAAAATAAAGTCT 1500 -43 1493 I
A~62 TAAAAAATAAAGTCT
AAAATACTGAGCTAAT -43 J
A~57 AAAATACTGAGCTAAT
CMV
5'UTR
ORF1
ORF2
NEO
GRR
SV40pA
L1spa_Dbl_smBB_GRRAN (spliced)

B.


1. L1spa_Dbl
2. L1spa_Dbl_SMBB
3. Restore_3'UTR
4. GRR_after_NEO
ORF2
GRR
NEO
SV40prom
SV40pA
Hygro
hsvTKpA
primer:
3'UTR
SV40 promoter
Hygro
1 2 3 4
700
500
400
300
200
75

572 **Figure 3: The GRR does not promote premature polyadenylation**

573 A. Structural hallmarks of ten insertions generated with the GRR_after_NEO construct are shown,
574 including the 5' TSD, 5' truncation position relative to the L1spaDbl_smBB_GRRAN, polyadenylation
575 position, poly-A tail length and 3'TSD. Below, the structure of GRR after NEO, with the spliced NEO
576 cassette in green, the GRR in red, and the SV40 polyadenylation signal in black. Schematic is not to
577 scale. Delta symbol indicates a target-site deletion. Insertion I contained a 5' inversion for which the
578 break-points are indicated. Insertions I and J arose from a transcript initiating upstream of the L1
579 5'UTR, at the transcription start site of the CMV promoter within the pCEP4 vector(67) denoted as
580 position -43. For two insertions bearing >100 bp TSDs, the TSD length, rather than TSD sequence, is
581 shown.

582 B. 3'RACE assay to identify sites of polyadenylation. Above, schematics representing the constructs 1)
583 L1spa_Dbl, 2) L1spa_Dbl_SMBB, 3) Restore_3'UTR and 4) GRR_after_NEO. The positions of gene-
584 specific primers annealing to the L1spa 3'UTR (blue), the antisense SV40 promoter driving NEO
585 expression (grey), and the hygromycin resistance gene in the pCEP4 vector backbone (orange) are
586 denoted by arrows. Schematics are not drawn to scale. Below, agarose gel showing 3'RACE products
587 detected using RNA isolated from cells transfected with the four above constructs, using each of the
588 primers shown.

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

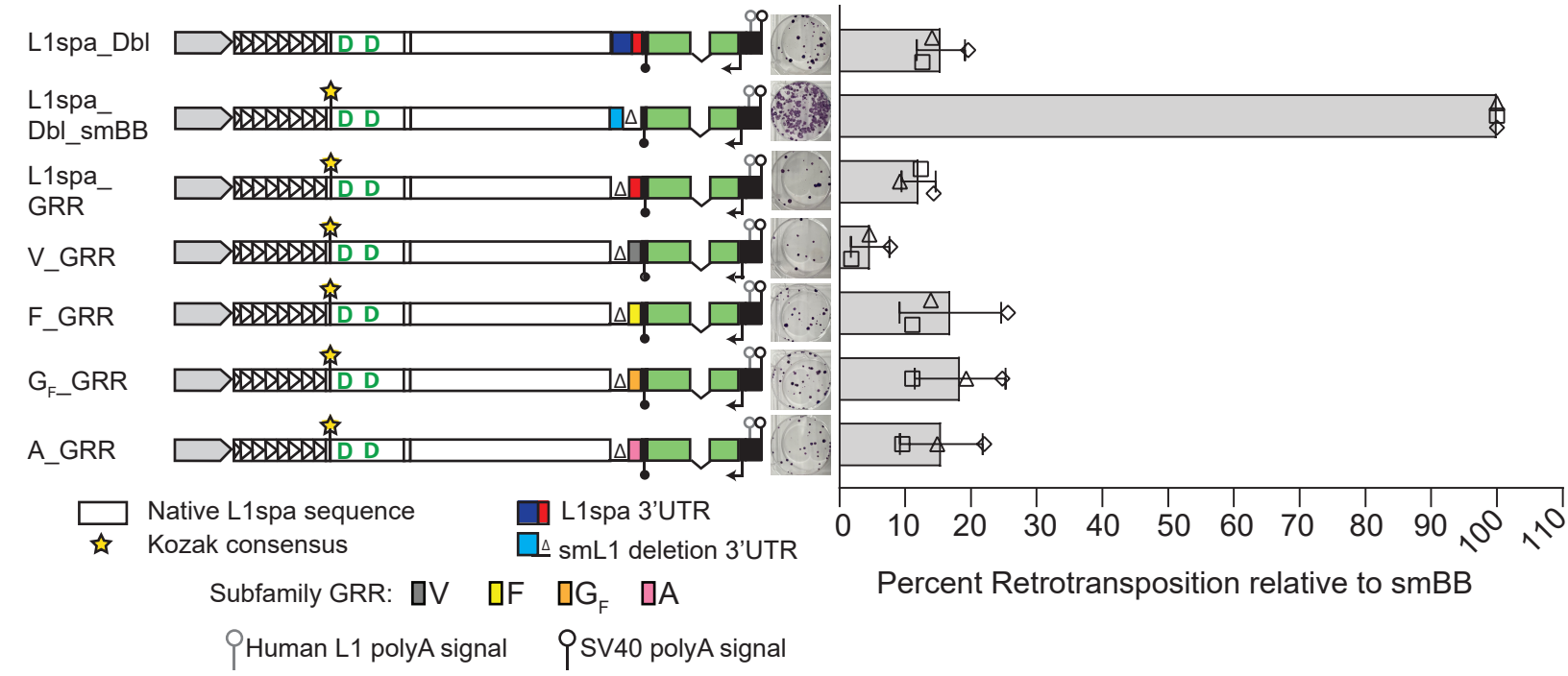
607

156 Finally, we investigated the sequence characteristics of the L1spa GRR responsible for
157 diminished engineered retrotransposition efficiency when placed upstream of the NEO cassette. An
158 L1spa_Dbl construct containing the L1spa GRR and lacking the remaining 3'UTR sequence
159 retrotransposed comparably to L1spa_Dbl, confirming that the L1spa GRR alone is sufficient to impede
160 engineered retrotransposition (L1spa_GRR, Figure 4A). To determine whether this effect is common
161 to the GRR sequences from ancient (V, F) and presently active (G_F, A) L1s in addition to the T_F subfamily
162 (represented by L1spa), we replaced the L1spa GRR with the consensus GRR sequence from each of
163 these subfamilies, which contain 9 (G_F, A) or 10 (V, F) runs of 3 or more guanine residues (6). In all
164 cases, retrotransposition was reduced to <20% of L1spa_Dbl_smBB (Figure 4A). To examine in greater
165 detail the role of the L1spa GRR sequence in retrotransposition inhibition, we generated a series of
166 deletion constructs. The L1spa GRR contains 9 runs of 3 or more guanine residues, separated by
167 variable stretches of other bases (Figure 4B). Relative to L1spa_Dbl_SMBB (no GRR sequence), the
168 sequential restoration of guanine runs in a 5'-to-3' direction resulted in a diminution of
169 retrotransposition efficiency, with the largest difference found between the constructs containing 5
170 and 6 guanine runs (Figure 4C). Thus, the position-dependent inhibitory effect of the GRR on
171 engineered L1 retrotransposition is shared by both ancient and presently active mouse L1s, and the
172 magnitude of inhibition is positively correlated with the length of the GRR.

173 **Discussion:** The highly active smL1 construct and its derivatives have facilitated numerous insights into
174 the mechanism, regulation, and consequences of L1 retrotransposition in vivo and in vitro (50-59).
175 Here we demonstrate that synthetic reduction in mouse L1 ORF adenosine content has little impact
176 on retrotransposition efficiency, at least in HeLa cells. We find instead that the elevated activity of
177 smL1 over pTN201/L1spa arises partly from the nonsynonymous ORF1 substitution H159D (44), and
178 to a larger extent from the deletion of the L1spa GRR from the smL1 3'UTR. Our results resolve a long-
179 standing incongruity regarding the impact of synthetic ORF sequences on L1 retrotransposition
180 efficiency, as the similarly re-coded synthetic human L1 ORFeus_Hs retrotransposes no more than 3-
181 fold more efficiently than the native parental element $L1_{RP}$ (42, 60). Thus, while both human and

Figure 4

A.



B. L1spa GRR nucleotide sequence

5′-TCAAAAGATGGCCTAGTCGGCCATCACTGGAAAGAGAGGCCCATTGGACACGCAGA

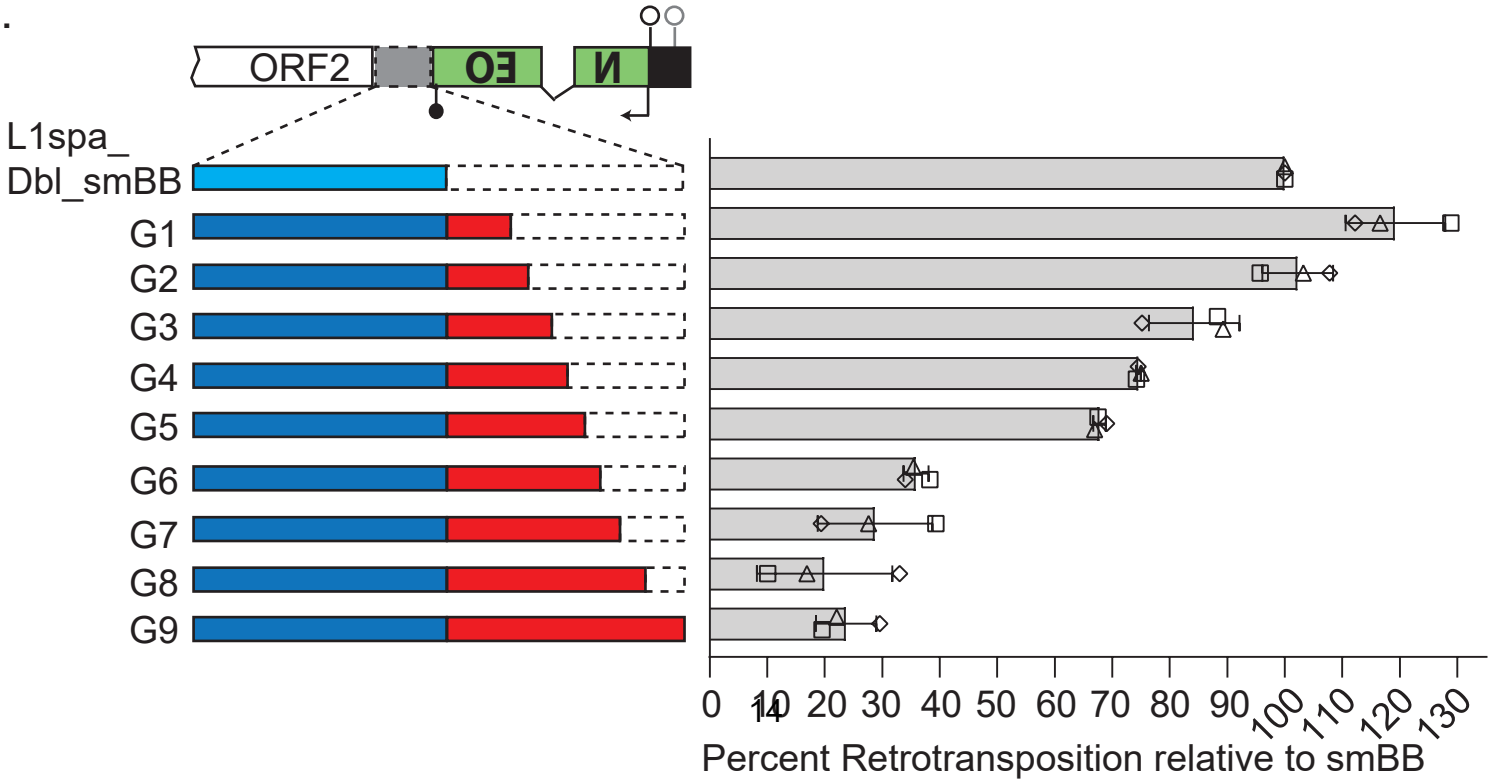
G1 G2 G3 G4 G5

CTTTGTGTGCCCCGGTACAGGGGAACGCCAGGGCCAAAGGGGGGGGAGTGGGTGGG

G6 G7 G8 G9

TAGGGGAGTGGGGGTAGGTGGGTAAGGGGGACTTTTGGTATAGCATT-3′

C.



608 **Figure 4: Variations in G-rich region sequence content differentially impact retrotransposition.**

609 A. From left to right: a schematic of each construct, a representative well from the retrotransposition
610 assay, and its retrotransposition efficiency relative to smL1. L1spa_GRR contains the GRR portion only
611 of the L1spa 3'UTR. V_GRR, F_GRR, GF_GRR and A_GRR contain the GRR portion of the 3'UTR from
612 the L1Md_V_I, L1Md_F_I, L1Md_G_{FI} and L1Md_A_I consensus sequences. The histogram displays the
613 mean of three biological replicate assays, each comprising three technical replicates per construct.
614 Triangle, square, and diamond shapes indicate the individual values for each biological replicate, and
615 the error bars represent standard deviation among biological replicates.

616 B. Nucleotide sequence of the 159bp region comprising the G-rich region (GRR) of the L1spa 3'UTR.
617 Runs of 3 or more consecutive G nucleotides are denoted as G1 through G9 and highlighted in red.

618 C. From left to right: a schematic of each construct, a representative well from the retrotransposition
619 assay, and its retrotransposition efficiency relative to L1spa_Dbl_SMBB. Runs of guanine residues have
620 been progressively restored to the GRR of constructs G1 through G8, with the full GRR present in
621 construct G9. The histogram displays the mean of three biological replicate assays, each comprising
622 three technical replicates per construct. Triangle, square, and diamond shapes indicate the individual
623 values for each biological replicate, and the error bars represent standard deviation among biological
624 replicates.

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

182 mouse synthetic L1 ORFs yield massively increased mRNA levels relative to native ORF sequences, this
183 elevated expression does not effect a proportionate increase in retrotransposition (1, 42). This result
184 is consistent with the modest increase in retrotransposition achieved by augmenting L1 expression
185 with a strong heterologous promoter (33), and suggests that L1 mRNA levels are not the rate-limiting
186 determinant of retrotransposition efficiency.

187 The molecular mechanism by which deleting or repositioning the L1spa GRR impacts
188 retrotransposition efficiency remains unclear, and does not appear to be caused by direct impediment
189 of RNA polymerase processing or by premature polyadenylation. Importantly, the impaired
190 retrotransposition efficiency observed when the L1spa GRR is placed upstream of the NEO cassette is
191 detected using GRR sequences from active and ancient mouse L1 elements, but based on previous
192 reports does not extend to human L1 reporter constructs. In contrast, deletions encompassing the
193 human poly-purine tract from NEO- and EGFP-based L1 constructs, in which this motif is also
194 positioned upstream of the reporter cassette, reduced retrotransposition by ~10% and ~30%,
195 respectively (33, 61). Furthermore, experimental evidence suggests that the human poly-purine tract
196 can form a G-quadruplex (G4) secondary structure (19, 62) and that stabilization of this structure very
197 modestly stimulates engineered retrotransposition (61). The decreased retrotransposition efficiency
198 of constructs incorporating the L1spa GRR upstream of the NEO cassette may represent an artifact
199 specific to the L1spa GRR sequence composition and its ability to form secondary structures in the
200 context of the reporter construct (19), rather than evidence that this motif intrinsically inhibits
201 retrotransposition. Indeed, the possibility that disruption of the L1 3'UTR by an indicator cassette
202 could impact retrotransposition has long been acknowledged (63) and, to speculate, the conservation
203 of a self-attenuating motif is difficult to reconcile with the evolutionary imperative of L1 as a selfish
204 genetic element (19, 20).

205 We observed no difference in retrotransposition efficiency when the L1spa GRR was deleted
206 entirely, compared to when it is repositioned downstream of the NEO cassette. This result suggests

207 that a potential function for the GRR in endogenous mouse L1 retrotransposition is likely to be subtle,
208 perhaps influenced by the stabilization of RNA or DNA G4 secondary structures and dependent on the
209 cellular milieu. The extent to which the deletion of the GRR from the smL1 construct and its derivatives
210 impacts the conclusions drawn from previous in vitro and in vivo studies using these reporters remains
211 to be evaluated (50-59); at least some transgenic mouse models generated using smL1 based
212 constructs lack the GRR (50, 53, 57, 64). Engineered mouse L1 constructs reconfigured with the GRR
213 downstream of the reporter cassette represent useful tools for investigating the function of this
214 conserved motif, and for studies of mouse L1 biology in general, as high efficiency retrotransposition
215 and structurally normal L1 integrants can be achieved using a construct containing the entire native
216 mouse L1 sequence.

217

218 **Materials and Methods:**

219 Constructs used in this study:

220 pTN201/L1spa: Has been described previously (5). It consists of the pCEP4 backbone with the element
221 L1spa truncated at position 7418 in the 3'UTR (removing the native mouse L1 polyadenylation signal
222 but leaving the G-rich region (GRR) of the 3'UTR in place). The L1spa 3'UTR is followed by the mneoI
223 retrotransposition indicator cassette (33, 36). The mneoI cassette is followed by the human L1
224 polyadenylation signal which is directly upstream of the SV40 polyadenylation signal in the pCEP4
225 backbone.

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

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

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

230 smL1: Has been described previously (1). It consists of the pCEP4 backbone with a synthetic mouse L1
231 element based on L1spa. The initiation codon of ORF1 has been placed into an engineered Kozak
232 consensus. The amino acid sequence of smL1 is identical to that of L1spa except for two residues in
233 ORF1 (position 53 and position 159). The 3'UTR of smL1 contains several single nucleotide
234 substitutions relative to L1spa, as well as two single-nucleotide deletions, and is truncated at position
235 7259 of L1spa. This truncation deletes the GRR of the L1spa 3'UTR. The smL1 3'UTR is followed by the
236 mneoI retrotransposition indicator cassette. The mneoI cassette is followed by the human L1
237 polyadenylation signal which is directly upstream of the SV40 poly-A signal in the pCEP4 backbone.

238 smL1_mut2: Has been described previously (1). It is identical to smL1 but with inactivating mutations
239 in the endonuclease and reverse transcriptase active sites of ORF2.

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

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

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

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

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

251 L1spa_Dbl_SMBB_L1spaGRR: Identical to L1spa_Dbl_SMBB, but the GRR sequence from the L1spa
252 3'UTR placed between the end of ORF2 and the NEO cassette, using an existing AgeI site and
253 introducing an engineered SbfI site.

254 L1spa_Dbl_SMBB_V_GRR, L1spa_Dbl_SMBB_F_GRR, L1spa_Dbl_SMBB_G_F_GRR, and
255 L1spa_Dbl_SMBB_A_GRR: Identical to L1spa_Dbl_SMBB_L1spaGRR, but with the GRR sequence from
256 the L1Md_VI, L1Md_FI, L1Md_GFI and L1Md_AI consensus sequences (6) placed between the end of
257 ORF2 and the NEO cassette, using AgeI and SbfI sites described above.

258 L1spa_Dbl_SMBB_G1 through L1spa_Dbl_SMBB_G9: Identical to L1spa_Dbl_SMBB, but with
259 increasing portions of the GRR sequence from the L1spa 3'UTR, each restoring an additional run of 3
260 or more guanine residues, introduced between the end of ORF2 and the NEO cassette, using the AgeI
261 and SbfI sites as described for L1spa_Dbl_SMBB_L1spaGRR.

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

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

268 Cell culture: HeLa-JVM cells (33) were maintained in Dulbecco's Modified Eagle's Medium (DMEM)
269 high glucose supplemented with 1% L-glutamine, 1% pen-strep and 10% fetal bovine serum. Cells
270 were passaged every 3-5 days at 70-80% confluence using 0.25% Trypsin-EDTA (Life Technologies).
271 Cells were grown at 37°C in 5% CO_2.

272 Retrotransposition assays: For transient retrotransposition assays performed in 6-well plates as
273 previously described(34), $5x10^3$ HeLa-JVM cells were plated per well. 24 hours later, each well was
274 transfected with 1 µg of plasmid DNA using 4 ul FuGene and 96 uL opti-MEM. Each construct was

275 transfected in triplicate. The media was replaced 24 hours post-transfection. At 72 hours post-
276 transfection, selection was initiated with 400 µg/ml G418 (Geneticin, Life Technologies). G418 media
277 was replaced every other day. At 12 days post-transfection, cells were fixed with 2%
278 formaldehyde/0.4% glutaraldehyde solution and stained with 0.1% crystal violet solution.

279 Assays to measure transfection efficiency for each L1 reporter plasmid were carried out as follows:
280 $4x10^4$ HeLa cells were plated per well of a 6-well dish. 24 hours later, each well was transfected with
281 1 µg total plasmid DNA, consisting of 0.5 µg L1 reporter plasmid and 0.5 µg pCEP4-GFP, using 4 ul of
282 FuGene and 96 ul opti-MEM. Each L1 reporter+GFP plasmid was transfected in duplicate. The media
283 was replaced 24 hours post-transfection. At 72 hours post-transfection, cells were trypsinized,
284 resuspended in PBS (Life Technologies) and assessed for GFP expression by flow cytometry using a
285 Cytoflex flow cytometer (Translational Research Institute flow cytometry core). The average percent
286 GFP-positive cells was determined for each L1 reporter construct, and used to normalize the results
287 of the retrotransposition assay.

288 Clonal cell lines for inverse PCR characterization of L1spa_Dbl_smBB_GRRAN insertions were
289 generated by plating $5x10^3$ HeLa cells per 10 cm dish, and 18 hours later transfecting with 1 ug plasmid
290 DNA using 4 ul FuGene and 96 ul opti-MEM. Media was replaced 24 hours after transfection. Selection
291 with 400 ug/ml Geneticin was initiated at 72 hours post-transfection and continued for 12 days to
292 generate isolated colonies harboring retrotransposition events. Individual colonies were manually
293 picked and expanded to generate clonal cell lines. Approximately $5x10^6$ cells were harvested per clonal
294 cell line and subjected to phenol-chloroform DNA extraction. DNA concentration was measured via
295 NanoDrop. Inverse PCR was carried out as described previously (65, 66). Briefly, 4 ug from each cell
296 line was digested with 25 units EcoRI (New England Biolabs) in 100 ul total volume. Reactions were
297 heat inactivated and digested DNA was ligated under dilute conditions to promote intramolecular
298 ligation (1 ml total volume) using 3200 units T4 DNA ligase (NEB) for 24 hours. Ligations were purified
299 by chloroform extraction and DNA was resuspended in 40 ul water. For each ligation, 4 ul of DNA was

300 used as a template in the first-round inverse PCR reaction using the Roche Expand long-template PCR
301 system. Each reaction consisted of 10 pmol primer NEO210as (5' GACCGCTTCCTCGTGCTTTACG 3), 10
302 pmol primer NEO1720s (5' TGCGCTGACAGCCGGAACACG 3'), 20 nM each dNTP, and 2.5 units of
303 enzyme. Cycling conditions were 95°C for 2 minutes, followed by 30 cycles of 94°C for 15 seconds,
304 64°C for 30s, and 68°C for 15 min, followed by a 30 min extension at 68°C. For the second-round
305 inverse PCR reaction, 4 ul of the product from the first-round inverse PCR was used directly as
306 template, with 10 pmol primer NEO173as (5' CATCGCCTTCTATCGCCTTCTTG 3') 10 pmol primer
307 NEO1808s (5' GCGTGCAATCCATCTTGTTCAATG 3'), 20 nM each dNTP, and 2.5 units of enzyme. Cycling
308 conditions were identical to first-round inverse PCR. Products were run on a 1% agarose gel and bands
309 were excised and purified using the Qiagen Min-Elute gel extraction kit. Bands <2kb were cloned using
310 the pGEM-T Easy kit (Promega) and bands >2kb were cloned using the Topo-XL2 PCR kit (Life
311 Technologies) following manufacturer's instructions. Cloned PCR products were sequenced using M13
312 forward and reverse primers Australian Genome Research Facility, Brisbane). Validation primers
313 (Integrated DNA Technologies) were designed based on the putative genomic location of each
314 insertion detected by inverse PCR, and insertions were validated as empty/filled and/or by 5' and 3'
315 junction PCRs using MyTaq HS DNA polymerase (Bioline). Validation products were run on 1% agarose
316 gels, purified using the Qiagen Min-Elute gel extraction kit, and capillary sequenced (Australian
317 Genome Research Facility, Brisbane) to verify the genomic location and structural hallmarks of each
318 insertion.

319 3'RACE: For each construct, $1x10^5$ HeLa-JVM cells were plated per well of a 6-well dish and transfected
320 ~18 hours later with 1ug of plasmid DNA using 3ul of FugeneHD and 97ul of Opti-MEM as described
321 above. Cells were selected for 7 days using 200ug/ml of Hygromycin B, and total RNA was extracted
322 using the Qiagen RNeasy Mini kit (cat no. 74104). 3'RACE was performed using the 3'RACE module of
323 the Invitrogen First Choice RLM-RACE kit (cat no. AM1700) according to manufacturer's instructions.
324 Briefly, 1ug of total RNA was used in a reverse transcription reaction with kit provided M-MLV reverse
325 transcriptase, RNase inhibitor, and 3'RACE adapter (5'-

326 GCGAGCACAGAATTAATACGACTCACTATAGGT12VN-3'). RNA was reverse transcribed at 42°C for 1
327 hour. PCR amplification of 3'RACE products was performed using gene-specific primers to the 5' region
328 of the L1spa 3'UTR (3'RACE_3'UTR-3 CGAAAGGACCCAGATGTAGC), the antisense SV40 early promoter
329 within the NEO retrotransposition indicator cassette (asSV40prom_F CGGGACTATGGTTGCTGACT),
330 and the 3' region of the hygromycin resistance gene in the pCEP4 plasmid backbone (Hygro_F3
331 ATTTCGGCTCCAACAATGTC) in combination with the adaptor specific primer from the RLM-RACE kit
332 (3'RACE Outer Primer 5'-GCGAGCACAGAATTAATACGACT-3'). PCR cycling conditions were as follows:
333 3 minutes at 94°C, 35 cycles of 94°C 30 seconds, 55°C 30 seconds, 72°C 30 seconds followed by 7
334 minutes at 72°C. PCR products were run on a 2% agarose gel, and bands were excised and purified
335 using the Qiagen MinElute gel extraction kit (cat no. 28604). Purified products were cloned using TA
336 Cloning Kit (Invitrogen, K204001.) and sequenced using primers T7 TAATACGACTCACTATAGGG and
337 Neosplice_F2 AGTGACAACGTCGAGCACAG by AGRF Brisbane.

338 **Acknowledgements:** HeLa-JVM cells were a gift from John V. Moran. The pTN201/L1spa construct was
339 a gift from Haig H. Kazazian. We thank members of the Richardson and Faulkner laboratories for
340 helpful discussion and critical reading of the manuscript. J.D.B. is a Founder and Director of CDI Labs,
341 Inc., a Founder of Neochromosome, Inc, a Founder and SAB member of ReOpen Diagnostics, and
342 serves or served on the Scientific Advisory Board of the following: Sangamo, Inc., Modern Meadow,
343 Inc., Sample6, Inc. and the Wyss Institute.

344 **References**

345 1. Han JS, Boeke JD. A highly active synthetic mammalian retrotransposon. Nature.
346 2004;429(6989):314-8.
347 2. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The influence
348 of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiology Spectrum. 2015;3(2).
349 3. Kazazian HH, Jr., Moran JV. Mobile DNA in Health and Disease. N Engl J Med. 2017;377(4):361-
350 70.
351 4. Goodier JL, Ostertag EM, Du K, Kazazian HH, Jr. A novel active L1 retrotransposon subfamily in
352 the mouse. Genome Res. 2001;11(10):1677-85.
353 5. Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, Seldin MF, et al. An actively
354 retrotransposing, novel subfamily of mouse L1 elements. EMBO J. 1998;17(2):590-7.

355 6. Sookdeo A, Hepp CM, McClure MA, Boissinot S. Revisiting the evolution of mouse LINE-1 in
356 the genomic era. Mobile DNA. 2013;4(1):3.
357 7. DeBerardinis RJ, Goodier JL, Ostertag EM, Kazazian HH, Jr. Rapid amplification of a
358 retrotransposon subfamily is evolving the mouse genome. Nat Genet. 1998;20(3):288-90.
359 8. Penzkofer T, Jager M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, et al. L1Base 2: more
360 retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 2017;45(D1):D68-
361 D73.
362 9. Zhou M, Smith AD. Subtype classification and functional annotation of L1Md retrotransposon
363 promoters. Mobile DNA. 2019;10:14.
364 10. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed
365 by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell.
366 1993;72(4):595-605.
367 11. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD. Human L1 retrotransposon encodes a conserved
368 endonuclease required for retrotransposition. Cell. 1996;87(5):905-16.
369 12. Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A. Reverse transcriptase encoded by
370 a human transposable element. Science. 1991;254(5039):1808-10.
371 13. Kolosha VO, Martin SL. In vitro properties of the first ORF protein from mouse LINE-1 support
372 its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A.
373 1997;94(19):10155-60.
374 14. Kolosha VO, Martin SL. High-affinity, non-sequence-specific RNA binding by the open reading
375 frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). J Biol Chem.
376 2003;278(10):8112-7.
377 15. Khazina E, Weichenrieder O. Non-LTR retrotransposons encode noncanonical RRM domains
378 in their first open reading frame. Proc Natl Acad Sci U S A. 2009;106(3):731-6.
379 16. Khazina E, Truffault V, Buttner R, Schmidt S, Coles M, Weichenrieder O. Trimeric structure and
380 flexibility of the L1ORF1 protein in human L1 retrotransposition. Nat Struct Mol Biol. 2011;18(9):1006-
381 14.
382 17. Baldwin ET, van Eeuwen T, Hoyos D, Zalevsky A, Tchesnokov EP, Sanchez R, et al. Structures,
383 functions and adaptations of the human LINE-1 ORF2 protein. Nature. 2024;626(7997):194-206.
384 18. Thawani A, Ariza AJF, Nogales E, Collins K. Template and target-site recognition by human
385 LINE-1 in retrotransposition. Nature. 2024;626(7997):186-93.
386 19. Howell R, Usdin K. The ability to form intrastrand tetraplexes is an evolutionarily conserved
387 feature of the 3' end of L1 retrotransposons. Mol Biol Evol. 1997;14(2):144-55.
388 20. Boissinot S, Sookdeo A. The Evolution of LINE-1 in Vertebrates. Genome Biol Evol.
389 2016;8(12):3485-507.
390 21. Goodier JL, Ostertag EM, Kazazian HH, Jr. Transduction of 3'-flanking sequences is common in
391 L1 retrotransposition. Hum Mol Genet. 2000;9(4):653-7.
392 22. Moran JV, DeBerardinis RJ, Kazazian HH, Jr. Exon shuffling by L1 retrotransposition. Science.
393 1999;283(5407):1530-4.
394 23. Pickeral OK, Makalowski W, Boguski MS, Boeke JD. Frequent human genomic DNA
395 transduction driven by LINE-1 retrotransposition. Genome Res. 2000;10(4):411-5.
396 24. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian
397 retroposons. Proc Natl Acad Sci U S A. 1997;94(5):1872-7.
398 25. Cost GJ, Boeke JD. Targeting of human retrotransposon integration is directed by the
399 specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry.
400 1998;37(51):18081-93.
401 26. Gilbert N, Lutz S, Morrish TA, Moran JV. Multiple fates of L1 retrotransposition intermediates
402 in cultured human cells. Mol Cell Biol. 2005;25(17):7780-95.
403 27. Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition.
404 Cell. 2002;110(3):315-25.

28. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, et al. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. Mol Cell. 2019;74(3):555-70 e7.
29. Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea GO, Munoz-Lopez M, et al. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res. 2017;27(8):1395-405.
30. Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, et al. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. 2019;29(10):1567-77.
31. Hancks DC, Kazazian HH, Jr. Roles for retrotransposon insertions in human disease. Mobile DNA. 2016;7:9.
32. Gagnier L, Belancio VP, Mager DL. Mouse germ line mutations due to retrotransposon insertions. Mobile DNA. 2019;10:15.
33. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. High frequency retrotransposition in cultured mammalian cells. Cell. 1996;87(5):917-27.
34. Wei W, Morrish TA, Alisch RS, Moran JV. A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem. 2000;284(2):435-8.
35. Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, et al. A mouse model of human L1 retrotransposition. Nat Genet. 2002;32(4):655-60.
36. Freeman JD, Goodchild NL, Mager DL. A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques. 1994;17(1):46, 8-9, 52.
37. Rangwala SH, Kazazian HH, Jr. The L1 retrotransposition assay: a retrospective and toolkit. Methods. 2009;49(3):219-26.
38. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004;429(6989):268-74.
39. Ilik IA, Glazar P, Tse K, Brandl B, Meierhofer D, Muller FJ, et al. Autonomous transposons tune their sequences to ensure somatic suppression. Nature. 2024;626(8001):1116-24.
40. Seczynska M, Bloor S, Cuesta SM, Lehner PJ. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. Nature. 2022;601(7893):440-5.
41. Kingsmore SF, Giros B, Suh D, Bieniarz M, Caron MG, Seldin MF. Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. Nat Genet. 1994;7(2):136-41.
42. An W, Dai L, Niewiadomska AM, Yetil A, O'Donnell KA, Han JS, et al. Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs. Mobile DNA. 2011;2(1):2.
43. Schauer SN, Carreira PE, Shukla R, Gerhardt DJ, Gerdes P, Sanchez-Luque FJ, et al. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Research. 2018;28(5).
44. Martin SL, Bushman D, Wang F, Li PW, Walker A, Cummiskey J, et al. A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. Nucleic Acids Res. 2008;36(18):5845-54.
45. Belotserkovskii BP, Liu R, Tornaletti S, Krasilnikova MM, Mirkin SM, Hanawalt PC. Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. Proc Natl Acad Sci U S A. 2010;107(29):12816-21.
46. Belotserkovskii BP, Neil AJ, Saleh SS, Shin JH, Mirkin SM, Hanawalt PC. Transcription blockage by homopurine DNA sequences: role of sequence composition and single-strand breaks. Nucleic Acids Res. 2013;41(3):1817-28.
47. Broxson C, Beckett J, Tornaletti S. Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. Biochemistry. 2011;50(19):4162-72.
48. Belotserkovskii BP, Soo Shin JH, Hanawalt PC. Strong transcription blockage mediated by R-loop formation within a G-rich homopurine-homopyrimidine sequence localized in the vicinity of the promoter. Nucleic Acids Res. 2017;45(11):6589-99.

455 49. Beaudoin JD, Perreault JP. Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both
456 alternative polyadenylation and mRNA shortening. Nucleic Acids Res. 2013;41(11):5898-911.
457 50. Newkirk SJ, Lee S, Grandi FC, Gaysinskaya V, Rosser JM, Vanden Berg N, et al. Intact piRNA
458 pathway prevents L1 mobilization in male meiosis. Proc Natl Acad Sci U S A. 2017;114(28):E5635-E44.
459 51. Kannan M, Li J, Fritz SE, Husarek KE, Sanford JC, Sullivan TL, et al. Dynamic silencing of somatic
460 L1 retrotransposon insertions reflects the developmental and cellular contexts of their genomic
461 integration. Mobile DNA. 2017;8:8.
462 52. Richardson SR, Narvaiza I, Planegger RA, Weitzman MD, Moran JV. APOBEC3A deaminates
463 transiently exposed single-strand DNA during LINE-1 retrotransposition. eLife. 2014;2014(3).
464 53. O'Donnell KA, An W, Schrum CT, Wheelan SJ, Boeke JD. Controlled insertional mutagenesis
465 using a LINE-1 (ORFeus) gene-trap mouse model. Proc Natl Acad Sci U S A. 2013;110(29):E2706-13.
466 54. Monot C, Kuciak M, Viollet S, Mir AA, Gabus C, Darlix JL, et al. The specificity and flexibility of
467 l1 reverse transcription priming at imperfect T-tracts. PLoS Genet. 2013;9(5):e1003499.
468 55. Xie Y, Mates L, Ivics Z, Izsvak Z, Martin SL, An W. Cell division promotes efficient
469 retrotransposition in a stable L1 reporter cell line. Mobile DNA. 2013;4(1):10.
470 56. Grandi FC, Rosser JM, An W. LINE-1-derived poly(A) microsatellites undergo rapid shortening
471 and create somatic and germline mosaicism in mice. Mol Biol Evol. 2013;30(3):503-12.
472 57. An W, Han JS, Schrum CM, Maitra A, Koentgen F, Boeke JD. Conditional activation of a single-
473 copy L1 transgene in mice by Cre. Genesis. 2008;46(7):373-83.
474 58. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. Unconventional translation of
475 mammalian LINE-1 retrotransposons. Genes Dev. 2006;20(2):210-24.
476 59. Wagstaff BJ, Barnerssoi M, Roy-Engel AM. Evolutionary conservation of the functional
477 modularity of primate and murine LINE-1 elements. PLoS One. 2011;6(5):e19672.
478 60. Kimberland ML, Divoky V, Prchal J, Schwahn U, Berger W, Kazazian HH, Jr. Full-length human
479 L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol
480 Genet. 1999;8(8):1557-60.
481 61. Sahakyan AB, Murat P, Mayer C, Balasubramanian S. G-quadruplex structures within the 3'
482 UTR of LINE-1 elements stimulate retrotransposition. Nat Struct Mol Biol. 2017;24(3):243-7.
483 62. Lexa M, Steflova P, Martinek T, Vorlickova M, Vyskot B, Kejnovsky E. Guanine quadruplexes
484 are formed by specific regions of human transposable elements. BMC Genomics. 2014;15:1032.
485 63. Moran JV. Human L1 retrotransposition: insights and peculiarities learned from a cultured cell
486 retrotransposition assay. Genetica. 1999;107(1-3):39-51.
487 64. An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, Ye P, et al. Active retrotransposition by a
488 synthetic L1 element in mice. Proc Natl Acad Sci U S A. 2006;103(49):18662-7.
489 65. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair
490 mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002;31(2):159-65.
491 66. Li J, Shen H, Himmel KL, Dupuy AJ, Largaespada DA, Nakamura T, et al. Leukaemia disease
492 genes: large-scale cloning and pathway predictions. Nat Genet. 1999;23(3):348-53.
493 67. Isomura H, Stinski MF, Kudoh A, Nakayama S, Murata T, Sato Y, et al. A cis element between
494 the TATA Box and the transcription start site of the major immediate-early promoter of human
495 cytomegalovirus determines efficiency of viral replication. J Virol. 2008;82(2):849-58.

496

497

498

499

500

643 **Supplemental Figure 1: Alignment of the 3'UTRs and flanking sequence of pTN201/L1spa and smL1.**
644 Alignment depicts the 3' end of ORF2 (L1spa ORF2, dark red; smL1 ORF2, bright red), followed by the
645 3'UTR (L1spa 3'UTR dark blue; smL1 3'UTR, light blue). The smL1 3'UTR deletion is highlighted in
646 yellow, and the runs of guanine bases in the L1spa 3'UTR are highlighted in red. The positions of
647 engineered restriction sites in each construct are shown in pink, and the NdeI sites resulting in the
648 GRR deletion are shown in green. The HSV Tk polyadenylation signal belonging to the NEO cassette
649 (purple) and the second exon of the NEO cassette (orange) are identical between pTN201/L1spa and
650 smL1.

651 **Supplemental Figure 2: Deletion of the GRR during cloning of the smL1 reporter construct.** A. The
652 pTN201/L1spa construct(5) and an intermediate pBluescript clone containing the smL1 sequence, with
653 features annotated as in Figure 1. The positions of known NdeI and BamHI sites are shown in black.
654 Our map of pTN201 did not include 15 base pairs of human L1 sequence present in pTN201
655 immediately following the L1spa 3'UTR, which was carried over from subcloning described in(5). This
656 15 bp of human L1 sequence contained an NdeI site (shown in red).

657 B) Deletion of the GRR-containing NdeI fragment during cloning. Left: we intended to move an
658 NdeI/BamHI fragment containing the GRR of the L1spa element and the NEO cassette downstream of
659 the smL1 ORFs and 5' portion of the 3'UTR in pBluescript. Right: due to the unanticipated NdeI site, a
660 180 bp NdeI fragment containing the terminal GRR-containing 159 bp of the L1spa 3'UTR and an
661 engineered PacI site (not shown) was deleted from the resulting construct. Owing to its small size, the
662 deletion was not detected when screening the constructs with diagnostic restriction digests.

663 **Supplemental Figure 3: Detailed structures of L1spa_Dbl_smBB_GRRAN integrants characterized via**
664 **inverse PCR.** The structure of each insertion is presented as follows: L1spa 5'UTR (triangles), ORFs
665 (white rectangles), first part of the smL1 3'UTR (light blue), spliced NEO cassette (green), sv40
666 promoter (grey), GRR (red) sv40 polyadenylation signal (purple). The 5' truncation point for each
667 insertion is shown. Insertions I and J arose from a transcript initiating upstream of the L1 5'UTR, at the
668 transcription start site of the CMV promoter within the pCEP4 vector(67) denoted as position -43.
669 Insertion J had a 5' inversion with breakpoint at position 1493 shown. Flanking genomic DNA is shown
670 to the right and left, with target-site duplications in red. The empty site is shown below, with TSD
671 sequence in red, the first-strand cleavage site indicated by a black arrow and the second-strand
672 cleavage site indicated by a grey arrow. Insertion A had a 2 bp target-site deletion, depicted in blue.
673 The chromosomal coordinates of the TSD sequence and the endonuclease cleavage motif for each
674 insertion is shown.

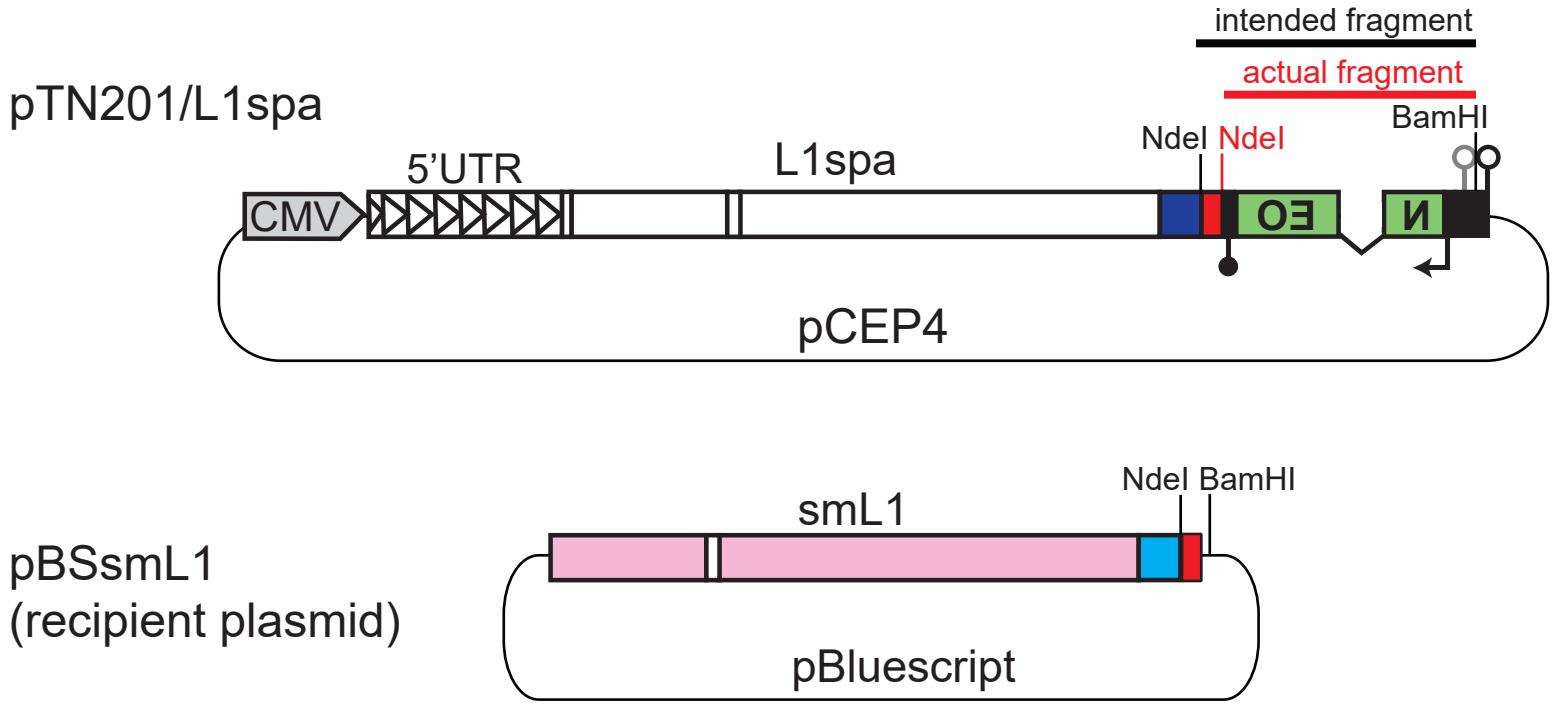
675 **Supplemental Table 1:** The first tab contains the sequences of L1spa (pTN201) and smL1 as
676 determined by capillary sequencing. The second tab contains the sequences (iPCR_L1_rev and
677 iPCR_L1_fwd) of inverse PCR products characterizing of L1spa_Dbl_smBB_GRRAN integrants and the
678 corresponding validation primers. The third tab contains retrotransposition assay data (average
679 colony counts from three technical replicate wells, normalized to transfection efficiency) and
680 mean/standard deviation values for three biological replicates used to generate the graphs shown in
681 Figure 1C, Figure 2, and Figure 4 A and C.

Supplemental Figure 1

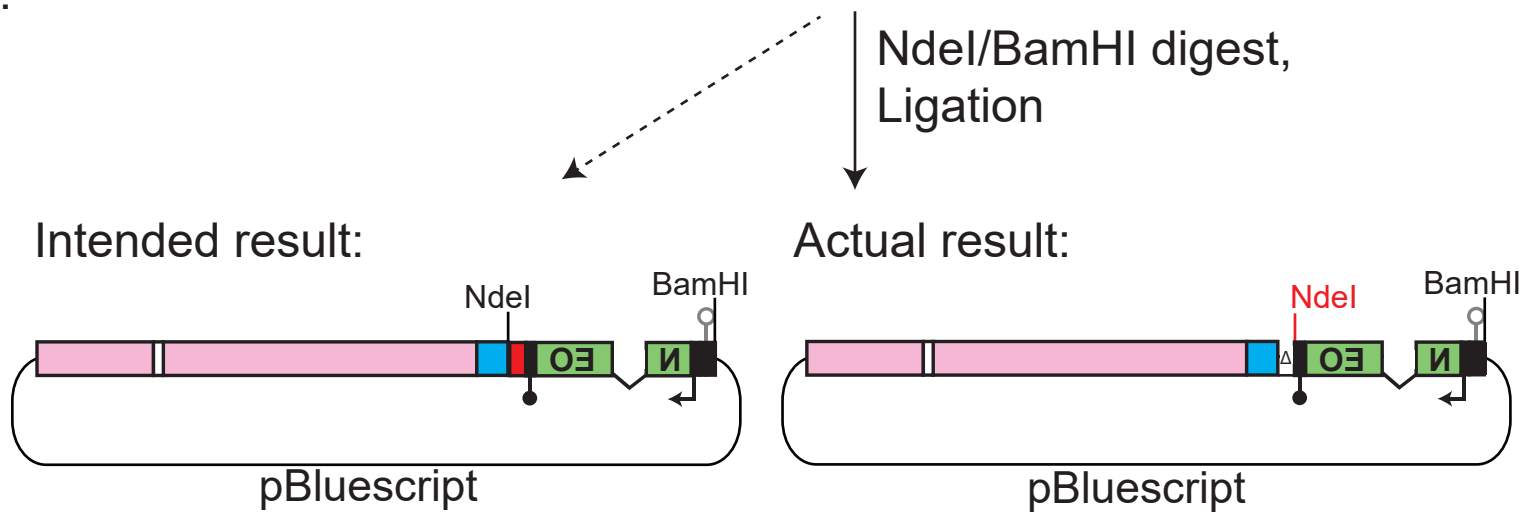
```
                L1Spa Orf2
L1spa   6742    TCCTGAGTGAGGTAACACAATCACAAAGGAACTCACACAATATGTACTCACTGATAAGTG   6801
                ||||||| ||||| || ||    ||  | |||   ||||| ||||||   ||||| || |
L1SM    6777    TCCTGAGCGAGGTGACCCAGAGCCAGCGCAACAGCCACAACATGTACAGCCTGATCAGCG   6836
                L1SM Orf2
                          SacI  L1spa 3'UTR
L1spa   6802    GATACTAGCGAGCTC-CAAAACCTAGGATACCCACGATATAAGATACAATTTCCTAAACA   6860
                | |||||||       |         |||||||||||||||||||||||||| ||||||||
L1SM    6837    GCTACTAGCCTAAAACCGGT------GATACCCACGATATAAGATACAATTGCCTAAACA   6890
                              AgeI        L1SM 3'UTR

L1spa   6861    CATGAAACTCAAGAAAAATGAAGACTGAAGTGTGGACACTATGCCCCTCCTTAGAAGTGG   6920
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
L1SM    6891    CATGAAACTCAAGAAAAATGAAGACTGAAGTGTGGACACTATGCCCCTCCTTAGAAGTGG   6950

L1spa   6921    GAACAAAACACCCATGGAAGGAGTTACAGAAACAAAGTTTGGAGCTGAGATGAAAGGAGG   6980
                |||||||||| |||| ||| |||||||||||||||||||||||||||||||||||||||
L1SM    6951    GAACAAAACAGCCATAGAAAGAGTTACAGAAACAAAGTTTGGAGCTGAGATGAAAGGAGA   7010

L1spa   6981    GACCATGTAGAGACTGCCATATCCAGGGATCCACCCCATAATCAGCATCCAAACGC--TG   7038
                ||||||||||| |||||||||||||||| |||||||||||||||||||||||||  ||  ||
L1SM    7011    GACCATGTAGAAACTGCCATATCCAGGGTTCCACCCCATAATCAGCATCCAA--GCTTTG   7068

L1spa   7039    ACACCATTGCATATACTAGCAAGATTTTATCGAAAGGACCCAGATGTAGCTGTCTCTTGT   7098
                |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||
L1SM    7069    ACACCATTGCATATACTAGGAAGATTTTATCGAAAGGACCCAGATGTAGCTGTCTCTTGT   7128

L1spa   7099    GAGACTATGCCGGGGCCTAGCAAACACAGAAGTGGATGCTCACAGTCAGCTAATGGATGG   7158
                |||||||||||||||||||||||| |||||||||||||||||||||||||||||| |||||||
L1SM    7129    GAGACTATGCCGGGGCCTAGCAA-CACAGAAGTGGATGCTCACAGTCAGCTA-TGGATGG   7186

L1spa   7159    ATCACAGGGCTCCCAATGGAGGAGCTAGAGAAAGTACCCAAGGAGCTAAAGGGATCTGCA   7218
                |||||||||||  |||||||||||||||||||||||||||||||||||||||||||||||||
L1SM    7187    ATCACAGGGCTTTCAATGGAGGAGCTAGAGAAAGTACCCAAGGAGCTAAAGGGATCTGCA   7246

L1spa   7219    ACCCTATAGGTGGAACAACATTATGAAC-TAACCAGTACCCCTGAGCTCTTGACTCTAGC   7277
                || |||||||||||||||||||||||   |||||||||||| |||||||||| | |||||
L1SM    7247    ACTCTATAGGTGGAACAACATTATGA-GTTAACCAGTACCCTTGAGCTCTTGTCGCTAGC   7305

L1spa   7278    TGCATATGTATCAAAAGATGGCCTAGTCGGCCATCACTGGAAAGAGAGGCCCATTGGACA   7337
                ||||||||||
L1SM    7306    TGCATATGTA-------------------------------------------------   7315

L1spa   7338    CGCAGACTTTGTGTGCCCCGGTACAGGGGAACGCCAGGGCCAAAGGGGGGGGAGTGGGTG   7397

L1SM    7316    ------------------------------------------------------------   7315

                                                                PacI
L1spa   7398    GGTAGGGGAGTGGGGGTAGGTGGGTAAGGGGGACTTTTGGTATAGCATTAATTAACTCGA   7457

L1SM    7316    ------------------------------------------------------------   7315

                                        HSV Tk pA (antisense)
L1spa   7458    TACATATGTAACTAACCCGATCCGAACAAACGACCCAACACCCGTGCGTTTTATTCTGTC   7517
                          |||||||||||||||||||||||||||||||||||||||||||||||||
L1SM    7316    ----------ACTAACCCGATCCGAACAAACGACCCAACACCCGTGCGTTTTATTCTGTC   7365

                                    NEO (antisense)
L1spa   7518    TTTTTATTGCCGATCCCCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTG   7577
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
L1SM    7366    TTTTTATTGCCGATCCCCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTG   7425
```

Supplemental Figure 2

A.

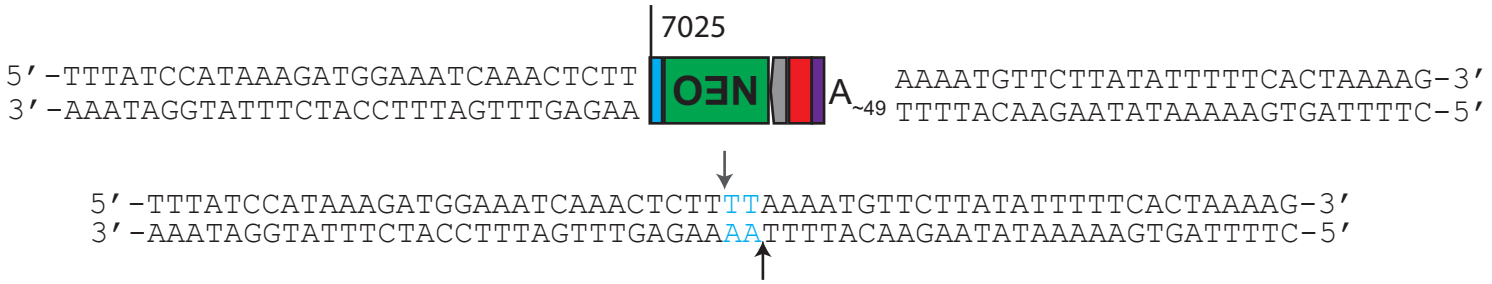


B.



Supplemental Figure 3

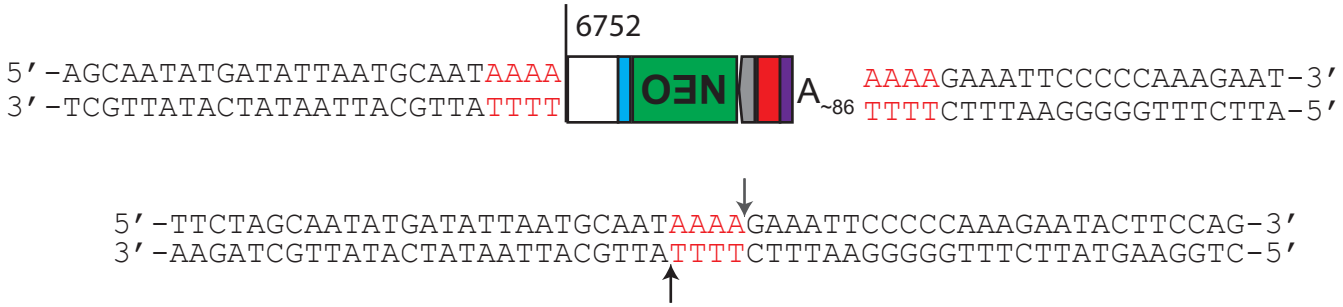
Insertion A



Target-site deletion: Chr3:130997869-130997870

EN cleavage: 5’-TTTT/AA-3’

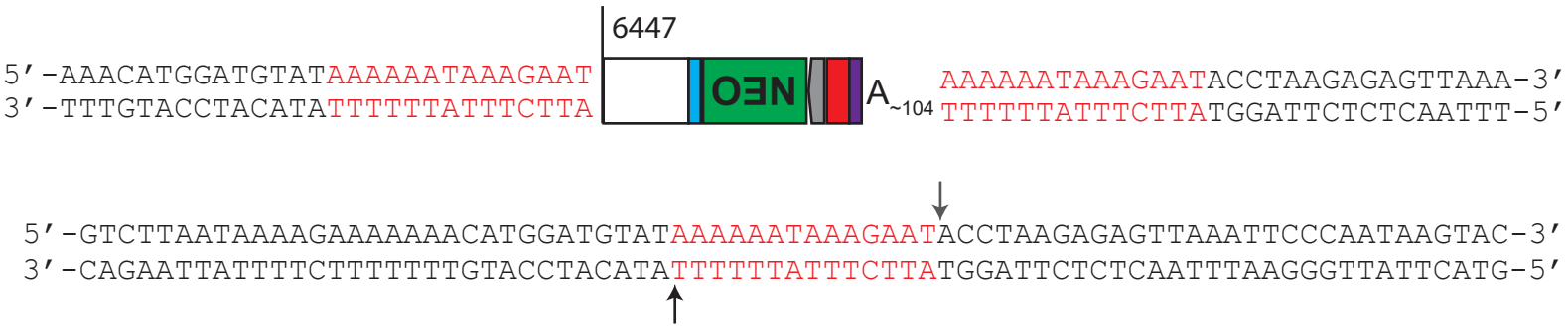
Insertion B



TSD: Chr3:148918423-148918426

EN cleavage: 5’-TTTT/AT-3’

Insertion C



TSD: Chr5: 21539952-21539965

EN cleavage: 5’-TTTT/AT-3’

Supplemental Figure 3

Insertion D

5′-ATAATTTTTGTGTCATGGTTTATGTTTTT
3′-TATTAAAAACACAGTACCAAATACAAAAA

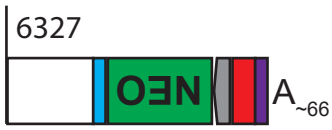
AAAAAAACTTACCTTTTAACATGCTAAAAAAG
TTTTTTTGAATGGAAAATTGTACGATTTTTTC

ACAGTACTTCTTGAATGGTAGAATAAAGACCA
TGTCATGAAGAACTTACCATCTTATTTCTGGT

TCAAAAATCTGCTCATCCATAAAAGCAATGAG
AGTTTTTAGACGAGTAGGTATTTTCGTTACTC

ATCAGTGGCAAAAATTGTCAAAATTAACTTTT
TAGTCACCGTTTTTAACAGTTTTAATTGAAAA

TCAGAACTCTGGAAATTAACTAAAAGATTGC
AGTCTTGAGACCTTTAATTGATTTTCTAACG



AAAAAAACTTACCTTTTAACATGCTAAAAAAG
TTTTTTTGAATGGAAAATTGTACGATTTTTTC

ACAGTACTTCTTGAATGGTAGAATAAAGACCA
TGTCATGAAGAACTTACCATCTTATTTCTGGT

TCAAAAATCTGCTCATCCATAAAAGCAATGAG
AGTTTTTAGACGAGTAGGTATTTTCGTTACTC

ATCAGTGGCAAAAATTGTCAAAATTAACTTTT
TAGTCACCGTTTTTAACAGTTTTAATTGAAAA

TCAGAACTCTGGAAATTAACTAAAAGATTGC
AGTCTTGAGACCTTTAATTGATTTTCTAACG

AAAATGTCAAGGACTGTTTAGTCAGGAAAA-3′
TTTTACAGTTCCTGACAAATCAGTCCTTTT-5′

TSD: Chr21: 29204205-29204047

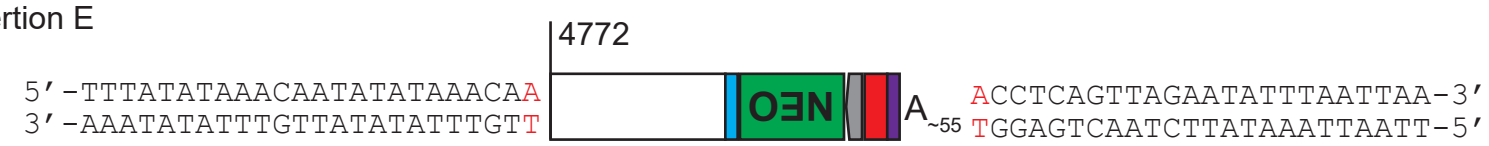
EN cleavage: 5’-TTTT/AA-3’

5′-TGTCATGGTTTATGTTTTTTAAAAAAACTTACCTTTTAACATGCTAAAAAAGACAGTACTTCTTGAATGGTAGAATAAAG
3′-ACAGTACCAAATACAAAAAATTTTTTTGAATGGAAAATTGTACGATTTTTTCTGTCATGAAGAACTTACCATCTTATTTC

ACCATCAAAAATCTGCTCATCCATAAAAGCAATGAGATCAGTGGCAAAAATTGTCAAAATTAACTTTTTCAGAACTCTGG
TGGTAGTTTTTAGACGAGTAGGTATTTTCGTTACTCTAGTCACCGTTTTTAACAGTTTTAATTGAAAAAGTCTTGAGACC

AAATTAACTAAAAGATTGCAAAATGTCAAGGACTGTTT-3′
TTTAATTGATTTTCTAACGTTTTACAGTTCCTGACAAA-5′

Insertion E



5′-TTTATATAAACAATATATAAACAACCTCAGTTAGAATATTTAATTAA-3′
3′-AAATATATTTGTTATATATTTGTTGGAGTCAATCTTATAAATTAATT-5′

TSD: chr7: 153197463

EN cleavage: 5’-AGGT/TG-3’

Supplemental Figure 3

Insertion F

```
5′-AATTCTTTGGAAAAAAGACAAAGTTGAAATAATGTCAGAGAGAATACATCAGGAAAAAGGTTGAAAATGTAAGTTCTA
3′-TTAAGAAACCTTTTTTCTGTTTCAACTTTATTACAGTCTCTCTTATGTAGTCCTTTTTCCAACTTTTACATTCAAGAT

 TTAAACACAACATACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGAC
 AATTTGTGTTGTATGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTG

 TCACATTTTTAATTTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAA
 AGTGTAAAAATTAAATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTT

 AGAGAATAAAAGACCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATT
 TCTCTTATTTTCTGGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAA

 GAACTTTACAGTGGTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAACACATAGATTTCATCGAAG
 CTTGAAATGTCACCACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTC

                                   4157
 TAAAAACTTCTGTGCATTAAGTAACATTC                                  GAAAATGTAAGTTCTATTAAACACAACAT
 ATTTTTGAAGACACGTAATTCATTGTAAG                                  CTTTTACATTCAAGATAATTTGTGTTGTA

 ACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGACTCACATTTTTAAT
 TGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTGAGTGTAAAAATTA

 TTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAAAGAGAATAAAAGA
 AATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTTTCTCTTATTTTCT

 CCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATTGAACTTTACAGTG
 GGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAACTTGAAATGTCAC

 GTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAACACATAGATTTCATCGAAGTAAAAACTTCTGT
 CACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTCATTTTTGAAGACA

 GCATTAAGTAACATTCTCAACAGAGTGAAAAGTCACAGAGGAAATATTTGCAAATCATATATCTGATAAGGGATTAATA-3′
 CGTAATTCATTGTAAGAGTTGTCTCACTTTTCAGTGTCTCCTTTATAAACGTTTAGTATATAGACTATTCCCTAATTAT-5′
```

```
5′-AATTCTTTGGAAAAAAGACAAAGTTGAAATAATGTCAGAGAGAATACATCAGGAAAAAGGTTGAAAATGTAAGTTCTA
3′-TTAAGAAACCTTTTTTCTGTTTCAACTTTATTACAGTCTCTCTTATGTAGTCCTTTTTCCAACTTTTACATTCAAGAT
                                                                  ↑
 TTAAACACAACATACTTAGAGGGGGAGAAAATCACAGAAGAGCCAGTTATGGCAGCTGAAAGAGGGGCAATGCTAAAGAC
 AATTTGTGTTGTATGAATCTCCCCCTCTTTTAGTGTCTTCTCGGTCAATACCGTCGACTTTCTCCCCGTTACGATTTCTG

 TCACATTTTTAATTTAAAAATCAGTGTTCCCTGGGCAGCAAGAAATTTTTGCCTATACTATATACAAAAAATTCATTCAA
 AGTGTAAAAATTAAATTTTTAGTCACAAGGGACCCGTCGTTCTTTAAAAACGGATATGATATATGTTTTTTAAGTAAGTT

 AGAGAATAAAAGACCTAAACCTGAGAGCTAAAACTCTAAAACCTTTAGAAGAAAACATAAGGGAAATGCTTCATGACATT
 TCTCTTATTTTCTGGATTTGGACTCTCGATTTTGAGATTTTGGAAATCTTCTTTTGTATTCCCTTTACGAAGTACTGTAA

 GAACTTTACAGTGGTGTCTTGGATATGATACCAAAAGCATGGACAATGAAAGAAAAAAAAACACATAGATTTCATCGAAG
 CTTGAAATGTCACCACAGAACCTATACTATGGTTTTCGTACCTGTTACTTTCTTTTTTTTTGTGTATCTAAAGTAGCTTC
                                  ↓
 TAAAAACTTCTGTGCATTAAGTAACATTCTCAACAGAGTGAAAAGTCACAGAGGAAATATTTGCAAATCATATATCTG-3′
 ATTTTTGAAGACACGTAATTCATTGTAAGAGTTGTCTCACTTTTCAGTGTCTCCTTTATAAACGTTTAGTATATAGAC-5′
```

TSD: Chr12:77126920-77127284

EN cleavage: 5’-TTTC/AA3’