THE FLOW OF GIBBON LAVA ELEMENT IS FACILITATED BY THE LINE-1 RETROTRANSPOSITION MACHINERY

by

Thomas Joshua Meyer

A CAPSTONE

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree

of

Master of Biomedical Informatics

December 2016

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Thomas J. Meyer

"The Flow of the Gibbon LAVA Element is Facilitated by the LINE-1 Retrotransposition Machinery"

Has been approved

Lucia Carbone, Capstone Advisor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is with much gratitude that I recognize that I owe my success in this Masters program to the excellent mentorship and guidance of my advisor, Lucia Carbone, and to the faculty and staff of the Department of Medical Informatics and Clinical Epidemiology of the Oregon Health & Sciences University. None of that success would have been possible, of course, without the support and love of my wife, Jennifer, my parents, the Nimbians, my friends, and my classmates. Without this group of people, I would have flagged on the course long ago. I am eternally indebted to each and every one of them.

The authors thank John Moran for providing plasmids pJM101/L1.3, pJM101/L1RP, and pJM101/L1RPΔneo and the cell lines HeLa-HA. We are grateful to John Goodier and Haig Kazazian who provided us with the reliable αL1ORF2p-N antibody. We are indebted to Kay-Martin Hanschmann for statistical analyses. This work was supported by the Deutsche Forschungsgemeinschaft [DA 545/2-1 and SCHU 1014/8-1 to G.G.S.] and the National Institute of Health [NCRR P51, RR000163 to LC]. Moreover, T.J.M. was supported by the National Library of Medicine of the National Institutes of Health under Award Number T15LM007088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ABSTRACT

LAVA (**L**INE-***A****lu*-**V**NTR-***A****lu*-like) elements comprise a family of non-autonomous, composite, non-LTR retrotransposons specific to gibbons and may have played a role in the evolution of this lineage. A full-length LAVA element consists of portions of repeats found in most primate genomes: CT-rich, *Alu*-like, and VNTR regions from the SVA retrotransposon, and portions of the *Alu*Sz and L1ME5 elements. To evaluate whether the gibbon genome currently harbors functional LAVA elements capable of mobilization by the endogenous LINE-1 (L1) protein machinery and which LAVA components are important for retrotransposition, we established a *trans*-mobilization assay in HeLa cells. Specifically, we tested if a full-length member of the older LAVA subfamily C that was isolated from the gibbon genome and named LAVA_{C}, or its components, can be mobilized in the presence of the human L1 protein machinery. We show that L1 proteins mobilize the LAVA_{C} element at frequencies exceeding processed pseudogene formation and human SVA_{E} retrotransposition by >100-fold and ≥3-fold, respectively. We find that only the SVA-derived portions confer activity, and truncation of the 3’ L1ME5 portion increases retrotransposition rates by at least 100%. Tagged *de novo* insertions integrated into intronic regions in cell culture, recapitulating findings in the gibbon genome. Finally, we present alternative models for the rise of the LAVA retrotransposon in the gibbon lineage.

The flow of the gibbon LAVA element is facilitated by the LINE-1 retrotransposition machinery

Thomas J. Meyer[1,2], Ulrike Held[3], Kimberly A. Nevonen[1], Sabine Klawitter[3,#], Thomas Pirzer[3], Lucia Carbone[1,2,4,*], and Gerald G. Schumann[3,*]

[1]Division of Neuroscience, Oregon National Primate Research Center, Beaverton, OR 97006, USA; [2]Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239-3098, USA; [3]Division of Medical Biotechnology, Paul-Ehrlich-Institut, Paul-Ehrlich-Strasse 51-59, D-63225 Langen, Germany; [4]Department of Medicine, Oregon Health & Science University, Portland, OR 97239-3098, USA.

[#]Current address: Division of Inborn Metabolic Diseases, University Children's Hospital, D-69120 Heidelberg, Germany.

[*]Authors for Correspondence:

Gerald G. Schumann, Division of Medical Biotechnology, Paul-Ehrlich-Institut, Paul-Ehrlich-Strasse 51-59, D-63225 Langen, Germany; Tel:+49-6103-773105; Fax:+49-6103-771280; Email: Gerald.Schumann@pei.de

Lucia Carbone, Division of Neuroscience, Oregon National Primate Research Center, 505 NW 185th ave, Beaverton, OR 97006, USA; Phone: +1 503 629 4057; Email: carbone@ohsu.edu

ABSTRACT

LAVA (**L**INE-***A****lu*-**V**NTR-***A****lu*-like) elements comprise a family of non-autonomous, composite, non-LTR retrotransposons specific to gibbons and may have played a role in the evolution of this lineage. A full-length LAVA element consists of portions of repeats found in most primate genomes: CT-rich, *Alu*-like, and VNTR regions from the SVA retrotransposon, and portions of the *Alu*Sz and L1ME5 elements. To evaluate whether the gibbon genome currently harbors functional LAVA elements capable of mobilization by the endogenous LINE-1 (L1) protein machinery and which LAVA components are important for retrotransposition, we established a *trans*-mobilization assay in HeLa cells. Specifically, we tested if a full-length member of the older LAVA subfamily C that was isolated from the gibbon genome and named LAVA$_C$, or its components, can be mobilized in the presence of the human L1 protein machinery. We show that L1 proteins mobilize the LAVA$_C$ element at frequencies exceeding processed pseudogene formation and human SVA$_E$ retrotransposition by >100-fold and ≥3-fold, respectively. We find that only the SVA-derived portions confer activity, and truncation of the 3’ L1ME5 portion increases retrotransposition rates by at least 100%. Tagged *de novo* insertions integrated into intronic regions in cell culture, recapitulating findings in the gibbon genome. Finally, we present alternative models for the rise of the LAVA retrotransposon in the gibbon lineage.

Key words: gibbon, retrotransposon, LAVA, LINE-1, *trans*-mobilization assay.

INTRODUCTION

Repetitive and repeat-derived sequences have been estimated to comprise 66-69% of the human genome (Lander et al. 2001; de Koning et al. 2011), with similar proportions being annotated in other non-human primate genomes (The Chimpanzee Sequencing and Analysis Consortium 2005; Gibbs et al. 2007; Locke et al. 2011). The majority of these sequences are non-long terminal repeat (LTR) retrotransposons, which include the LINE-1 (L1) and *Alu* families, as well as the composite SVA (SINE-R-VNTR-*Alu*) family (Cordaux and Batzer 2009). In addition to being the largest contributors of genetic variation during species evolution, insertions of these elements have been linked to human diseases, including cancer (Callinan and Batzer 2006; Belancio et al. 2009; Belancio et al. 2010; Konkel and Batzer 2010; Hancks and Kazazian 2012; Kaer and Speek 2013). Further, the high sequence similarity between copies within each family can result in large-scale structural changes by inducing non-homologous allelic recombination (NAHR) events that can result in duplications, deletions, and inversions of the intervening sequence, including genes and other functional sequences, hence resulting in a disease phenotype (Han et al. 2007b; Belancio et al. 2008; Han et al. 2008; Lee et al. 2008; Belancio et al. 2010).

LAVA elements are a newly discovered family of non-autonomous, composite, non-LTR retrotransposons that can be classified in 22 subfamilies which originated and expanded exclusively in one group of primates: the gibbons, or small apes (Carbone et al. 2012; Carbone et al. 2014) (Supplementary Figure S1). Gibbons are part of the superfamily Hominoidea and shared a common ancestor with hominids (humans and other great apes) around 17 million years ago (mya) (Carbone et al. 2014). Although evolutionarily close to the hominids, gibbons have some very distinct genomic traits, including highly divergent karyotypes as a consequence of numerous chromosomal rearrangements that occurred during a relatively short period of time. As a consequence, each of the four gibbon genera

(*Nomascus*, *Hylobates*, *Hoolock,* and *Symphalangus*) carries a different karyotype with chromosome numbers ranging from 38 to 52 (Muller et al. 2003) although the genera diverged from each other as recently as about 5 mya (Veeramah et al. 2015). The recent analysis of the first gibbon genome assembly uncovered evidence that the appearance of the LAVA element might have been associated with the accelerated karyotype evolution in the gibbon lineage through the disruption of genes linked to chromosome segregation (Carbone et al. 2014). Finally, one of the gibbon genera, the *Hoolock*, displays a remarkable expansion of the LAVA element detectable by fluorescent *in situ* hybridization in almost all centromeres (Carbone et al. 2012), further highlighting a link between the LAVA element and the karyotype variability observed in gibbons. Altogether, these observations suggest that the LAVA element was an evolutionary novelty that may have greatly impacted the evolution of the gibbon lineage through several different avenues.

The LAVA element can be described as a “composite-of-composites” because, together with fragments from other repeats, it includes a portion from the hominoid-specific SVA element, a composite non-LTR retrotransposon itself (Figure 1A). Specifically, it consists of the CT-rich, *Alu*-like, and VNTR (variable number tandem repeat) regions of the SVA element, a portion of an *Alu*Sz element, and a portion of an L1ME5 element. Two small spacer sequences, derived from a unique genomic locus in the genome of the anthropoid common ancestor, separate SVA and *Alu*Sz components (28-bp U1 sequence) and *Alu*Sz and L1ME5 modules (103-bp U2 sequence) (Figure 1A). Given their unique location in the other primate genomes, these spacers allow the identification of the putative progenitor locus for the 3’ end of the LAVA element. Just such a locus containing the U1-*Alu*Sz-U2-L1ME5 sequence can be found on gibbon chromosome 1a, human chromosome 9, and the orthologous loci in all other great apes, old world monkeys (i.e. macaque), and marmoset genomes. The steps leading to the origin of the LAVA element, however, are still not entirely clear.

During the analysis and annotation of the gibbon genome assembly, all 1097 full-length LAVA elements were retrieved and classified into 22 subfamilies (Carbone et al. 2014) based on the presence/absence of diagnostic mutations (Supplementary Figure S1). These 22 subfamilies were further classified into six larger subfamilies designated LAVA_A through LAVA_F. The age of LAVA copies of the oldest subfamilies (LAVA_A1 and LAVA_A2) was estimated to be about 17-18 my, and therefore very close to the inferred date of divergence of the small ape and great ape lineages (Carbone et al. 2014). This observation suggests that the LAVA element was present from the very beginning of the gibbon lineage and the rise of LAVA might actually be one of the events involved in gibbon divergence. Subfamily LAVA_B consists of a large and diverse expansion of ten small and midsized subfamilies that evolved between 12-18 mya. One early branch of this LAVA_B expansion, LAVA_B1G, led to the lineage of the remaining middle-aged and younger LAVA subfamilies. LAVA_C and LAVA_D consist of three and two subfamilies aged 16-18 and 13-15 my, respectively, while LAVA_E and LAVA_F contain the most recently-evolved subfamilies. The single large LAVA_E subfamily and the three LAVA_F subfamilies are estimated to have evolved between 10-11 mya and 6-9 mya, respectively (Carbone et al. 2014). Analyses of the evolutionary history of the LAVA elements identified in the gibbon reference genome revealed that about 50% of LAVA insertions have been fixed in all gibbon species, indicating that this retrotransposon was active before the split of the four gibbon genera about 5 mya (Carbone et al. 2014; Veeramah et al. 2015). Additionally, the discovery of polymorphic insertions specific to only some individuals of one species (*Nomascus leucogenys*) (Carbone et al. 2014) implies that LAVA activity was ongoing, at least until recently, and suggests the possibility that some source elements could be currently active.

Endogenous LAVA insertions display all the hallmarks of L1-mediated retrotransposition by target-primed reverse transcription (TPRT) (Carbone et al. 2014), including a consensus target sequence closely matching the human L1 endonuclease (EN) cleavage motif (5'-TTTT/AA-3'), target site duplications (TSDs) flanking each insertion, variable length poly(A) tails following putative polyadenylation signals, and the existence of 5'-truncated LAVA insertions (Szak et al. 2002; Ostertag et al. 2003). It has been demonstrated experimentally that both *Alu* and SVA elements, which lack any protein-coding capacity, accomplish their retrotransposition by hijacking the L1-encoded protein machinery in *trans* (Dewannieux et al. 2003; Raiz et al. 2012). In order to test the hypothesis that the gibbon genome harbours functional LAVA elements that can be *trans*-mobilized by the L1-encoded protein machinery, we established a LAVA retrotransposition reporter assay in HeLa cells.

Taking advantage of the modular structure of full-length LAVA elements, we set out to test the retrotransposition competence of its components relative to the full-length element in the presence of human L1 proteins. To this end, we picked one (GL397285: 9072637-9074632 in Nleu1.5) of the 207 members of the LAVA subfamily LAVA_C4B (Supplementary Figure S1) from the gibbon genome (Carbone et al. 2014). Here we show that this full-length LAVA element (hereafter referred to as LAVA$_C$) is able to mobilize in human cells at frequencies exceeding those of a human SVA$_E$ retrotransposon. Data demonstrate that the SVA-derived portion makes the LAVA$_C$ element a preferred substrate for *trans*-mobilization by the human L1 protein machinery. Surprisingly, removal of the L1ME5-derived portion increased the mobilization rate of LAVA$_C$, indicating that the 3'-truncated element is mobilized more efficiently than the full-length LAVA$_C$. We also found that *mneo*I-tagged LAVA *de novo* insertions integrate into intronic regions of human genes in cell culture experiments recapitulating endogenous LAVA insertion preferences in the gibbon genome. Our results are in agreement with the inferred evolutionary origin of the LAVA element and indicate that there are genomic LAVA elements that are still “hot and flowing”, explaining why numerous insertions are polymorphic within the gibbon lineage.

MATERIALS AND METHODS

Selection and isolation of the LAVA element to be used for *trans*-mobilization assays

The full-length LAVA$_C$ element studied in this work was recently identified within the insert of the BAC clone CHORI 271-458C4 (Carbone et al. 2014) of the *Nomascus leucogenys* genome library (CHORI 271, https://bacpac.chori.org/library.php?id=228) that was also used for the gibbon genome assembly (Nleu1.0). A glycerol stock of CHORI271-458C4 was obtained from BACPAC Resources (https://bacpac.chori.org/). Using the publicly available BAC sequence (AC202765.2) as a template, we designed primers (GS_LAVA_19_F and GS_LAVA_19_R, Supplementary Table S1) flanking LAVA$_C$ and performed long-range PCR on isolated BAC DNA using standard protocols (Carbone et al. 2014). The resulting 2418-bp PCR product harbouring the full-length LAVA$_C$ element was isolated after separation in a 1% agarose gel, and sub-cloned in the pDRIVE vector (Qiagen) leading to the plasmid pDRIVE.LAVA$_C$. Subsequently, primer walking was carried out to validate the isolated LAVA$_C$ element in pDRIVE.LAVA$_C$ by sequence analysis (Supplementary Figure S2).

LAVA$_C$-specific filled site/empty site PCR

Genomic DNA was isolated from blood of 11 individuals of the genus *Nomascus leucogenys*, 3 individuals of the genus *N. gabriellae*, and 1 individual of each of the genera *Hoolock leuconedys*, *Hylobates moloch*, and *Symphalangus syndactylus* applying the PureGene DNA Isolation Kit (Qiagen) according to the manufacturers instructions. To test for presence/absence of LAVA$_C$, 25 µl PCR reactions were made each containing 50 ng genomic DNA, 1X Pfu Turbo Buffer, 0.32 µM LAVA19F primer, 0.32 µM LAVA19R (Supplementary Table S1), 0.3 µM dNTPs, 1 U Platinum Taq (Invitrogen), 0.1 U Pfu Turbo Cx, and water. Cycling parameters for the reactions were: 95°C for 3 min, 30 cycles of 95°C for 30s, 61°C for 30s, 72°C for 8 min, and a final extension at 72°C for 10 min. Samples in order of loading on the 1% agarose gel presented in Supplementary Figure S3 are from *N. leucogenys* individuals Vok, Asia, China, Nancy, Johannes, Ricky, Bobby, B09007, Melouprey, Gibson, and Khao, *N. gabriellae* individuals Victor, China, Enick, and *H. leuconedys, H. moloch,* and *S. syndactylus* individuals Drew, Khusus, and Monty, respectively.

Retrotransposition reporter constructs and L1 protein donor plasmids

The 2418-bp, full-length LAVA$_C$ element was isolated by PCR amplification from pDRIVE.LAVA$_C$ using primers GS.LAVA1 and GS.LAVA2 (Supplementary Table S1, Supplementary Figure S2). LAVA$_C$ fragments of 2,096 bp, 819 bp and 473 bp including different portions of the LAVA$_C$ element (Figure 1A), were isolated by PCR using primer combinations GS.LAVA1/GS.LAVA3, GS.LAVA1/GS.LAVA4, and GS.LAVA2/GS.LAVA5 (Supplementary Table S1, Supplementary Figure S2), respectively. Each of the four different PCR products was cloned into the pGEM-Teasy vector (Promega), sequenced, and the obtained sequences were compared with the LAVA$_C$ sequence of pDRIVE-LAVA$_C$ in order to check for PCR artifacts (data not shown). After validation of the correct nucleotide sequences of the PCR fragments, each of them was cloned as KpnI/NheI fragment into the KpnI/NheI digested pCEPneo plasmid (Raiz et al. 2012) as KpnI and NheI restriction sites are part of primers GS.LAVA1, GS.LAVA5 and GS.LAVA2, GS.LAVA3, GS.LAVA4, respectively (Supplementary Table S1). Using this strategy, each of the four PCR products was inserted between CMV$_P$ and the *mneo*I indicator cassette of pCEPneo via KpnI/NheI, yielding pLC10/LAVA$_C$ (2418-bp fragment), pLC5/LAVA$_C$ (2096-bp fragment), pLC21/LAVA$_C$ (819-bp fragment), and pLC23/LAVA$_C$ (473-bp fragment) (Figure 1B). The construct pJM101/L1$_{RP}$ΔneoΔORF1 (L1$_{RP}$ΔORF1) was generated by introducing a 330-bp in-frame deletion in L1 ORF1 of pJM101/L1$_{RP}$Δneo (Wei et al. 2001; Raiz et al. 2012) (Figure 1C).

LAVA retrotransposition reporter assays and statistical analyses

HeLa-HA cells (Hulme et al. 2007) were cultured in DMEM High Glucose (Biochrom AG, Berlin, Germany) supplemented with 10% FCS (Biowest, Nuaillé, France), 100 µg/ml streptomycin and 100 U/ml penicillin. To perform retrotransposition reporter assays, 1.8x10^{6} cells were plated on T75-flasks. Plated cells were co-transfected with 3 µg of each of the four LAVA$_{C}$ retrotransposition reporter plasmids, the SVA reporter plasmid, or pCEPneo and 3 µg of an L1 protein donor construct (pJM101/L1$_{RP}$Δneo, pJM101/L1$_{RP}$ΔneoΔORF1; Figure 1C) or pCEP4 (Thermo Fisher Scientific) using FUGENE HD reagent (Promega) according to the manufacturer's instructions. The L1 *cis*-retrotransposition rate observed after co-transfection of 3 µg of the L1$_{RP}$-encoding pJM101/L1$_{RP}$ reporter construct with 3 µg pCEP4, served as positive control and was defined as 100% retrotransposition frequency. Each co-transfection was performed in biological triplicates using three independent DNA preparations of each plasmid, and each biological replicate was executed on a different day. Each cotransfection performed on the same day was done in quadruplicate. In each case, three co-transfections of the same biological replicate were used to quantify retrotransposition rates of the LAVA, SVA and L1 reporter elements and pseudogene formation rates of the pCEPneo construct. The fourth co-transfection was used to isolate cell lysates and total RNA in order to analyze expression of L1 proteins expressed from the L1 donor plasmids and transcription of retrotransposition reporter cassettes, respectively. Consequently, the arithmetic mean of the retrotransposition frequency of each analyzed reporter element was calculated from nine cotransfection experiments (N=9). Starting 24 hours post-transfection, cells were subjected to hygromycin

(200 µg/ml, Invitrogen) selection for 12 days. After trypsinization and re-seeding, cells were selected for L1-mediated retrotransposition events in medium containing 400 µg/ml G418 (Invitrogen). After 11-12 days of selection, G418^R colonies were either fixed and stained with Giemsa (Merck) to quantify retrotransposition events as described previously (Moran et al. 1996), or individual G418^R colonies were isolated and expanded to characterize individual LAVA *de novo* retrotransposition events.

To gain countable results for retrotransposition in *cis*, cells co-transfected with pJM101/L1$_{RP}$ and pCEP4 were trypsinized after 12 days of hygromycin selection. Obtained cells were diluted 1:10 and 1:100, and diluted cells were reseeded in T75 flasks. 24 hours later, cells were selected for L1 retrotransposition events in medium containing 400 µg/ml G418. Statistical evaluation was performed by means of an Analysis of Variance (ANOVA). To control the overall type I error (α=0.05), p-values were adjusted according to Dunnett for multiple comparisons. The statistical analysis was performed with SAS®/STAT software, version 9.2 SAS system for Windows.

Immunoblot analysis

To assess L1 ORF1p and L1 ORF2p expression, HeLa cells were co-transfected as described above and harvested after 12 days of hygromycin selection. Cells were lysed in RIPA buffer (25 mM Tris, pH 8, 137 mM NaCl, 1% glycerol, 0.5% sodium deoxycholate, 1% Nonidet P-40, 2 mM EDTA, pH 8, 0.1% SDS, and protease inhibitors), and lysates were cleared by centrifugation. Protein concentrations were quantified by applying the Pierce™ BCA Protein Assay Kit (Cat.# 23225, Thermo Fisher Scientific) according to the manufacturers instructions. The particular amounts of protein lysates (described in the following) were boiled in Laemmli buffer, loaded on 12% polyacrylamide gels, subjected to SDS-PAGE, and electroblotted onto nitrocellulose membranes. In the case of the immunoblot analyses presented in Figures 2A and 3B/left panel (αL1 ORF1p blot), 20 µg of cell lysate of each differently transfected HeLa-HA culture were loaded per lane. In the case of the gel presented in Figure 3B/right panel (αL1 ORF2p blot), 40 µg of total cell lysate of each differently transfected HeLa-HA culture were loaded per lane. Only 2 µg of the cell extracts isolated from the human embryonal carcinoma cell line NTERA.2 cl.D1 (NTERA-2; ATCC#: CRL-1973; (Andrews et al. 1984)) were loaded on one lane of each of the three immunoblot analyses presented in Figures 2A and 3B and served as positive control for L1-ORF1p and L1-ORF2p detection. After protein transfer, membranes were blocked for two hours at room temperature in a 10% solution of non-fat milk powder in 1x PBS-T (137 mM NaCl, 3 mM KCl, 16.5 mM Na_2HPO_4, 1.5 mM KH_2PO_4, 0.05% Tween 20 (Sigma)), washed in 1xPBS-T, and incubated overnight with the respective primary antibody at 4°C. To detect L1 ORF1p, the polyclonal rabbit-anti-L1 ORF1p antibody #984 (Raiz et al. 2012) was used in a 1:2000 dilution in 1xPBS-T containing 5% milk powder. L1 ORF2p expression was verified using a rabbit anti-L1 ORF2p-N antibody (Goodier et al. 2004) at a 1:1000 dilution in 1xPBS, 5% milk, 0.05% Tween 20. Membranes were washed thrice in 1xPBS-T and incubated with an HRP-conjugated, secondary anti-rabbit IgG antibody (Amersham Biosciences) at a dilution of 1:30,000 in 1xPBS-T/5% milk powder for 1 hour. Subsequently, the membrane was washed six times for 10 min in 1xPBS-T. ß-actin expression was detected using a monoclonal anti-ß-actin antibody (clone AC-74, Sigma-Aldrich) as primary antibody at a dilution of 1:30,000. Anti-mouse HRP-linked species-specific antibody (from sheep) at a dilution of 1:10,000 served as secondary antibody specific for anti-ß-actin. Immunocomplexes were visualized using lumino-based ECL immunoblot reagent (Amersham Biosciences).

Analysis of LAVA *de novo* insertions

Genomic DNA from expanded G418^R-HeLa colonies was isolated applying the Qiagen DNeasy® Tissue Kit according to the manufacturer’s protocol. To test for the presence of the spliced *mneo*I indicator cassette, a diagnostic PCR was performed using the intron-flanking primer pair GS86/GS87 (Figure 4A, Supplementary Table S1). PCR cycling conditions were as follows: 3 min at 96°C, (30 sec at 96°C, 15 sec at 56°C; 2 min at 72°C) 25 cycles, 7 min at 72°C. To determine genomic pre- and post-integration sites of LAVA *de novo* insertions, we used a modified version (Kirilyuk et al. 2008) of a previously published **e**xtension **p**rimer **t**ag **s**election preceding solid-phase **l**igation-**m**ediated **PCR** (EPTS/LM-PCR) (Schmidt et al. 2001) to isolate 3’ junctions of these insertions. Products of the final PCR were separated in a 1% agarose gel, isolated from the gel using the QIAquick Gel Extraction Kit (Qiagen), and sequenced either directly or after sub-cloning into pGEM-T Easy. Obtained sequences were mapped to the human genome using the UCSC genome browser at http://genome.ucsc.edu. To characterize 5’-junctions of each LAVA *de novo* insertion, primers specific for the genomic sequence adjacent to the 5’ end of the *de novo* insertions were designed (Supplementary Table S2). The second PCR primers used were designed to bind specifically to the retrotransposed LAVA$_C$ reporter cassette (LAVA19_5pSVA1 and LAVA19_5pU2A). All oligonucleotides used in this study are listed in Supplementary Table S1. Genomic pre-integration sites and surrounding sequences were characterized using the UCSC genome browser annotation for genes (Supplementary Table S3).

Sequence logos

Using the WebLogo utility (Schneider and Stephens 1990; Crooks et al. 2004), sequence logos were generated for the genomic integration sites of LAVA elements, L1PA elements found in the gibbon genome, and L1-Ta elements reported for the human genome.

The 34 LAVA target sites used in this analysis included 33 loci described in a recent report (Carbone et al. 2012) and the original LAVA$_C$ locus used to design the constructs in this study. The 70 genomic target sequences of human endogenous L1Ta elements were listed in a previous study (Raiz et al. 2012) (Figure 4C, Supplementary Table S4). In order to extract genomic target sequences of L1PA elements from the gibbon genome, we first used the RepeatMasker (Smit et al. unpublished) annotation of Nleu3.0 available from UCSC's Table Browser, to select 200 full-length L1PA insertions randomly. Next, 100-bp marginal sequences consisting of +/- 50 bp upstream and downstream of the start and end position of each element were extracted. These pairs of marginal sequences were then aligned using Clustal Omega (Sievers et al. 2011) and the alignments were visualized in Geneious version 5.6.5 (http://www.geneious.com) (Kearse et al. 2012) and inspected to call TSDs and EN cleavage sites, if possible, for each locus. Only 91 of the 200 L1PA loci inspected showed identifiable TSDs, which allowed inference of EN cleavage sites (Figure 4C, Supplementary Table S4).

Quantitative Real-Time RT-PCR to analyze transcription of reporter cassettes after cotransfection of driver and reporter plasmid

To quantify *mneo*I-transcripts expressed from the various reporter plasmids in our assay system, plated HeLa cells were cotransfected with 3 µg of each of the four LAVA$_C$ reporter plasmids, the SVA reporter plasmid, or pCEPneo, and 3 µg of the L1 driver pJM101/L1$_{RP}$Δneo, or with 3 µg pCEP4 and 3 µg of pJM101/L1$_{RP}$. Hygromycin selection started 24 hours post-transfection. Each co-transfection was performed in three biological replicates with each biological replicate being executed with a separate set of independent plasmid preparations. 13 days after co-transfection of HeLa cells, total RNA was extracted from hygromycin-selected HeLa cells using the RNeasy® Mini Kit (Qiagen) with On-Column

DNase digestion following the manufacturer’s instructions. Using the *SuperScript III*® First-Strand Synthesis System for RT-PCR (Invitrogen) in combination with a hexamer primer, first-strand cDNA was synthesized from 0.5 µg of DNaseI-digested, total RNA according to the manufacturer’s instructions. To quantify levels of spliced transcripts expressed from the *mneo*I-tagged reporter elements in pAD3/SVA$_E$, pLC5/LAVA$_C$, pLC10/LAVA$_C$, pLC21/LAVA$_C$, pLC23/LAVA$_C$, pCEPneo and pJM101/L1$_{RP}$, real-time PCR was performed as technical triplicates applying TaqMan® chemistry (Applied Biosystems) in an Applied Biosystems 7900HT Fast Real-Time PCR System base unit. We used a primer/probe combination (Neofor: 5’-GCTATTCGGCTATGACTGG-3’; Neorev: 5’-GCCACGATAGCCGCGCTGC-3’; probe: 5´-FAM-CCTCGTCCTGAAGCTCATTC-3´) specifically recognizing the spliced *mneo*I cassette. Spliced *mneo*I transcript levels were normalized against cellular 18srRNA levels. Cycling conditions were as follows: 95°C for 15 min (initial denaturation), 95°C for 15s and 60°C for 1 min (40 cycles). The software applied to analyze real-time and end point fluorescence was RQ manager 1.2. Relative quantification of RNA expression was carried out using the $\Delta\Delta C_t$ method (Livak and Schmittgen 2001).

RESULTS

Identification of a potentially active LAVA source element in the *Nomascus leucogenys* genome

In order to investigate the competency of non-autonomous LAVA retrotransposons to be *trans*-mobilized by the L1 protein machinery, we chose a full-length LAVA element from the *Nomascus leucogenys* genome. The selected LAVA insertion was chosen among 48 elements characterized in a previous study (Carbone et al. 2012; Carbone et al. 2014) because it does not contain any obvious indels, includes all complete LAVA-specific modules, a 3' poly(A) tail, and is flanked by clearly identifiable TSDs covering 12 nts. A PCR-based assay over a larger panel of gibbon individuals confirmed that the same element inserted specifically in the *Nomascus* genus, as it can be found both in *Nomascus leucogenys* and its sister taxon *Nomascus gabriellae*, but is absent from the remaining genera (Supplementary Figure S3). This indicates that this specific insertion occurred less than 5 mya after the four gibbon genera diverged from each other but had the time to be fixed in the *Nomascus* genus. Consistently, comparison of its sequence to the consensus sequences identified for the 22 LAVA subfamilies (Carbone et al. 2014) indicated similarity to the older LAVA_C4B subfamily which, based on previous age analysis, emerged ~16-18 mya (Carbone et al. 2014). We hence named this element LAVA$_C$ and set out to investigate whether this element, whose sequence differs from the youngest subfamilies LAVA_F and LAVA_E, is retrotransposition-competent. Given that older LAVA elements such as LAVA$_C$ represent the bulk of LAVA insertions in the gibbon genome, the finding of its activity would have significant biological implications (Supplementary Figure S1).

Engineering functional LAVA$_C$ reporter elements to be applied in a *trans*-mobilization assay

To investigate if the full-length LAVA$_C$ element can be mobilized in *trans* by the human L1 protein machinery and assess which of the LAVA$_C$–encoded modules are relevant for such mobilization, we generated reporter plasmids expressing the full-length element (pLC10/LAVA$_C$) and different combinations of its modules (pLC5/LAVA$_{C,}$ pLC21/LAVA$_C$, and pLC23/LAVA$_C$), respectively (Figure 1B). To this end, the full-length LAVA$_C$ and each of the three LAVA$_C$ fragments were inserted separately between the CMV promoter (CMV$_P$) and the *mneo*I retrotransposition reporter cassette of the pCEPneo plasmid (Raiz et al. 2012). Each of the resulting LAVA reporter constructs was co-transfected into HeLa-HA cells (Hulme et al. 2007) with the L1 driver construct pJM101/L1$_{RP}$Δneo (Figure 1C) (Wei et al. 2001) that provides the overexpressed human L1 protein machinery (Figure 2A) in the presence of *mneo*I-tagged LAVA$_C$ transcripts. To directly compare retrotransposition activities of the four different LAVA$_C$-derived reporter cassettes (Figure 1B) with the recently determined *trans*-mobilization frequency of the human full-length SVA$_E$ element, we also co-transfected the SVA$_E$ reporter plasmid pAD3/SVA$_E$ (Figure 1B,C) (Raiz et al. 2012) with the L1 driver plasmid in HeLa-HA cells in parallel. Concurrently, we used this co-transfection with pAD3/SVA$_E$ also as positive control for the established *trans*-mobilization assay. Processed pseudogene formation rate was assessed by measuring the *trans*-mobilization frequency of the reporter cassette encoded by the empty pCEPneo plasmid (Raiz et al. 2012) in the presence of the co-transfected L1 driver plasmid (Figure 1C). As pCEPneo differs from the LAVA and SVA reporter plasmids exclusively in the absence of any retrotransposon sequence (Figure 1B) and expresses solely the *mneo*I cassette, it allows obtaining a rough estimate of the *trans*-mobilization frequency of PolII genes in our experimental setup (Esnault et al. 2000). More details on the generation of the LAVA retrotransposition reporter constructs and the LAVA retrotransposition reporter assay are outlined in the *Material and Methods* section.

Each co-transfection was performed in three biological replicates with each replicate being executed as technical triplicates (for details see Materials and Methods). After co-transfection, cells were cultivated in the presence of hygromycin to select for the presence of the expression plasmids, and subsequently selected for G418 resistance (Figure 1C). Each of the retrotransposition reporter cassettes used in this study was tagged with the selectable marker gene *mneo*I coding for neomycin phosphotransferase, in an antisense orientation. This combination assures that G418-resistant (G418^R) cells only originate if a transcript initiated from the CMV promoter (CMV$_P$) driving LAVA*mneo*I, SVA*mneo*I, CEP*mneo*I or L1*mneo*I expression is spliced, reverse transcribed, reintegrated into the genomic DNA, and expressed from promoter P' (Figure 1B) (Heidmann et al. 1988; Freeman et al. 1994; Moran et al. 1996).

Retrotransposition frequencies of the different LAVA$_C$ reporter elements in the presence of the overexpressed L1 driver plasmid ranged from 0.02% to 4.7% relative to the L1$_{RP}$ *cis*-retrotransposition frequency set as 100% (Figure 2B, and Supplementary Figures S4 and S5). The absence of G418^R colonies after co-transfection of LAVA$_C$ or SVA$_E$ reporter plasmids or pCEPneo with the empty episomal expression vector pCEP4 substantiates that overexpression of L1-encoded proteins is essential for the *trans*-mobilization of LAVA$_C$, SVA$_E$ and for pseudogene formation and consistent with previous reports on *trans*-mobilization of SVA elements (Hancks et al. 2011; Raiz et al. 2012) and processed pseudogene formation (Esnault et al. 2000).

Human L1$_{RP}$ proteins *trans*-mobilize full-length LAVA$_C$ at a frequency exceeding processed pseudogene formation significantly

First, we asked if L1 proteins favor the LAVA$_C$ reporter-encoded mRNA over any random RNA polymerase II transcript as substrate for *trans*-mobilization. In this case, *trans*-

mobilization frequency of the LAVA reporter element should significantly exceed the frequency of processed pseudogene formation of a random PolII gene. In order to determine the processed pseudogene formation rate, we used the pCEPneo reporter plasmid. Since transcripts expressed from the CMV$_P$ of pCEPneo consist exclusively of the *mneo*I indicator cassette in antisense orientation and do not include any retrotransposon sequences (Raiz et al. 2012), these transcripts should be *trans*-mobilized as frequent as random mRNAs encoded by host PolII genes. *Trans*-mobilization frequencies of the canonical full-length LAVA$_C$ element expressed from pLC10/LAVA$_C$ exceeded pseudogene formation of the reverse *mneo*I cassette encoded by pCEPneo by ~117-fold (Figure 2B). LAVA$_C$ reporter plasmids expressing exclusively the CT-rich and *Alu*-like regions (pLC21/ LAVA$_C$) or CT-rich region, *Alu*-like region, VNTR module, and the *Alu*Sz-derived component (pLC5/ LAVA$_C$) lead to *trans*-mobilization frequencies outreaching processed pseudogene formation frequency by 15- or 226-fold, respectively (Figure 2B, and Supplementary Figures S4 and S5). In contrast, *trans*-mobilization frequency of the LAVA$_C$ reporter element comprising only the 3’ end modules, U1, *Alu*Sz, U2, and L1ME5, (pLC23/LAVA$_C$) does not significantly exceed processed pseudogene formation frequency of pCEPneo (Figure 2B). Reporter constructs pLC10/LAVA$_C$ (full-length) and pLC5/LAVA$_C$ (ΔL1ME5 module), whose expression results in the highest retrotransposition frequencies, are characterized by the presence of the 5’ portion of the full-length LAVA$_C$ element ranging from the CT-rich region to the U2 region (Figure 1A) suggesting that the included modules make LAVA$_C$ a preferred substrate for *trans*-mobilization by the L1 protein machinery. Interestingly, the pAD3/SVA$_E$–encoded SVA$_E$ reporter element that is mobilized at a frequency exceeding pseudogene formation by ~33-fold has CT-region, *Alu*-like region, and VNTR module in common with LAVA$_C$ (Figure 1A). This observation strengthens the hypothesis that the SVA-derived modules account for at least some of the preference the L1 protein machinery has for both LAVA and

SVA elements, while LAVA-specific U1, *Alu*Sz, U2, and L1ME5 modules do not support *trans*-mobilization by L1 proteins.

Quantification of overexpressed human L1 proteins by immunoblot analysis demonstrates that the observed differences in G418^R colonies are not a consequence of varying amounts of overexpressed L1 proteins (Figure 2A). To evaluate if the observed differences in G418^R colonies resulted from discrepancies in mRNA production or stability between the different LAVA$_C$ reporter cassettes, we tested for comparable amounts of spliced tagged LAVA$_C$ mRNAs by quantitative real-time RT-PCR (qRT-PCR) (Figure 3A). To this end, we isolated total RNA from each of the three biological replicates of each cotransfection experiment after 12 days of hygromycin selection, applied primer/probe combinations specifically recognizing the spliced *mneo*I reporter cassette (Raiz et al. 2012) (see Materials and Methods) and quantified the relative amounts of spliced mRNA expressed from the reporter plasmids by Real-Time Quantitative Reverse Transcription PCR (qRT-PCR) (Figure 3A). Data show that the total quantities of spliced *mneo*I-carrying reporter mRNAs expressed in pAD3/SVA$_E$-, pLC21/LAVA$_C$-, pLC23/LAVA$_C$-, or pJM101/L1$_{RP}$-transfected cells are comparable, and reduced by only ~22, 39, 44 or 24%, respectively, relative to the reporter mRNA level expressed from pLC10/LAVA$_C$ (Figure 3A). These comparatively small differences in reporter transcript levels have only a negligible effect on the relative differences observed between *trans*-mobilization frequencies of the respective reporter elements (Figure 2B and Supplementary Figure S5). However, we observed *mneo*I–harbouring reporter transcript levels expressed from pLC5/LAVA$_C$ that are significantly reduced by 55 to 75% relative to the transcript levels expressed from the remaining transfected LAVA, SVA and L1 retrotransposition reporter plasmids (Figure 3A).

Considering the reduced spliced reporter transcript levels expressed in pLC5/LAVA$_C$-transfected cells (Figure 3A), we would expect that the observed 122-fold increase in *trans*-

mobilization frequency of the LAVA$_C$ reporter encoded by pLC5/LAVA$_C$ relative to pLC23/LAVA$_C$ (Figure 2B, Supplementary Figure S5), is even an underestimate of the relative retrotransposition frequency of the pLC5/LAVA$_C$-encoded reporter element. The reduced level of transcripts expressed in pLC21/LAVA$_C$- relative pLC10/LAVA$_C$-transfected cells (Figure 3A) has only a negligible effect on the ~8-fold difference in *trans*-mobilization frequency measured between pLC21/LAVA$_C$- and pLC10/LAVA$_C$-transfected cells (Figure 2B). In order to distinguish the impact of variations in transfection efficiencies on *mneo*I-harbouring reporter transcript levels from the influence of variations in reporter gene transcription rates on these transcript levels, we normalized the spliced *mneo*I RNA levels against RNA amounts expressed from the Hyg^R gene located on the pCEP4 backbone (Figure 1B) of each reporter plasmid that was cotransfected before with a donor plasmid in the retrotransposition reporter assay (Supplementary Figure S6). Transcription rates of pAD3/ SVA$_E$, pLC21/LAVA$_C$ and pLC23/LAVA$_C$ differ from pLC5/LAVA$_C$ by only 45, 36 and 10%, respectively, and are therefore comparable (Supplementary Figure S6B). In contrast, transcription rates of the reporter cassettes of pCEPneo and pLC10/LAVA$_C$ deviate from pLC5/LAVA$_C$ more significantly by 3.1- and 2.7-fold, respectively. While the comparatively high transcription rate of the pCEPneo reporter cassette can be explained by the absence of any retrotransposon sequence (Figure 1A), it is surprising that the full-length LAVA$_C$ in pLC10/LAVA$_C$ that is identical to the LAVA$_C$ element located in the gibbon genome and differs from the pLC5/ LAVA$_C$ reporter exclusively in the presence of the L1ME5 module, is transcribed almost as efficiently as the pCEPneo reporter cassette (Supplementary Figure S6B). These data demonstrate that variations in individual transcription rates specific for each retrotransposition reporter cassette play a role in the observed overall differences (Figure 3A) in spliced *mneo*I-harbouring reporter transcript levels in the variably co-transfected cells. Considering the differences in overall *mneo*I transcript levels between pCEPneo- and the

LAVA$_C$ reporter-transfected cells, as demonstrated in Figure 3A, data indicate that *trans*-mobilization frequencies of the reporter elements encoded by pLC23/LAVA$_C$, pLC21/LAVA$_C$, pLC5/ LAVA$_C$ and pLC10/LAVA$_C$ exceed processed pseudogene formation frequencies in HeLa cells by up to 7-, 46-, 1860-, and 238-fold, respectively.

L1 ORF1p is essential for *trans*-mobilization of LAVA elements

Retrotransposition in *cis* and in *trans* requires reverse transcriptase (RT) activity encoded by L1 ORF2. While it was shown previously that non-autonomous *Alu* elements require only L1 ORF2p for their mobilization in *trans* (Dewannieux et al. 2003), SVA retrotransposition and processed pseudogene formation were demonstrated to require the presence of both L1 ORF1p and L1 ORF2p (Esnault et al. 2000; Hancks et al. 2011; Raiz et al. 2012). In order to investigate if L1 ORF1p is also required for the *trans*-mobilization of LAVA elements, we co-transfected the LAVA reporter elements with the L1 protein donor pJM101/L1$_{RP}$ΔneoΔORF1 (Figure 1C) which differs from the L1 driver pJM101/L1$_{RP}$Δneo exclusively in a 330-bp in-frame deletion in L1 ORF1 causing the expression of a mutant ORF1p that lacks ORF1p amino acid positions 99-208, and measured the resulting *trans*-mobilization frequency (Raiz et al. 2012). Initiation of ORF2 translation within the bicistronic RNA encoded by pJM101/L1$_{RP}$ΔneoΔORF1 is not perturbed. Immunoblot analysis of cell extracts isolated from the differently co-transfected HeLa cells with anti-L1 ORF1p and anti-L1 ORF2p antibodies (Figure 3B) confirmed absence of overexpressed L1 ORF1p and presence of comparable amounts of transiently expressed ORF2p after co-transfection of the L1 protein donor pJM101/L1$_{RP}$ΔneoΔORF1 with each of the four LAVA$_C$ reporter constructs. Co-expression of pJM101/L1$_{RP}$ΔneoΔORF1 (Figure 3B) did not result in *trans*-mobilization of any of the four LAVA$_C$ reporter elements (Figure 2B) although similar amounts of L1 ORF2p are expressed from both protein donor plasmids L1_{RP} and $\text{L1}_{\text{RP}}\Delta$ORF1, indicating that L1ORF1p expression is essential for LAVA *trans* mobilization.

Marked LAVA_{C} *de novo* insertions exhibit hallmarks of L1-mediated Target Primed Reverse Transcription

To confirm that *mneo*I-tagged LAVA_{C} *de novo* retrotransposition events resulted in the observed G418^{R} HeLa colonies, we next analyzed those colonies that followed from co-transfection of the passenger plasmid $\text{pLC10/LAVA}_{\text{C}}$ expressing the full-length LAVA_{C} reporter element and the L1 driver plasmid for the presence of *mneo*I-tagged insertions. To this end, we isolated genomic DNA from 17 randomly chosen single expanded G418^{R} HeLa colonies that arose from this co-transfection experiment. We first performed diagnostic PCR to specifically test for the presence of the functional neo^{R} gene that should result from splicing of the $\text{pLC10/LAVA}_{\text{C}}$-encoded *mneo*I cassette presuming retrotransposition. For this purpose, we used primers specifically binding to the intron-flanking *mneo*I sequences. As shown in Figure 4A, PCR on each of the 17 genomic DNA samples generated a PCR product of ~792 bp indicating the presence of a spliced *mneo*I cassette that resulted from *trans*-mobilization of full-length LAVA_{C} encoded by $\text{pLC10/LAVA}_{\text{C}}$.

Next, we isolated three *mneo*I-tagged LAVA_{C} *de novo* insertions applying EPTS/LM-PCR (Schmidt et al. 2001) to analyze them for the presence of hallmarks of L1-mediated retrotransposition. Sequence analysis of pre- and post-integration sites of these $\text{pLC10/LAVA}_{\text{C}}$-derived *de novo* insertions uncovered that insertions 6 and 14 are full-length and encompass 4,692 bp including 25 bp of CMV promoter sequence, the full length LAVA_{C} element (2,357 bp) and the spliced *mneo*I reporter cassette (2,310 bp) (Figure 4B, Supplementary Table S2). The 5’ ends of both full-length *de novo* insertions are identical and match with position 4 downstream of the transcription initiation site of the CMV promoter

(Raiz et al. 2012) controlling expression of the LAVA$_C$ reporter element. Insertions 6 and 14 occurred into introns 1 and 11 of the genes *NUP50* (Nucleoporin 50kDa) and *TRPC4AP* (Transient Receptor Potential Cation Channel, Subfamily C, Member 4 Associated Protein), respectively. Further, both retrotransposition events inserted in antisense orientation relative to the transcription direction of the host gene, mimicking the arrangement observed for the majority of intragenic LAVA elements in the gibbon genome (Carbone et al. 2014). Both *de novo* insertions exhibit hallmarks of L1-mediated retrotransposition such as 14-nt and 16-nt TSDs, 13-nt and 76-nt poly(A) tails, and genomic target sequences resembling the human L1 EN consensus target sequence 5'-TTTT/AA-3' of pre-existing human L1-Ta, *Alu*Ya5, and SVA$_{E/F}$ insertions (Raiz et al. 2012) and gibbon L1PA and LAVA insertions (Figure 4C). The accordance of the genomic target sequences of both insertions with the human L1 EN consensus target sequence and endogenous gibbon L1 and LAVA target sequences indicates that the LAVA integration sites are determined by L1 EN activity. Also, the untemplated G nucleotide at the 5' end of insertion 14 and the microcomplementarity at the 5' end of insertion 6 are structural features reported previously for L1-mediated retrotransposition events (Zingler et al. 2005; Gilbert et al. 2005; Raiz et al. 2012). The intergenic 3,853 bp insertion 10 occurred into the poly(A) tail of a pre-existing *Alu*Sx1 element, is 5'-truncated and includes only the 1,543 bp of the 3' portion of the full-length LAVA$_C$ element and the subsequent 2,310 bp of the *mneo*I reporter gene (Figure 4B). The extensive genomic poly(A) stretch that is located adjacent to the integration site hampered sequence analysis and identification of the 3' junction sequence.

DISCUSSION

The LAVA element represents an evolutionary novelty in the gibbon lineage. The recent annotation and analysis of the gibbon genome revealed a possible key role for this composite retrotransposon, as LAVA has been found to preferentially insert into introns of genes involved in chromosome segregation processes (Carbone et al. 2014). It has therefore been hypothesized that disruption of these genes through LAVA insertions might have been the trigger of the unusually high rate of chromosomal rearrangements that characterize gibbon species (Carbone et al. 2009). This hypothesis was substantiated by functional studies showing that LAVA insertions can alter host gene transcription by introducing a premature termination site into a given gene (Carbone et al. 2014), similarly to the gene-breaking event described for L1 elements (Han et al. 2004; Wheelan et al. 2005). For this reason it is important to understand how this new retrotransposon has been mobilizing and if functional LAVA copies are still present in the gibbon genome. Alignments of whole-genome sequences from different gibbon species have shown that multiple LAVA insertions are polymorphic ((Carbone et al. 2014) and data not shown), suggesting that the gibbon genome harbours functional LAVA source elements that are currently retrotransposition-competent and mobilized.

Here we show that a gibbon genome-encoded, full-length LAVA element (LAVA$_C$), a member of the older LAVA_C4B subfamily (Supplementary Figure S1), is *trans*-mobilized by transiently expressed human L1 proteins in HeLa cells at frequencies exceeding processed pseudogene formation rates by at least 117-fold. Similar to SVA retrotransposition (Hancks et al. 2011; Raiz et al. 2012) and processed pseudogene formation (Esnault et al. 2000), LAVA mobilization relies on the presence of both L1-encoded proteins, ORF1p and ORF2p. Analysis of genomic pre- and post-integration sites of three isolated *mneo*I-tagged LAVA$_C$ *de novo* insertions show all hallmarks of TPRT-mediated retrotransposition events, including the presence of TSDs, a 3’ poly(A) tail, and an L1 EN-specific consensus target sequence 5’-TTTT/AA-3’ (Figure 4C), all of which were also shown to be characteristic of pre-existing genomic LAVA insertions (Carbone et al. 2012; Carbone et al. 2014) and *de novo* retrotransposition events launched from both human L1 and SVA reporter elements (Gilbert et al. 2002; Symer et al. 2002;Ostertag et al. 2003; Gilbert et al. 2005;Hancks et al. 2011; Raiz et al. 2012). Although the small number of isolated insertions is a limitation of this part of our study, the insertions provide evidence for the *trans*-mobilization of the $\mathrm{LAVA_C}$ reporter element by the human L1-encoded protein machinery. Unfortunately, additional insertions that we attempted to isolate had occurred into highly repetitive sequences that hindered retrieval and thorough characterization of these insertions via PCR and sequence analysis. The integration events we capture with this kind of cell culture assay are merely a reflection of selection, and it is unlikely to identify *de novo* insertions disrupting genes, as cells with these insertions would probably not survive. Our finding that two of the three characterized $\mathrm{LAVA_C}$ *de novo* insertions are full-length and occurred into introns of host genes in antisense orientation is consistent with a previous report in which six out of seven isolated tagged LAVA_F1 *de novo* insertions were shown to be full-length and four of these insertions occurred into introns of host genes (Ianc et al. 2014). Additionally, our results are in line with the observation that ~70% of the endogenous LAVA insertions characterized in the gibbon genome assembly are full-length and 98.5% of the intragenic insertions are intronic (Carbone et al. 2014).

The *trans*-mobilization frequency of full-length $\mathrm{LAVA_C}$ exceeds the processed pseudogene formation rate by ~117-238-fold and exceeds human full-length $\mathrm{SVA_E}$ retrotransposition frequencies by ~3-3.6-fold. In fact, this $\mathrm{LAVA_C}$ *trans*-mobilization frequency is within the range of the human *Alu* retrotransposition frequency that was reported to be 100-1000-fold higher than control mRNAs (Dewannieux et al. 2003). Considering that

$LAVA_C$ is a member of one of the oldest LAVA sub-families aged 16-18 my, this is a noteworthy result indicating that not only members of the youngest sub-families, LAVA_E (10-11 mya) and LAVA_F (6-9 mya) are preferred substrates of the human L1 protein machinery (Ianc et al. 2014). Clearly, retrotransposition reporter assays in a homologous system in which the reporter plasmids are co-transfected with driver plasmids expressing functional gibbon-encoded LINE-1 retrotransposons into cultured gibbon cells would allow to draw more reliable conclusions on the *in vivo* retrotransposition activity of LAVA elements in the gibbon genome. For instance, it was reported recently that in a similar assay established to test transposition activity of the DNA transposon *piggyBat*, the number of *de novo* transposition events that occurred in HeLa cells, was much higher than those identified in bat fibroblasts (98,816 vs 4,264) (Mitra et al. 2013). However, since, to our knowledge, there is neither any applicable gibbon cell culture system nor any isolated functional gibbon LINE-1 retrotransposon available to date, we decided to assess the $LAVA_C$ retrotransposition activity in the well-established human cell culture system. One reason for these experimental limitations is that many gibbon species are critically endangered making biological materials extremely difficult to come by. Undoubtedly, we can neither suggest the actual *in vivo* LAVA retrotransposition frequency based on the LAVA *trans*-mobilization frequencies observed in the human cell culture system, nor compare $LAVA_C$ retrotransposition frequencies with the recently reported *trans*-mobilization rates of LAVA_E and LAVA_F1 members (Ianc et al. 2014), due to the overall variability of the applied cell culture based assays between different laboratories. Nevertheless, it is surprising that the copy numbers of the LAVA subfamilies C, E and F constitute only 404, 350 and 773 members, respectively. Therefore, we speculate that the number of functional and transcribed LAVA_C master copies is likely to be relatively small in the modern genome, otherwise we would observe much higher copy numbers of this subfamily.

The modular structure of the LAVA element allowed testing of different combinations of its portions in our *trans*-mobilization assay to identify which modules make LAVA RNAs preferred substrates for L1 proteins. We found that the SVA-derived module containing the CT-rich, *Alu*-like, and VNTR regions provides the competence to retrotranspose at rates beyond that of processed pseudogene formation. Recent studies investigating the structural features of SVA elements that affect their *trans*-mobilization frequency (Hancks et al. 2012) demonstrated that removal of the *Alu*-like region in the context of a full-length SVA has little to no effect, whereas removal of the CT-hexamer or the VNTR region can result in a 75% decrease in activity. This is consistent with our observation that removal of the SVA-specific fragment covering CT-rich region, *Alu*-like region, and VNTR module from the full-length LAVA$_{C}$ reporter element (pLC10/LAVA$_{C}$) reduces *trans*-mobilization frequency by ~98% (Figure 2B). The VNTR region is indeed likely to be crucial for both SVA and LAVA mobilization given that SVA2 elements found in rhesus macaque (Gibbs et al. 2007; Han et al. 2007a) consists exclusively of VNTRs and show signatures of L1-mediated retrotransposition (Damert et al. 2009).

Interestingly, the LAVA$_{C}$ reporters exhibiting the highest *trans*-mobilization frequencies include both the VNTR and *Alu*Sz modules. Removal of the VNTR-U1-*Alu*Sz-U2 module results in reduction of the retrotransposition frequency by at least ~15-fold (Figure 2B and Supplementary Figure S5). It is plausible that this effect is caused by the absence of the VNTR region that has recently been demonstrated to be crucial for LAVA_F1 mobilization (Lupan et al. 2015). However, we cannot rule out the possibility that the *Alu*Sz module may also play a role in LAVA *trans*-mobilization as it has been shown that the presence of an *Alu*Sp module at the 3’ end of an SVA$_{F1}$ source element increases its *trans*-mobilization frequency by ~25-fold (Raiz et al. 2012). Our data also indicate that the presence of the 211-bp L1ME5 module reduces mobilization frequency of the full-length

$LAVA_C$ element by ~2-8-fold. It remains to be investigated, which function the *trans*-mobilization-attenuating L1ME5 module may have and why it was stably retained as part of functional endogenous LAVA elements during evolution, despite its apparent ability to attenuate LAVA mobilization rates. An exhaustive search of the gibbon genome retrieved only seven LAVA elements that lack exclusively the L1ME5 module. However, all these insertions are missing poly(A) tails suggesting that they derived from polyadenylated full-length LAVA transcripts that were 3'-truncated upon insertion by a mechanism different from TPRT.

The 3'-end portion of $LAVA_C$ comprising the modules U1, *Alu*Sz, U2, and L1ME5, is *trans*-mobilized at frequencies exceeding processed pseudogene formation rate by only ~2-7-fold. This is consistent with our comparative data showing that only one copy of this sequence was present in the anthropoid common ancestor genome. Since this sequence was present without apparent mobility for a very long time in primate genomes, it is obviously not a preferred substrate for *trans*-mobilization by L1 proteins. In the gibbon genome, this sequence is found at position chr1a:39006175-39006659, and we refer to it as '*LAVA 3' Progenitor Locus Allele A*'. We speculate that this locus is the original source of the LAVA 3' sequence and propose two possible models for how this sequence was united with the SVA-derived module we find at the 5' end of the full-length LAVA (Figure 5).

Model A is based on the fortuitous retrotransposition of the SVA2- or SVA_A-derived sequences precisely to the 5' end of the U1 component at LAVA 3' Progenitor Locus Allele A (Figure 5A). The inserting sequence could have been donated either from a 3'-truncated hypothetical precursor element comprised of an SVA2 united with CT-rich and *Alu*-like sequences (Pathway 1) (Ianc et al. 2014) or a 3'-truncated SVA_A element (Pathway 2) (Damert et al. 2009). Insertions of either type could have resulted in the full-length LAVA element at the chromosome 1a locus (Allele B) and this could have served as the first functional LAVA master copy.

Model B is based on the fact that the LAVA 3’ Progenitor Locus is found within Intron 2 of the gene *HSD17B3* and that the two nucleotides immediately adjacent to the 5’ end of the U1 component at this locus are 5’-AT-3’. We hypothesize that this 5’-AT-3’ dinucleotide may have mutated at some point to 5’-AG-3’, resulting in a cryptic splice acceptor site. A subsequent *trans*-splicing event could then have occurred between this splice acceptor site within an *HSD17B3* transcript and a splice donor site in the 3’ end of the VNTR region of an SVA_A mRNA molecule, resulting in an mRNA molecule which, when reverse transcribed and integrated into the genome, produced the first LAVA master element (Figure 5B). We investigated whether a *cis*-splicing event could explain the origin of LAVA as well. However, this would require a copy of SVA upstream of the LAVA 3’ Progenitor Locus on chromosome 1a and in the same orientation. Only one such SVA copy exists upstream on this chromosome (at position chr1a: 26376513-26377921), and its VNTR sequences do not match those in the LAVA element. Additionally, this SVA copy is located ~12.6 Mb upstream of the LAVA 3’ Progenitor Locus, a distance making *cis*-splicing with this locus unlikely.

In the current gibbon genome assembly (Nleu3.0), we do not find evidence of a full-length LAVA at the LAVA 3’ Progenitor Locus Allele B on chromosome 1a (Model A) nor do we find the cryptic splice acceptor sequence (Model B) (Figure 5). This suggests that under either model these variants were lost to genetic drift over time, but only after having given rise to other master copies in the genome from which the LAVA family could propagate. Since the probability of any new mutation in a population of diploid organisms being lost to genetic drift is $1 - (1/2N_e)$ in which N_e is the effective population size, reversion to the ancestral sequence is a highly likely event for any new mutation (Kimura 1955). However, we cannot discount the possibility that one of the alleles that gave rise to the first master copy is present in the genome, but is obfuscated in the current assembly which, like all draft genomes, contains gaps and assembly errors.

In conclusion, we have demonstrated the ability of a full-length member of the old LAVA_C subfamily of gibbon retrotransposons to mobilize in HeLa-HA cells by recruiting the human L1 protein machinery. Our results indicate that the 5' SVA-derived module is essential for efficient retrotransposition, likely due to the presence of the VNTR region. The *Alu*Sz module may also confer some activity, while the L1ME5 module appears to reduce retrotransposition efficiency. Analysis of the sequence architecture of *de novo* LAVA$_C$ insertions supports the conclusion that LAVA is a non-autonomous family that relies upon L1-mediated TPRT as its mechanism of propagation. Considering the recently reported *trans*-mobilization of members of the LAVA_E and LAVA_F1 subfamilies (Ianc et al. 2014), LAVA$_C$ represents a member of the evolutionary older LAVA_C subfamily that is the third LAVA element demonstrated to be a preferred substrate for *trans*-mobilization by the L1 protein machinery.

ACKNOWLEDGEMENTS

We thank John Moran for providing plasmids pJM101/L1.3, pJM101/L1$_{RP,}$ and pJM101/L1$_{RP}$Δneo and the cell lines HeLa-HA. We are grateful to John Goodier and Haig Kazazian who provided us with the reliable αL1ORF2p-N antibody. We are indebted to Kay-Martin Hanschmann for statistical analyses.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft [DA 545/2-1 and SCHU 1014/8-1 to G.G.S.] and the National Institute of Health [NCRR P51, RR000163 to LC]. Moreover, T.J.M. was supported by the National Library of Medicine of the National

Institutes of Health under Award Number T15LM007088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Andrews PW, et al. 1984. Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2. Differentiation in vivo and in vitro. Lab Invest 50:147-162.

Belancio VP, Deininger PL, Roy-Engel AM. 2009. LINE dancing in the human genome: transposable elements and disease. Genome Med 1:97.

Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18:343-358.

Belancio VP, Roy-Engel AM, Deininger PL. 2010. All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 20:200-210.

Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. Genome Dyn 1:104-115.

Carbone L, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature 513:195-201.

Carbone L, et al. 2012. Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. Genome Biol Evol 4:648-658.

Carbone L, et al. 2009. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genet 5:e1000538.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet 10:691-703.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-1190.

Dai L, LaCava J, Taylor MS, Boeke JD. 2014. Expression and detection of LINE-1 ORF-encoded proteins. Mob Genet Elements 4:e29319.

Damert A, et al. 2009. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res 19:1992-2008.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7:e1002384.

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

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

Freeman JD, Goodchild NL, Mager DL. 1994. A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques 17:46, 48-49, 52.

Gibbs RA, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Gilbert N, Lutz S, Morrish TA, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 25:7780-7795.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110:315-325.

Goodier JL, Ostertag EM, Engleka KA, Seleme MC, Kazazian HH, Jr. 2004. A potential role for the nucleolus in L1 retrotransposition. Hum Mol Genet 13:1041-1048.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429:268-274.

Han K, et al. 2007a. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science 316:238-240.

Han K, et al. 2008. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A 105:19366-19371.

Han K, et al. 2007b. Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet 3:1939-1949.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH, Jr. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum Mol Genet 20:3386-3400.

Hancks DC, Kazazian HH, Jr. 2012. Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22:191-203.

Hancks DC, Mandal PK, Cheung LE, Kazazian HH, Jr. 2012. The minimal active human SVA retrotransposon requires only the 5'-hexamer and Alu-like domains. Mol Cell Biol 32:4718-4726.

Heidmann T, Heidmann O, Nicolas JF. 1988. An indicator gene to demonstrate intracellular transposition of defective retroviruses. Proc Natl Acad Sci U S A 85:2219-2223.

Hulme AE, Bogerd HP, Cullen BR, Moran JV. 2007. Selective inhibition of Alu retrotransposition by APOBEC3G. Gene 390:199-205.

Ianc B, Ochis C, Persch R, Popescu O, Damert A. 2014. Hominoid composite non-LTR retrotransposons-variety, assembly, evolution, and structural determinants of mobilization. Mol Biol Evol 31:2847-2864.

Isomura H, et al. 2008. 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.

Kaer K, Speek M. 2013. Retroelements in human disease. Gene 518:231-241.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649.

Kimura M. 1955. Solution of a Process of Random Genetic Drift with a Continuous Model. Proc Natl Acad Sci U S A 41:144-150.

Kirilyuk A, et al. 2008. Functional endogenous LINE-1 retrotransposons are expressed and mobilized in rat chloroleukemia cells. Nucleic Acids Res 36:648-665.

Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. Semin Cancer Biol 20:211-221.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lee J, Han K, Meyer TJ, Kim HS, Batzer MA. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS One 3:e4047.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25:402-408.

Locke DP, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469:529-533.

Lupan I, Bulzu P, Popescu O, Damert A. 2015. Lineage specific evolution of the VNTR composite retrotransposon central domain and its role in retrotransposition of gibbon LAVA elements. BMC Genomics 16:389.

Mitra R, et al. 2013. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc Natl Acad Sci U S A 110:234-239.

Moran JV, et al. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87:917-927.

Muller S, Hollatz M, Wienberg J. 2003. Chromosomal phylogeny and evolution of gibbons (Hylobatidae). Hum Genet 113:493-501.

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73:1444-1451.

Raiz J, et al. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res 40:1666-1683.

Schmidt M, et al. 2001. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther 12:743-749.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18:6097-6100.

Sievers F, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

Smit A, Hubley R, Green P. unpublished. RepeatMasker. Current Version: open-4.0.5.

Symer DE, et al. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110:327-338.

Szak ST, et al. 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol 3:research0052.

Veeramah KR, et al. 2015. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. Genetics 200:295-308.

Wei W, et al. 2001. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21:1429-1439.

Wheelan SJ, Aizawa Y, Han JS, Boeke JD. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res 15:1073-1078.

Zingler N, et al. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. Genome Res 15:780-789.

FIGURE LEGENDS

Figure 1: Structures of LAVA$_C$ retrotransposition reporter plasmids and rationale of the LAVA *trans*-mobilization assay. (A) Organization of the non-autonomous non-LTR retrotransposon LAVA in the gibbon genome. The SVA-derived module consists of CT-rich (green), *Alu*-like (light blue), and VNTR (banded) region. It is separated from the *Alu*Sz-derived module (black) by Unique Region 1 (U1). The L1ME5-derived module (pink) is separated from the *Alu*Sz module by Unique Region 2 (U2). Black arrows in boxed *Alu*-like region and L1ME5 module indicate antisense orientation. Poly(A), poly(A) stretch; TSD, target site duplication. **(B)** Schematics of the retrotransposition reporter plasmid pLC10/LAVA$_C$ carrying the full-length LAVA$_C$ element, the expression cassettes of the reporter plasmids pLC5/LAVA$_C$, pLC21/LAVA$_C$, pLC23/LAVA$_C$ expressing truncated LAVA$_C$ versions, and pAD3/SVA$_E$ (Raiz et al. 2012). pCEPneo is used to measure processed pseudogene formation frequency. Each of the LAVA and SVA reporter elements and the processed pseudogene formation cassette were tagged with the indicator gene *mneo*I, and set under transcriptional control of the human CMV immediate early enhancer/promoter (CMV$_P$). Splice donor (SD) and splice acceptor (SA) sites of the oppositely oriented γ-globin intron are indicated. *mneo*I is flanked by an SV40 promoter (P') and polyadenylation signal (A').Transcripts starting from CMV$_P$ driving LAVA *mneo*I, SVA *mneo*I or pCEP *mneo*I transcription, can splice the intron, but contain an antisense copy of the *neo*^{R} gene. G418 resistant (G418^R) colonies accrue only if this transcript is reverse transcribed, integrated into chromosomal DNA, and expressed from its own promoter P'. LAVA or SVA sequences were inserted between CMV$_P$ and the *mneo*I cassette. pLC5/LAVA$_C$ differs from pLC10/LAVA$_C$ exclusively in the absence of the L1ME5 and U2 modules covering 322 bp. pLC21/LAVA$_C$ and pLC23/LAVA$_C$ encode the 5'-terminal 819 bp of LAVA$_C$ covering TSD, CT-rich and *Alu*-like region, and the 3'-terminal 473 bp covering U1, *Alu*Sz and L1ME5 modules,

respectively. pAD3/SVA$_E$ (Raiz et al. 2012) serves as positive control construct for *trans* mobilization. Transcriptional termination signals at the 3' ends of the L1ME5 (pLC10/LAVA$_C$, pLC23/LAVA$_C$) and SINE-R (pAD3/SVA$_E$) modules were deleted from the LAVA and SVA reporter cassettes (ΔAATAAA) to ensure transcriptional read-through into the *mneo*I cassette and polyadenylation at the pCEP4-encoded SV40 polyadenylation signal (pA). pCEPneo is distinguished from the remaining presented reporter constructs by the absence of any LAVA or SVA sequence. CMV$_P$ sequences are highlighted in grey. CMV$_P$ major and minor transcription start sites (Isomura et al. 2008) are indicated by arrows. TSD, target site duplication; CT-rich, *Alu*-like, VNTR (Variable number of tandem repeats), U1, *Alu*Sz, U2 and L1ME5 represent repeat modules, LAVA is composed of; pA, poly(A) tail. Hyg^R, hygromycin resistance gene serving as selectable marker for eukaryotic cells. **(C)** Design of the experimental approach to test for *trans*-mobilization of the *mneo*I-tagged LAVA$_C$ element by the human L1 protein machinery. LAVA or SVA retrotransposition reporter plasmids, or pCEPneo were each co-transfected with L1 protein donor plasmid pJM101/L1$_{RP}$Δneo or pJM101/L1$_{RP}$ΔneoΔORF1 (blue) into HeLa-HA cells that were subsequently selected for hygromycin resistance for 12 days. Hyg^R cells were assayed for retrotransposition events by selecting for 9-12 days for G418^R HeLa colonies.

Figure 2: *Trans*-mobilization of *mneo*I-tagged LAVA$_C$ reporter elements by the human L1 protein machinery requires the presence of L1 ORF1p. (A) Immunoblot analysis of L1 protein expression in HeLa-HA cells after co-transfection of the L1 protein donor plasmid pJM101/L1$_{RP}$Δneo (L1$_{RP}$) with the LAVA$_C$, SVA$_E$ or pCEPneo reporter plasmids. Cell lysates were isolated 13 days after co-transfection upon completion of hyg^R selection and subjected to immunoblot analysis using anti-L1 ORF1p antibodies (αORF1p). 20 µg of cell lysates from the differently transfected cells were loaded per lane. Only 2 µg of NTERA-2

cell lysates were separated on one lane of the gel. β-actin protein levels (~42 kDa) served as loading control. Lysates from pCEP4-transfected HeLa cells (HeLa + pCEP4) and from the human embryonal carcinoma cell line NTERA-2 served as negative and positive control for L1 ORF1p detection, respectively. **(B)** LAVA$_C$ retrotransposition reporter assay after hyg^R selection for the presence of expression plasmids. LAVA$_C$ **r**eporter plasmids pLC5/LAVA$_C$, pLC10/LAVA$_C$, pLC21/LAVA$_C$, pLC23/LAVA$_C$, were co-transfected with the L1 protein donors pJM101/L1$_{RP}$Δneo (L1$_{RP}$) or pJM101/L1$_{RP}$ΔneoΔORF (ΔORF) or the empty vector pCEP4. Control constructs pAD3/SVA$_E$ and pCEPneo were co-transfected with either pJM101/L1$_{RP}$Δneo or pCEP4. After hyg^R selection, G418^R selection for retrotransposition events followed and retrotransposition rates were determined by counting G418^R HeLa colonies. To quantify retrotransposition frequencies, each co-transfection was performed in biological triplicates with each biological replicate being executed on a separate day. Each biological replicate was conducted in technical triplicates. Absolute retrotransposition frequencies per 10^6 cells are listed and relative retrotransposition rates are presented as bar diagrams. *Cis* retrotransposition frequency of the L1 reporter element encoded by pJM101/L1$_{RP}$ after its co-transfection with pCEP4 was set as 100%. Each bar depicts the arithmetic mean ± SD of the relative retrotransposition frequencies that resulted from nine individual co-transfection experiments (n=9).

Figure 3: Analysis of the expression from L1 protein donor and retrotransposition reporter plasmids. (A) qRT-PCR analyses to quantify the relative amounts of spliced transcripts expressed from the different retrotransposition reporter cassettes. Total RNA was isolated after 12 days of hygromycin selection following co-transfection of pCEPneo, pAD3/SVA$_E$, pLC5/LAVA$_C$, pLC10/LAVA$_C$, pLC21/LAVA$_C$, or pLC23/LAVA$_C$ with the L1 protein donor plasmid pJM101/L1$_{RP}$Δneo and co-transfection of pJM101/L1$_{RP}$ with

pCEP4. Each co-transfection was performed in three biological replicates. The used primer/probe combination (see Materials and Methods) is specific for the spliced *mneo*I-cassette (black box with arrow). Real-time PCR of each biological replicate was conducted in technical triplicates. Relative amounts of mRNA expression refer to the signal obtained from total RNA from pCEP4(mock)-transfected HeLa cells which was set as 1 (pCEP4) and served as negative control. Bars depict arithmetic means ±SD of technical triplicates of three biological replicates. **(B)** Immunoblot analysis of L1 ORF1p (left panel) and L1 ORF2p (right panel) expression after co-transfection of the L1 protein donor plasmid pJM101/L1$_{RP}$ΔneoΔORF1 (L1$_{RP}$ΔORF1) with the LAVA$_{C}$ retrotransposition reporter plasmids. Whole-cell lysates were prepared 13 days after co-transfection upon completion of hygromycin selection and subjected to immunoblot analysis using antibodies against either L1 ORF1p (αORF1p) or L1 ORF2p (αORF2p). In the case of the anti-L1 ORF1p immunoblot analysis (left panel), 20 µg of the cell lysate isolated from each of the differently transfected HeLa-HA cell cultures, were loaded per lane on a 12% PAA gel. Since ORF2p (Predicted MW~150 kDa) is expressed at a significantly lower level than ORF1p (Dai et al. 2014), 40 µg of total cell lysate from each of the differently transfected HeLa-HA cells were loaded per lane on a 6% PAA gel to perform anti-L1 ORF2p immunoblot analysis (right panel). Detectable amounts of intact L1 ORF1-encoded proteins are absent from pJM101/L1$_{RP}$ΔneoΔORF1-transfected HeLa cells. 2 µg of the NTERA-2 cell extract were loaded on one lane of each of the two presented gels. Lysates from pCEP4-transfected HeLa cells (HeLa + pCEP4) and from NTERA-2 cells served as negative and positive control for L1 protein detection, respectively. β-actin protein levels (~42 kDa) were analyzed as loading control.

Figure 4: Marked LAVA$_C$ *de novo* insertions have typical attributes of L1-mediated retrotransposition events. (A) Diagnostic PCR to demonstrate splicing of the artificial γ-globin intron from the *mneo*I indicator cassette of marked LAVA$_C$ *de novo* insertions. Genomic DNA was extracted from 17 G418^R colonies (gDNA samples 1-15 and 17-18) that resulted from co-transfection of the LAVA reporter pLC10/LAVA$_C$ and the L1 driver construct pJM101/L1$_{RP}$Δneo, and used as template for PCR using *mneo*I cassette-specific primers GS86 and GS87 (Supplementary Table S1). Presence of the spliced and reverse-transcribed *mneo*I cassette resulting from retrotransposition gives rise to a 792-bp PCR product, while the unspliced reporter cassette which did not go through retrotransposition leads to a 1694-bp PCR product. PCR on gDNA sample #12 resulted in an additional product of ~700 bp whose origin is unclear. PCR performed on pcDNA™ 3.1(+) (Invitrogen) served as positive control for the presence of a *neo^R* gene lacking any intron (lane A). PCR analysis using pLC10/LAVA$_C$ as template resulted in a 1694 bp product specific for the unspliced *mneo*I cassette (lane B). M, 1 kb DNA ladder (New England Biolabs #N3232). **(B)** Structural organization of *mneo*I-tagged LAVA$_C$ *de novo* insertions. Both pre- and post-integration sites of *de novo* insertions 6, 10 and 14 are presented. Chromosomal nucleotide positions given for each locus indicate the sites into which the LAVA$_C$ elements inserted and refer to human reference genome hg19. Marked full-length insertions 6 and 14 cover 4692 bp and 5'-truncated insertion 10 comprises 3853 bp. The extra deoxyguanylate at the 5'-end of insertion 14 is indicated in green. The L1 EN target sequence on the bottom strand is presented in blue with the arrow indicating the L1 EN nicking site. CMV$_P$-derived sequences are highlighted in yellow. Transcriptional directions of host genes *NUP50* and *TRPC4AP* in which insertions 6 and 14 integrated are denoted with an arrow. TSDs could not be identified for insertion 10 because sequencing of the LAVA 3' junction was hampered by the directly adjacent poly(A) tail which is part of a pre-existing genomic *Alu*Sx1 element. Red lettering, TSD sequences;

*mneo*I, neomycin-phosphotransferase gene. **(C)** The nucleotide profile of LAVA insertion sites resembles the consensus target sequence of gibbon L1PA and human L1-Ta insertions. Target sequence logos were generated by multiple sequence alignments of genomic integration sites of LAVA insertions (top panel), gibbon L1PA (middle panel), and human L1-Ta (bottom panel) using the program WebLogo (Crooks et al. 2004). Logos for the top strand sequence cover two nucleotides of upstream and four nucleotides of downstream sequence relative to the L1 EN cleavage site on the bottom strand. Numbers on the x-axis denote nucleotide positions relative to the L1 EN nicking site. Target sequence logos were generated from genomic integration sites of 34 preexisting LAVA insertions (Carbone et al. 2012), target sequences of randomly selected L1PA insertions from the *Nomascus leucogenys* reference genome (n = 91), and target sequences of 70 preexisting L1-Ta insertions (Raiz et al. 2012). Integration site sequences were identified as described in the Materials and Methods section.

Figure 5: Alternative models for the assembly process resulting in functional LAVA master copies. The putative LAVA 3' Progenitor Locus Allele A consisting of the U1, *Alu*Sz, U2, and L1ME5 modules was present in the anthropoid common ancestor and is still present at orthologous loci in the human (chr9:99026506-99026986 in hg19) and gibbon (chr1a:39006175-39006659 in Nleu3.0) genomes, among others. **(A)** Model A entails the precise insertion of the CT-rich, *Alu*-like, and VNTR regions upstream of LAVA 3' Progenitor Locus Allele A. The inserting sequence may have been derived from the union of an SVA2 sequence with CT-rich and *Alu*-like region (Pathway 1; Intermediates I and II, within the dashed box, are speculative) or from a 3'-truncated SVA_A element (Pathway 2). Since the master copy of the LAVA 3' Progenitor Locus Allele B cannot be found at the 3' Progenitor Locus on gibbon chromosome 1a, we speculate that this allele persisted only long

enough to give rise to other master copies and was subsequently lost to genetic drift, leaving only the original allele (Allele A), which lacks the SVA-derived modules. **(B)** In Model B, a *trans*-splicing event gave rise to the first LAVA master locus. We hypothesize that a *trans*-splicing event occurred between a cryptic splice donor (SD) site within the 3’ end of an SVA_A mRNA (pre-mRNA 1) and a splice acceptor (SA) site immediately upstream of the chromosome 1a LAVA 3’ Progenitor Locus in the gibbon common ancestor. The mRNA from the LAVA 3’ Progenitor Locus is provided as the consequence of transcription of the gene *HSD17B3*, in which the LAVA 3’ Progenitor Locus is located in Intron 2 in transcriptional orientation of the gene. In all analyzed gibbon genomes, including Nleu3.0, the SA site adjacent to the LAVA 3’ Progenitor Locus does not have the necessary AG nucleotide sequence, and instead has an AT sequence. We speculate that a T-to-G point mutation at this position in an ancestral gibbon individual produced the necessary SA sequence. After giving rise to the first LAVA master element elsewhere in the genome as consequence of the *trans*-splicing event, this variant allele at the LAVA 3’ Progenitor Locus was subsequently lost to genetic drift.

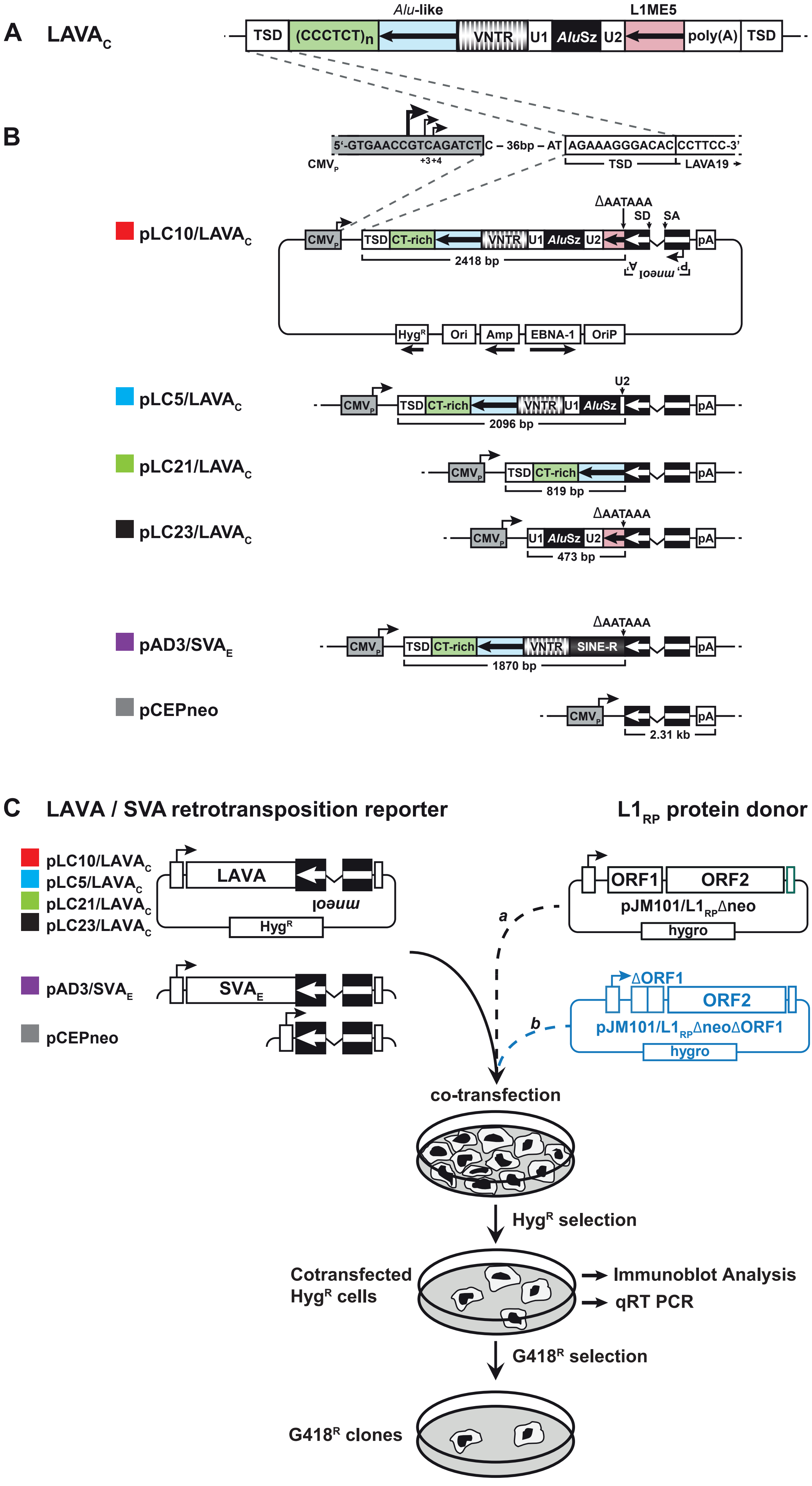

A
LAVA$_C$
Alu-like
L1ME5
TSD
(CCCTCT)n
VNTR
U1
AluSz
U2
poly(A)
TSD
B
5'-GTGAACCGTCAGATCT C – 36bp – AT AGAAAGGGACAC CCTTCC-3'
CMV$_P$
+3+4
TSD
LAVA19
pLC10/LAVA$_C$
ΔAATAAA
SD
SA
CMV$_P$
TSD
CT-rich
VNTR
U1
AluSz
U2
pA
2418 bp
P', mneoI A'
HygR
Ori
Amp
EBNA-1
OriP
pLC5/LAVA$_C$
U2
2096 bp
pLC21/LAVA$_C$
819 bp
pLC23/LAVA$_C$
ΔAATAAA
473 bp
pAD3/SVA$_E$
ΔAATAAA
SINE-R
1870 bp
pCEPneo
2.31 kb
C
LAVA / SVA retrotransposition reporter
L1$_{RP}$ protein donor
pLC10/LAVA$_C$
pLC5/LAVA$_C$
pLC21/LAVA$_C$
pLC23/LAVA$_C$
LAVA
mneoI
HygR
pAD3/SVA$_E$
SVA$_E$
pCEPneo
ORF1
ORF2
pJM101/L1$_{RP}$Δneo
hygro
a
ΔORF1
ORF2
pJM101/L1$_{RP}$ΔneoΔORF1
hygro
b
co-transfection
HygR selection
Cotransfected HygR cells
Immunoblot Analysis
qRT PCR
G418R selection
G418R clones

Figure 1

Figure 2

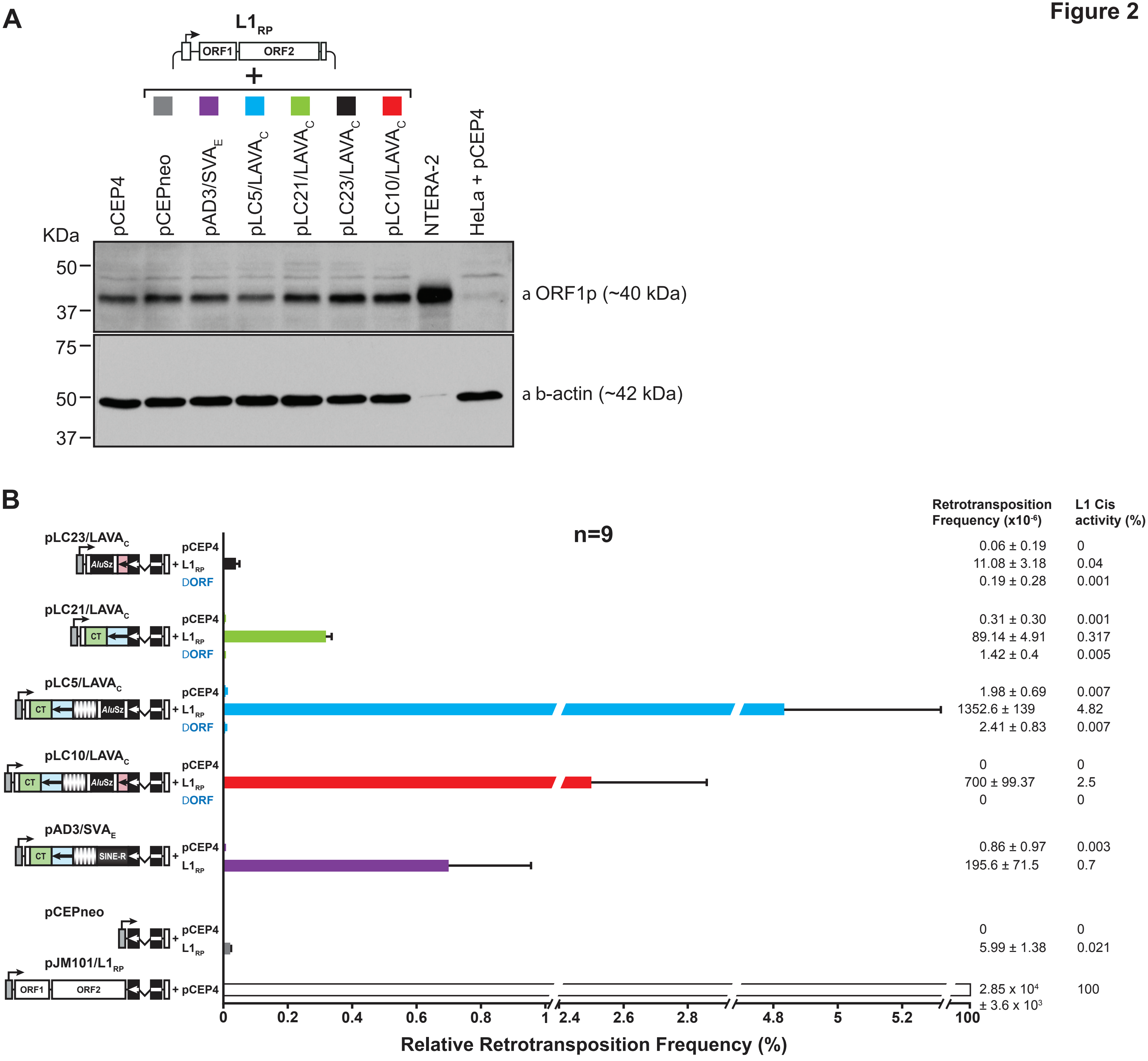
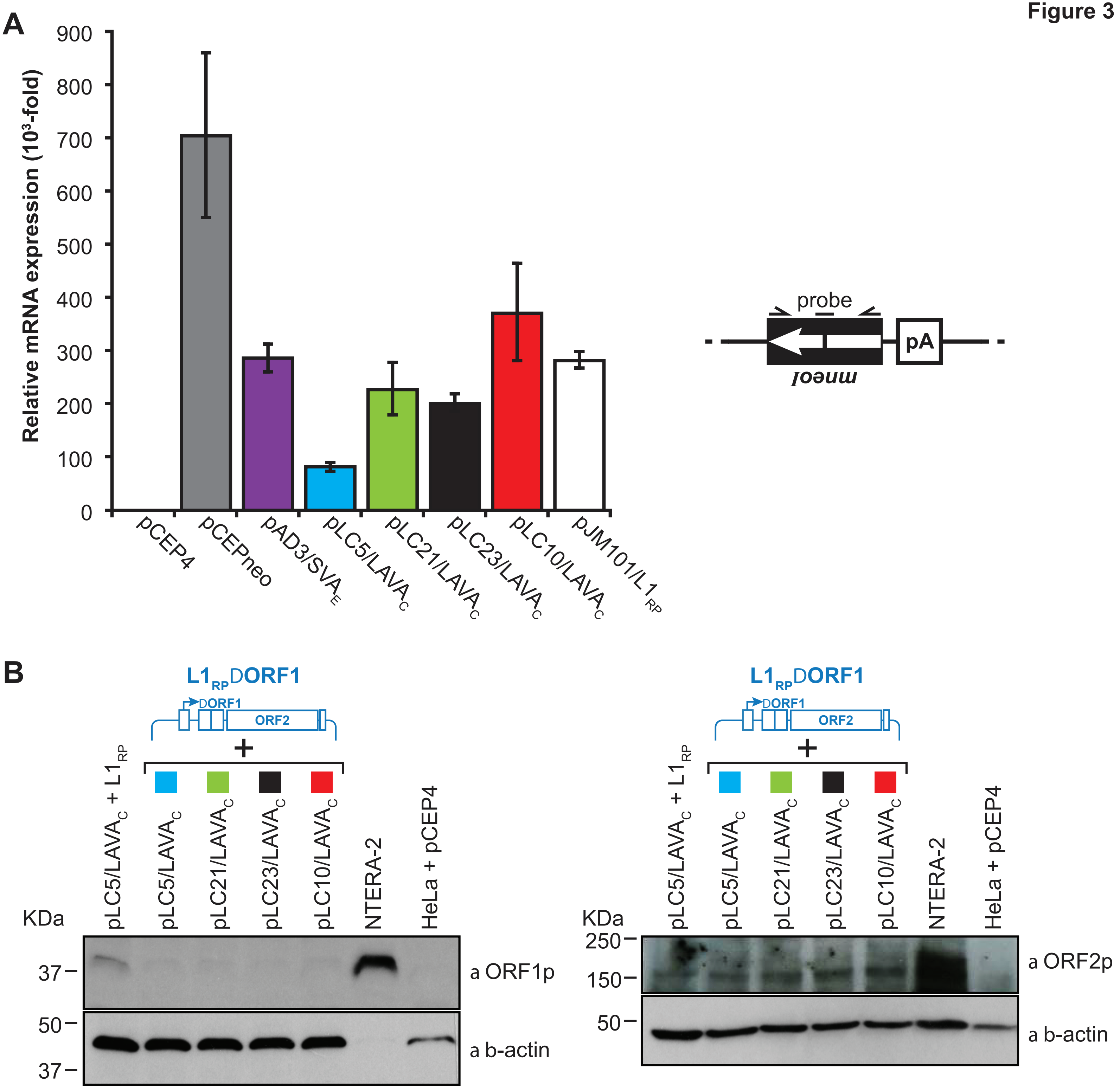

A
L1RP
ORF1
ORF2
+
pCEP4
pCEPneo
pAD3/SVAE
pLC5/LAVAC
pLC21/LAVAC
pLC23/LAVAC
pLC10/LAVAC
NTERA-2
HeLa + pCEP4
KDa
50
37
75
50
37
a ORF1p (~40 kDa)
a b-actin (~42 kDa)
B
n=9
Retrotransposition Frequency (x10-6)
L1 Cis activity (%)
pLC23/LAVAC
AluSz
pCEP4
+ L1RP
DORF
0.06 ± 0.19 0
11.08 ± 3.18 0.04
0.19 ± 0.28 0.001
pLC21/LAVAC
CT
pCEP4
+ L1RP
DORF
0.31 ± 0.30 0.001
89.14 ± 4.91 0.317
1.42 ± 0.4 0.005
pLC5/LAVAC
CT
AluSz
pCEP4
+ L1RP
DORF
1.98 ± 0.69 0.007
1352.6 ± 139 4.82
2.41 ± 0.83 0.007
pLC10/LAVAC
CT
AluSz
pCEP4
+ L1RP
DORF
0 0
700 ± 99.37 2.5
0 0
pAD3/SVAE
CT
SINE-R
pCEP4
+ L1RP
0.86 ± 0.97 0.003
195.6 ± 71.5 0.7
pCEPneo
pCEP4
+ L1RP
0 0
5.99 ± 1.38 0.021
pJM101/L1RP
ORF1
ORF2
+ pCEP4
2.85 x 10^4 ± 3.6 x 10^3 100
0
0.2
0.4
0.6
0.8
1
2.4
2.6
2.8
4.8
5
5.2
100
Relative Retrotransposition Frequency (%)

Figure 3



A
Relative mRNA expression (10³-fold)
900
800
700
600
500
400
300
200
100
0
pCEP4
pCEPneo
pAD3/SVA_E
pLC5/LAVA_C
pLC21/LAVA_C
pLC23/LAVA_C
pLC10/LAVA_C
pJM101/L1_RP
probe
pA
mneoI
B
L1_RP DORF1
DORF1
ORF2
+
pLC5/LAVA_C + L1_RP
pLC5/LAVA_C
pLC21/LAVA_C
pLC23/LAVA_C
pLC10/LAVA_C
NTERA-2
HeLa + pCEP4
KDa
37
50
37
a ORF1p
a b-actin
L1_RP DORF1
DORF1
ORF2
+
pLC5/LAVA_C + L1_RP
pLC5/LAVA_C
pLC21/LAVA_C
pLC23/LAVA_C
pLC10/LAVA_C
NTERA-2
HeLa + pCEP4
KDa
250
150
50
a ORF2p
a b-actin

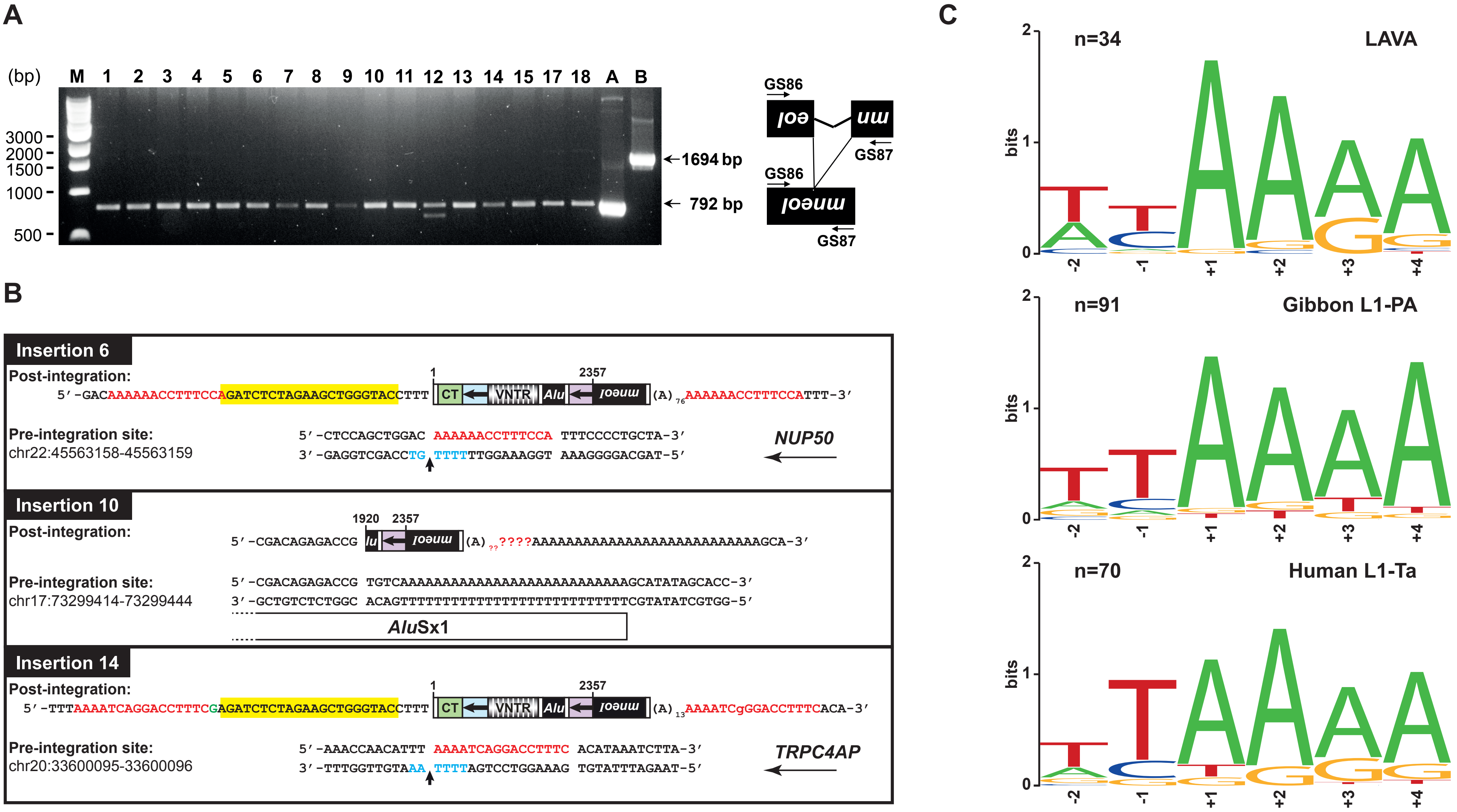

A
(bp)
M
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
18
A
B
3000
2000
1500
1000
500
1694 bp
792 bp
GS86
eol
mn
GS87
GS86
mneol
GS87
B
Insertion 6
Post-integration:
1
2357
5′-GACAAAAAAACCTTTCCAGATCTCTAGAAGCTGGGTACCTTT
CT
VNTR
Alu
mneol
(A)76AAAAAACCTTTCCATTT-3′
Pre-integration site:
chr22:45563158-45563159
5′-CTCCAGCTGGAC AAAAAACCTTTCCA TTTCCCCTGCTA-3′
3′-GAGGTCGACCTG TTTTTTGGAAAGGT AAAGGGGACGAT-5′
NUP50
Insertion 10
Post-integration:
1920
2357
5′-CGACAGAGACCG
lu
mneol
(A)??????AAAAAAAAAAAAAAAAAAAAAAAAAAGCA-3′
Pre-integration site:
chr17:73299414-73299444
5′-CGACAGAGACCG TGTCAAAAAAAAAAAAAAAAAAAAAAAAAAGCATATAGCACC-3′
3′-GCTGTCTCTGGC ACAGTTTTTTTTTTTTTTTTTTTTTTTTTTCGTATATCGTGG-5′
AluSx1
Insertion 14
Post-integration:
1
2357
5′-TTTAAAATCAGGACCTTTCGAGATCTCTAGAAGCTGGGTACCTTT
CT
VNTR
Alu
mneol
(A)13AAAATCgGGACCTTTCACA-3′
Pre-integration site:
chr20:33600095-33600096
5′-AAACCAACATTT AAAATCAGGACCTTTC ACATAAATCTTA-3′
3′-TTTGGTTGTAAA TTTTAGTCCTGGAAAG TGTATTTAGAAT-5′
TRPC4AP
C
n=34
LAVA
n=91
Gibbon L1-PA
n=70
Human L1-Ta
bits
2
1
0
-2
-1
+1
+2
+3
+4

Figure 4

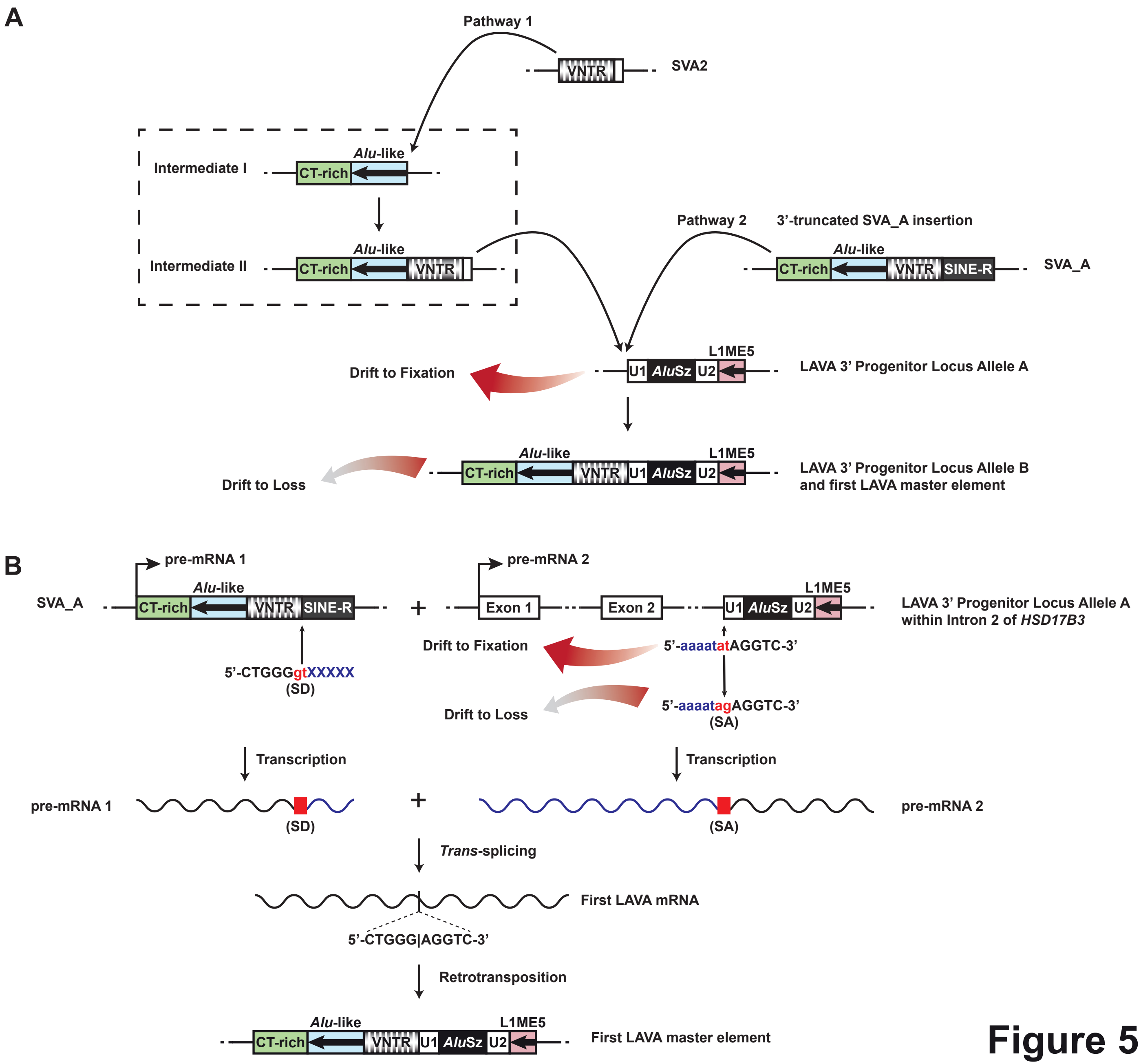
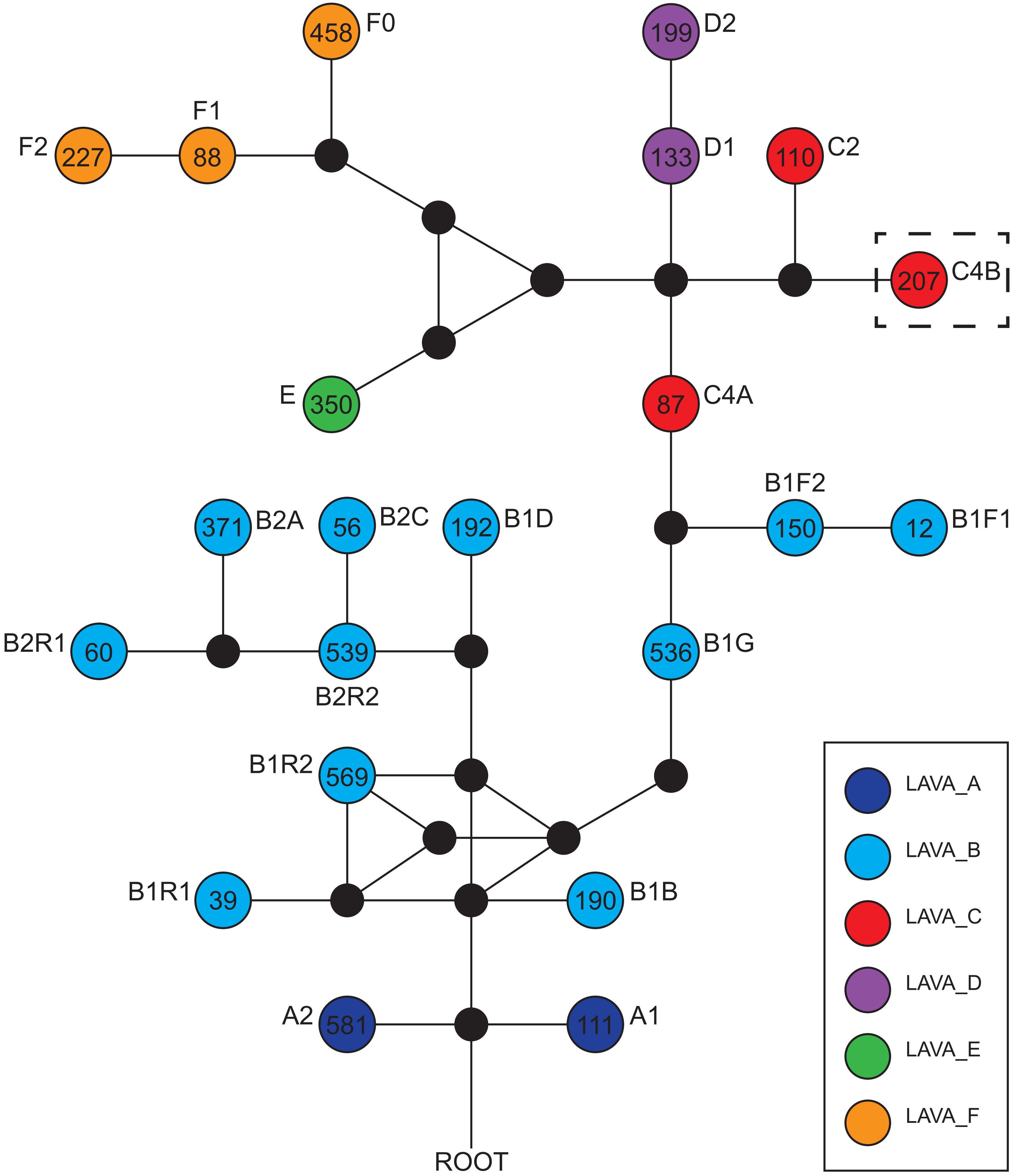

A
Pathway 1
VNTR
SVA2
Intermediate I
Alu-like
CT-rich
Intermediate II
Alu-like
CT-rich
VNTR
Pathway 2
3'-truncated SVA_A insertion
Alu-like
CT-rich
VNTR
SINE-R
SVA_A
L1ME5
Drift to Fixation
U1
AluSz
U2
LAVA 3' Progenitor Locus Allele A
Alu-like
L1ME5
Drift to Loss
CT-rich
VNTR
U1
AluSz
U2
LAVA 3' Progenitor Locus Allele B
and first LAVA master element
B
pre-mRNA 1
Alu-like
SVA_A
CT-rich
VNTR
SINE-R
+
pre-mRNA 2
L1ME5
Exon 1
Exon 2
U1
AluSz
U2
LAVA 3' Progenitor Locus Allele A
within Intron 2 of HSD17B3
5'-CTGGGgtXXXXX
(SD)
Drift to Fixation
5'-aaaatatAGGTC-3'
Drift to Loss
5'-aaaatagAGGTC-3'
(SA)
Transcription
Transcription
pre-mRNA 1
+
pre-mRNA 2
(SD)
(SA)
Trans-splicing
First LAVA mRNA
5'-CTGGG|AGGTC-3'
Retrotransposition
Alu-like
L1ME5
CT-rich
VNTR
U1
AluSz
U2
First LAVA master element

Figure 5



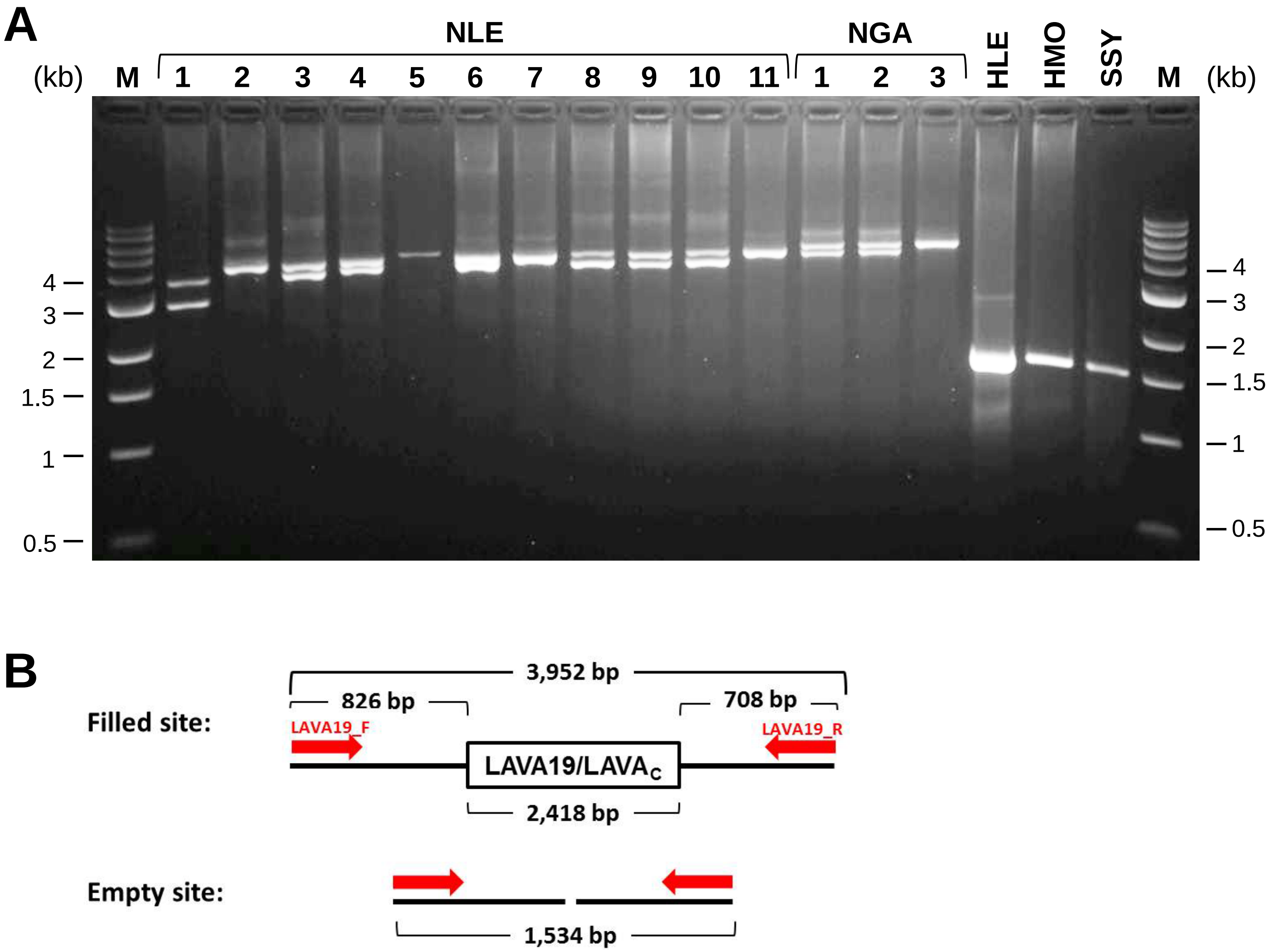
Supplementary Figure S1: LAVA subfamily network (Carbone et al. 2014). Colored circles are LAVA subfamilies. Black circles are hypothetical intermediate sequences speculated to have existed during LAVA evolution but not found in modern genomes. Counts of copies as determined by RepeatMasker analysis of Nleu3.0 are given within colored circles. The dashed box indicates LAVA_C4B, which is the subfamily of the element used in this *trans*-mobilization analysis.

```
               |    (5′ Flank)     || (5′ TSD) ||(SVA start)→
  Nleu3.0      TAGCCATTCCCTATTTCTTTAGAAAGGGACACCCTTCCTTCCTTCCTTCCCTCCCTCCCT 60
      BAC      TAGCCATTCCCTATTTCTTTAGAAAGGGACACCCTTCCTTCCTTCCTTCCCTCCTTCCCT 60
pDRIVE.LAVAc   TAGCCATTCCCTATTTCTTTAGAAAGGGACACCCTTCCTTCCTTCCTTCCCTCCTTCCCT 60
               ****************************************************** *****


  Nleu3.0      CCCTCCTTCCTTCC------------------------TTTCTTTCTTTTCCCTCTGTGG 96
      BAC      CCCTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTTCTTTCTTTCTTTTCCCTCTGTGG 120
pDRIVE.LAVAc   CCCTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTTCTTTCTTTCTTTTCCCTCTGTGG 120
               **************                        **********************


  Nleu3.0      CCCAGGCTGCAATCTACTCAGCTCGCTGCTACCCCGCACCACGGACTGCCTGGGACTTCC 156
      BAC      CCCAGGCTGCAATCTACTCGGCTCGCCGCTACCCCGCACCACGGACTGCCTGGGACTTCC 180
pDRIVE.LAVAc   CCCAGGCTGCAATCTACTCGGCTCGCCGCTACCCCGCACCACGGACTGCCTGGGACTTCC 180
               ******************* ****** *********************************


  Nleu3.0      GGCGCGCGCCACCACCGCTGCCTGCTTTTTCTCCTTTCGCTGCAGGCGCGGTTTCGCCAT 216
      BAC      GGCGCGCGCCACCACCGCTGCCTGCTTTTTCTCCTTTCGCTGCAGGCGCGGTTTCGCCAT 240
pDRIVE.LAVAc   GGCGCGCGCCACCACCGCTGCCTGCTTTTTCTCCTTTCGCTGCAGGCGCGGTTTCGCCAT 240
               ************************************************************


  Nleu3.0      GTTGGCCACGCTGGTCTCCAGCTCCTGACTCCGAGTGCTCTGCCCGCCTCGGCCTCCCGA 276
      BAC      GTTGGCCACGCTGGTCTCCAGCTCCTGACTCCGAGTGCTCTGCCCGCCTCGGCCTCCCGA 300
pDRIVE.LAVAc   GTTGGCCACGCTGGTCTCCAGCTCCTGACTCCGAGTGCTCTGCCCGCCTCGGCCTCCCGA 300
               ************************************************************


  Nleu3.0      GGTGCTGGGACTGCAGACGGACTCTCGCTCACTCGGTGCTCGGTGTTGCCCGGGCTGGAG 336
      BAC      GGTGCTGGGACTGCAGACGGACTCTCGCTCACTCGGTGCTCGGTGTTGCCCGGGCTGGAG 360
pDRIVE.LAVAc   GGTGCTGGGACTGCAGACGGACTCTCGCTCACTCGGTGCTCGGTGTTGCCCGGGCTGGAG 360
               ************************************************************


  Nleu3.0      TGCGGTGGCATGGTCTTGGCTCGCGGCAGCCTCCGCCTCCCAGCCGCCTGCCTTGGCCTA 396
      BAC      TGCGGTGGCATGGTCTTGGCTCGCGGCAGCCTCCGCCTCCCAGCCGCCTGCCTTGGCCTA 420
pDRIVE.LAVAc   TGCGGTGGCATGGTCTTGGCTCGCGGCAGCCTCCGCCTCCCAGCCGCCTGCCTTGGCCTA 420
               ************************************************************


  Nleu3.0      CCAGGGTGCTGGGATTGCAGCCCCTGCCCGGCCACCGCCCCGTCTGGGAGGTGGGGAGCG 456
      BAC      CCAGGGTGCTGGGATTGCAGCCCCTGCCCGGCCACCGCCCCGTCTGGGAGGTGGGGAGCG 480
pDRIVE.LAVAc   CCAGGGTGCTGGGATTGCAGCCCCTGCCCGGCCACCGCCCCGTCTGGGAGGTGGGGAGCG 480
               ************************************************************


  Nleu3.0      TCTCTGCTCGGCCGCCCCTCTGCCCGGCCTCCCCATCTGGGAAGTGAGGAGCGCCTCTGC 516
      BAC      TCTCTGCTCGGCCGCCCCTCTGCCCGGCCTCCCCATCTGGGAAGTGAGGAGCGCCTCTGC 540
pDRIVE.LAVAc   TCTCTGCTCGGCCGCCCCTCTGCCCGGCCTCCCCATCTGGGAAGTGAGGAGCGCCTCTGC 540
               ************************************************************


  Nleu3.0      CCGGCCACCCATCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGGA 576
      BAC      CCGGCCACCCATCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGGA 600
pDRIVE.LAVAc   CCGGCCACCCATCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGGA 600
               ************************************************************


  Nleu3.0      GGTGAGGAGAGCCTCTGCCCGGCCGCCCCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGC 636
      BAC      GGTGAGGAGAGCCTCTGCCCGGCCGCCCCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGC 660
pDRIVE.LAVAc   GGTGAGGAGAGCCTCTGCCCGGCCGCCCCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGC 660
               ************************************* **********************


  Nleu3.0      CACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCACCCACCGTCTGGGAAGTGA 696
      BAC      CACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCACCCACCGTCTGGGAAGTGA 720
pDRIVE.LAVAc   CACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCACCCACCGTCTGGGAAGTGA 720
               ************************************************************
```

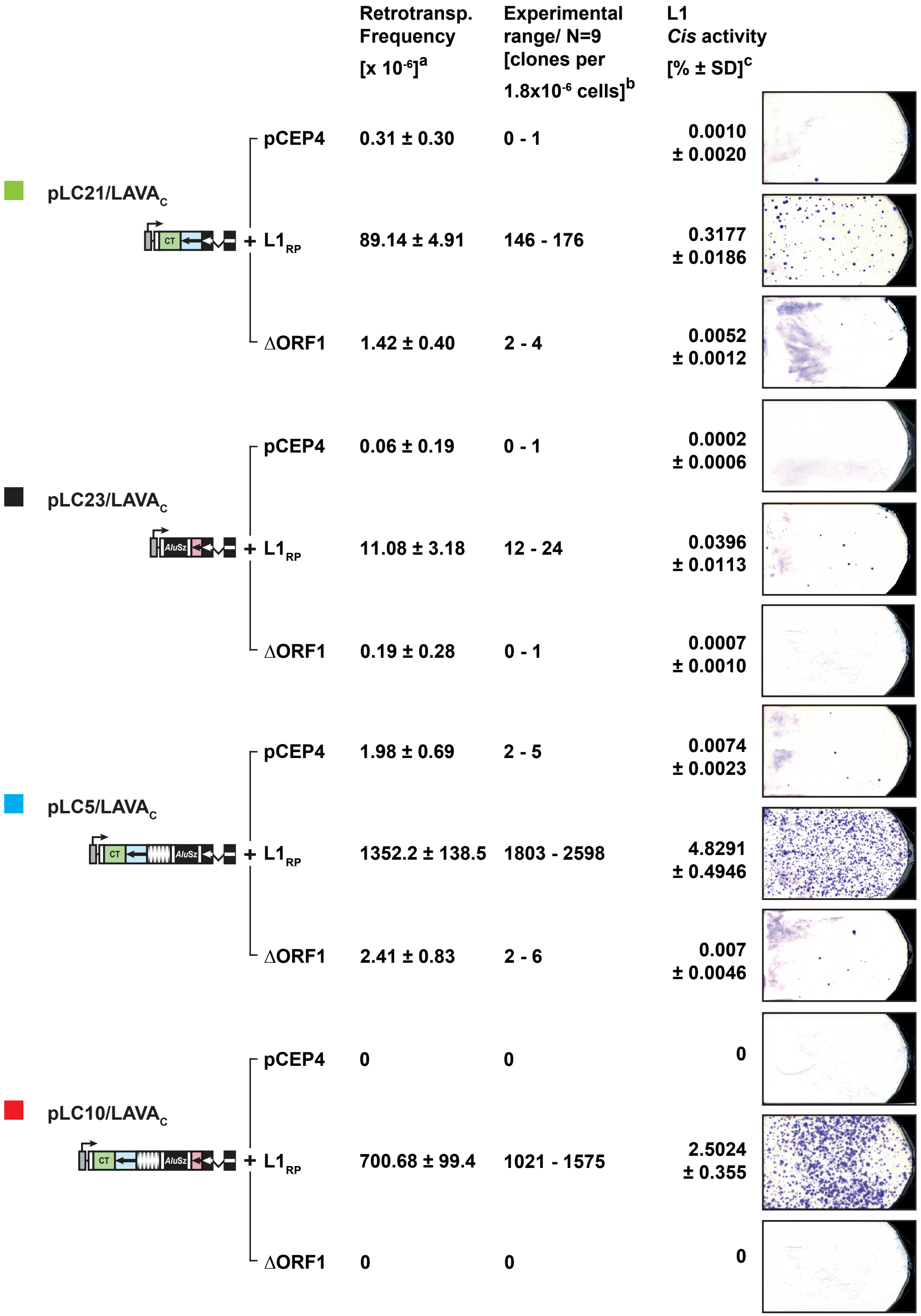
```
     Nleu3.0    GGAGCGCCTCGGCCCGGCCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCTGGCCGC 756
         BAC    GGAGCGCCTCGGCCCGGCCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCTGGCCGC 780
pDRIVE.LAVAc    GGAGCGCCTCGGCCCGGCCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCTGGCCGC 780
                ************************************************************


     Nleu3.0    CTCATCTGGGAAGAGAGGAGCGCCTCTGCCCGGCTGCCCCGTCCGGGGAGAATAGAGGAG 816
         BAC    CTCATCTGGGAAGAGAGGAGCGCCTCTGCCCGGCTGCCCCGTCCAGGAAGAATAGAGGAG 840
pDRIVE.LAVAc    CTCATCTGGGAAGAGAGGAGCGCCTCTGCCCGGCTGCCCCGTCCAGGAAGAATAGAGGAG 840
                ******************************************** ** ************


     Nleu3.0    CACCTCTGCCCGGCCGCCCCGTCCGGGAAGAAGTGAGGAGCGCCTCTGCCCGCCCACCCC 876
         BAC    CGCCTCTGCCCGGCCGCCCCGTCCGGGAAGAAGTGAGGAGCGCCTCTGCCCGCCCACCCC 900
pDRIVE.LAVAc    CGCCTCTGCCCGGCCGCCCCGTCCGGGAAGAAGTGAGGAGCGCCTCTGCCCGCCCACCCC 900
                * **********************************************************


     Nleu3.0    GTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCGCCCCGTCTGGGAAGTGAAGAGCGCCTC 936
         BAC    ATCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCGCCCCGTCTGGGAAGTGAAGAGCGCCTC 960
pDRIVE.LAVAc    ATCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCGCCCCGTCTGGGAAGTGAAGAGCGCCTC 960
                 ***********************************************************


     Nleu3.0    TGCCCGGCCGCCCCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGG 996
         BAC    TGCCCGGCCGCCCCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGG 1020
pDRIVE.LAVAc    TGCCCGGCCGCCCCGTCTGGGAGGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGG 1020
                ************************************************************


     Nleu3.0    AAGTGAGGAGCGCCTCTGCCCGGCCACCCTGTCTGGGAAGTGAGGAGCGCCTCGGCCCGG 1056
         BAC    AAGTGAGGAGCGCCTCTGCCCGGCCACCCCATCTGGGAAGTGAGGAACGCCTCTGCCCGG 1080
pDRIVE.LAVAc    AAGTGAGGAGCGCCTCTGCCCGGCCACCCCATCTGGGAAGTGAGGAACGCCTCTGCCCGG 1080
                *****************************  *************** ****** ******


     Nleu3.0    CCGCC--TCGTCTGGGAAGTGAGGAGCGCCTCGGCCCGGCCACCCATCGTCTGGGAAGTG 1114
         BAC    CCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGGAAGTG 1140
pDRIVE.LAVAc    CCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGGAAGTG 1140
                ** **  ************************* ***************************


     Nleu3.0    AGGAGCGCCTCTGCCCTGCCGCCCCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCTGCC 1174
         BAC    AGGAGCGCCTCTGCCCTGCCGCCCCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCGCC 1200
pDRIVE.LAVAc    AGGAGCGCCTCTGCCCTGCCGCCCCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCGCC 1200
                ******************************************************** ***


     Nleu3.0    TCGTCTGGGAAGAGAGGAGCGCCTCTGCCCGGCCACCCCGTCTGGGAAGTGAGGAGCACC 1234
         BAC    TCGTCTGGGAAGAGAGGAGCGCCTCTGCCCGGCCACCCCGTCTGGGAAGTGAGGAGCACC 1260
pDRIVE.LAVAc    TCGTCTGGGAAGAGAGGAGCGCCTCTGCCCGGCCACCCCGTCTGGGAAGTGAGGAGCACC 1260
                ************************************************************


     Nleu3.0    TCTGCCCGGCCACCCATCATCTGGGAAGTGAGGAGCGCCTCGGCCCGGCC---------- 1285
         BAC    TCTGCCCGGCCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCGCCCCGTCCG 1320
pDRIVE.LAVAc    TCTGCCCGGCCACCCATCGTCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCGCCCCGTCCG 1320
                ****************** ********************** ********


     Nleu3.0    ------------------------------------------------------------
         BAC    GGAAGAAGTGAGGAGCGCCTCTGCCCGCCCACCCCGTCTGGGAGGTGAGGAGCGCCTCTG 1380
pDRIVE.LAVAc    GGAAGAAGTGAGGAGCGCCTCTGCCCGCCCACCCCGTCTGGGAG-TGAGGAGCGCCTCTG 1379


     Nleu3.0    ------------------------------------------------------------
         BAC    CCCGGCCGCCCTGTCTGGGAAGTGAAGAGCGCCTCTGCCCGGCCACCCCGTCTGGGAGGT 1440
pDRIVE.LAVAc    CCCGGCCGCCCTGTCTGGGAAGTGAAGAGCGCCTCTGCCCGGCCACCCCGTCTGGGAGGT 1439
```

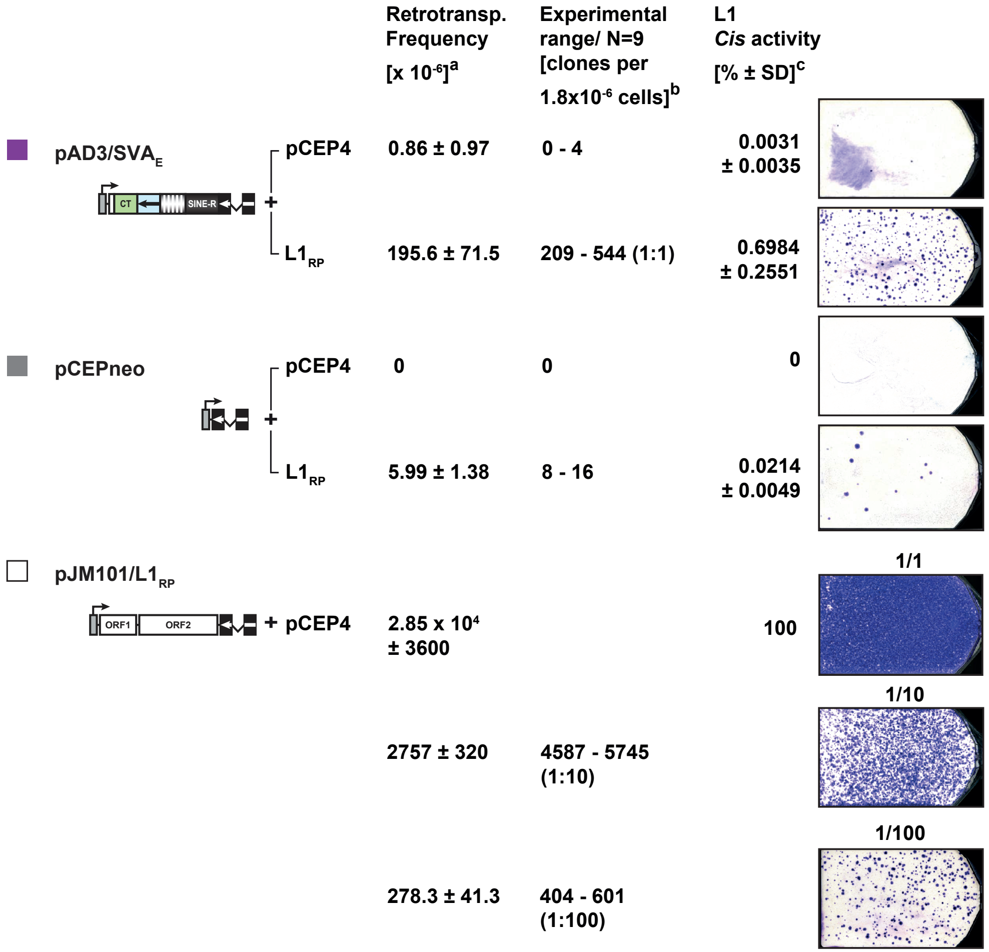
```
 Nleu3.0        ------------------------------------------------------------
     BAC        GAGGAGCGCCTCTGCCCGGCCGCCCCGTC-TGGGAGGTGAGGAGCGCCTCTGCCCGGCCA 1499
pDRIVE.LAVAc    GAGGAGCGC-TCTGCTCG--CGCCCCGTCCTGGGAAGTGAGGAGCGCCTCTGCC-GGTCA 1495


 Nleu3.0        ------------------------------------------------------------
     BAC        CCCATCATCTGGGAAGTGAGGA-GCCCCTCTGCCC-GGCCACCCATCGTCTGGGAAGTGA 1557
pDRIVE.LAVAc    CCCATCATCTGGGAAGTGAGGAGGCCCCTCTGCCCGGGCCACCCATCGTCTGGGAAGTGA 1555


 Nleu3.0        ------------------------------------------------------------
     BAC        GGAGCACCTCTG-CCTGGCCACCCATCGTCTGGGAAGTGAGGG-GCGCCTCTGCCCGGCC 1615
pDRIVE.LAVAc    GGAGCACCTCTGCCCTGGCCACCCATCGTCTGGGAAGTGAGGGAGCGCCTCTGCCCGGCC 1615


 Nleu3.0        ----------------------------------------ACCCATCGTCTGGG-AAGTG 1303
     BAC        ACCCATCGTCTG-GGAAGTGAGGAGCGCCTCTGCCCGGCCACCCATCGTCTGGG-AAGTG 1673
pDRIVE.LAVAc    ACCCATCGTCTGAGGAAGTGAGGAGCGCCTCTACCCGGCCACCCATCGTCTGGGAAAGTG 1675
                                                        ************** *****


 Nleu3.0        AGGAGCGCCTCTGCCCGGCCGCCCCATCCGGGAA-------------------------- 1338
     BAC        AGGAGCGCCTCTGCCCGGCTGCCCCATCTGGGAAGAAGTTAGGAGCGCCTCTGCCCGGCC 1733
pDRIVE.LAVAc    AGGAGCGCCTCTGCCCGGCTGCCCCATCTGGGAAGAAGTTAGGAGCGCCTCTGCCCGGCC 1735
                ******************* ******** *****


 Nleu3.0        --------------GAAGTGAGGAGTGTCTCTGCCCAGCCGCCCCGTCTGGGAAGAAATG 1383
     BAC        GCCCCGTCTGGAAAGAAGTGAGGAGTGTCTCTGCCCGGCTGCCCTGTCTGGGAAGAAATG 1793
pDRIVE.LAVAc    GCCCCGTCTGGAAAGAAGTGAGGAGTGTCTCTGCCCGGCTGCCCTGTCTGGGAAGAAATG 1795
                              ********************** ** **** ***************


 Nleu3.0        AGCAACGCGTCTACCCGGTCGCCCCATCTGGGAAGTGAGGAGCGCCTCTGCCCGGCCGCC 1443
     BAC        AGCAACGCGTCTACCCGGTCGCCCCATCTGGGAAGTGAGGAGCGCCTCTGCCCGG----- 1849
pDRIVE.LAVAc    AGCAACGCGTCTACCCGGTCGCCCCATCTGGGAAGTGAGGAGCGCCTCTGCCCGG----- 1851
                *******************************************************


 Nleu3.0        CCGTCTGGGAAGTGAGGAGTGCCTCTGCCCGGCCGCCCCGTCTGGGATGTAGGGAGTGCC 1503
     BAC        ------------------------------------------------------------
pDRIVE.LAVAc    ------------------------------------------------------------


                                →(SVA end)||(U1 start)        (U1 end)||(AluSz start)→
 Nleu3.0        TCTGCCCGACCGCCCCGTCTGGGAGGTCTACCACAGAGGCCAGAAGCAATGTGGGGGCTG 1563
     BAC        ---------CCGCCCCGTCTGGGAGGTCTACCACAGAGGCCAGAAGCAATGTGGGGGCTG 1899
pDRIVE.LAVAc    ---------CCGCCCCGTCTGGGAGGTCTACCACAGAGGCCAGAAGCAATGTGGGGGCTG 1901
                         ***************************************************


 Nleu3.0        GACGTGGCGGCTCACGCCTGTGGTCCCGGCACTCTGGGGGGCCAAGGCGGGTTGATCACT 1623
     BAC        GACGTGGCGGCTCACGCCTGTGGTCCCGGCACTCTGGGGGGCCAAGGCGGGTTGATCACT 1959
pDRIVE.LAVAc    GACGTGGCGGCTCACGCCTGTGGTCCCGGCACTCTGGGGGGCCAAGGCGGGTTGATCACT 1961
                ************************************************************


 Nleu3.0        TGGGGCTAGGAGTTCGAGACCAGTCTGGCCAACATGGCGAAACATATGAAAAATATGAAA 1683
     BAC        TGGGGCTAGGAGTTCGAGACCAGTCTGGCCAACATGGCGAAACATATGAAAAATATGAAA 2019
pDRIVE.LAVAc    TGGGGCTAGGAGTTCGAGACCAGTCTGGCCAACATGGCGAAACATATGAAAAATATGAAA 2021
                ************************************************************

          →(AluSz end)||(U2 start)
 Nleu3.0        AATACAACAGACAAACCAACCAACCAACTCAGTGACAACAAAACAGGTCTACCCTGGAGT 1743
     BAC        AATACAACAGACAAACCAACCAACCAACTCAGTGACAACAAAACAGGTCTACCCTGGAGT 2079
pDRIVE.LAVAc    AATACAACAGACAAACCAACCAACCAACTCAGTGACAACAAAACAGGTCTACCCTGGAGT 2081
                ************************************************************

                                                         (U2 end)||(L1ME5 end)←
     Nleu3.0   CATACTCTAATTTTTTCTATTTTCCTCCCTTTCTGATCCTTTATCCCACTTTCTTTTTCT 1803
         BAC   CATACTCTAATTTTTTCTATTTTCCTCCCTTTCTGATCCTTTATCCCACTTTCTTTTTCT 2139
pDRIVE.LAVAC   CATACTCTAATTTTTTCTATTTTCCTCCCTTTCTGATCCTTTATCCCACTTTCTTTTTCT 2141
               ************************************************************


     Nleu3.0   TCCTCTTCCTTCTCCTTCTTCTTTGTCAAATAGAGGATTGAGTTATTATCACTGATCTAT 1863
         BAC   TCCTCTTCCTTCTCCTTCTTCTTTGTCAAATAGAGGATTGAGTTATTATCACTGATCTAT 2199
pDRIVE.LAVAC   TCCTCTTCCTTCTCCTTCTTCTTTGTCAAATAGAGGATTGAGTTATTATCACTGATCTAT 2201
               ************************************************************


     Nleu3.0   ATAAAGTCCCTCTCTCATTTATTTTAATTCCCACTCCCCATTTCTATTCCCCGACTTCCC 1923
         BAC   ATAAAGTCCCTCTCTCATTTATTTTAATTCCCACTCCCCATTTCTATTCCCCGACTTCCC 2259
pDRIVE.LAVAC   ATAAAGTCCCTCTCTCATTTATTTTAATTCCCACTCCCCATTTCTATTCCCCGACTTCCC 2261
               ************************************************************


     Nleu3.0   ATGTGCAACCTTCCTAATATATTTGATACGCATCTTTTTGTTTGTATGTATTTTTAGAAA 1983
         BAC   ATGTGCAACCTTCCTAATATATTTGATACGCATCTTTTTGTTTGTATGTATTTTTAGAAA 2319
pDRIVE.LAVAC   ATGTGCAACCTTCCTAATATATTTGATACGCATCTTTTTGTTTGTATGTATTTTTAGAAA 2321
               ************************************************************

                    ←(L1ME5 start)||(poly-A start)        (poly-A end)|| (3’ TSD)
     Nleu3.0   ATGTTTATTGTTTTTGTGTGCAAAAAAAAAATTACTAAAAATAAAAAAAATAAAAAAAAAG 2043
         BAC   ATGTTTATTGTTTTTGTGTGCAAAAAAAAAATTAATAAAAATAAAAAAAATAAAAAAAAAG 2379
pDRIVE.LAVAC   ATGTTTATTGTTTTTGTGTGCAAAAAAAAAATTAATAAAAATAAAAAAAATAAAAAAAAAG 2381
               ********************************** *************************

               (3’ TSD) ||    (3’ Flank)    |
     Nleu3.0   AAAGGGACACTCTAATAAAGAATTCCTAGC 2073
         BAC   AAAGGGACACTCTAATAAAGAATTCCTAGC 2409
pDRIVE.LAVAC   AAAGGGACACTCTAATAAAGAATTCCTAGC 2411
               ******************************
```

Supplementary Figure S2: Alignment of the LAVA$_C$ sequence obtained from the gibbon reference genome (Nleu3.0) with the LAVA$_C$ sequence of the BAC clone gi_148298916_gb_AC202765.2 (BAC) and the consensus of the three sequences obtained from the pDRIVE.LAVA constructs. 20 bp of flanking genomic sequence is present both 5’ and 3’ of the LAVA$_C$ insertion. TSDs, U1, and U2 sequences are shown in green, yellow, and blue, respectively. Arrows indicate the orientation of the repeat components that make up the LAVA structure. SVA- and *Alu*Sz-components are both in sense, while *Alu*-like region and L1ME5-component are in antisense orientation. Binding sites of primers GS.LAVA1, GS.LAVA4, GS.LAVA5, GS.LAVA3 and GS.LAVA2 (listed in their 5’-3’ order they appear in the Clone 1 sequence) are marked in bold and underlined. Asterisks indicate positions at which there is no difference between the Nleu3.0, BAC, and clone sequences. Numbers are the cumulative de-gapped length of each sequence.



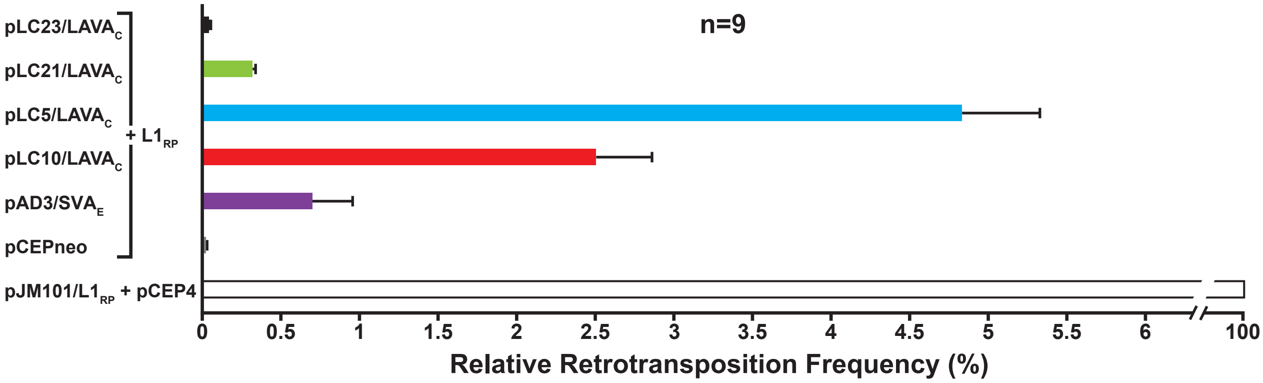
Supplementary Figure S3: PCR-based assay for presence/absence of LAVA$_C$ in the *Nomascus* genus. A) Gel electrophoresis of PCR products obtained using genomic DNA from different gibbon individuals as templates and primers flanking the LAVA$_C$ insertion site. NLE 1-11, *Nomascus leucogenys* individuals 1 -11; NGA 1-3, *Nomascus gabriellae* individuals 1-3 which are members of the same family; HLE, *Hoolock leuconedys; HMO*, *Hylobates moloch*; SSY, *Symphalangus syndactylus*; M, 1 kb DNA ladder (New England Biolabs #N3232); **B)** Primer binding sites (Supplementary Table S1) and expected lengths of PCR products resulting from filled and empty integration sites are schematically outlined. A ~4-kb PCR product corresponding to the presence of the LAVA$_C$ insertion is evident in both *Nomascus* species and absent from representatives of the remaining genera tested. In the genera *Hylobates moloch*, *Symphalangus syndactylus*, and *Hoolock leuconedys*, a 1.5-kb product corresponding to the empty site was amplified. The presence of a second PCR product detectable in NLE samples 1, 3, 4, 6, and 8-10 and NGA samples 1 and 2 is most likely due to the presence of two different alleles in the population whose LAVA$_C$-specific PCR products differ from each other probably as a consequence of variations in their VNTR lengths. Red arrows, binding sites of primers LAVA19_F and LAVA19_R used for filled site/empty site PCR.





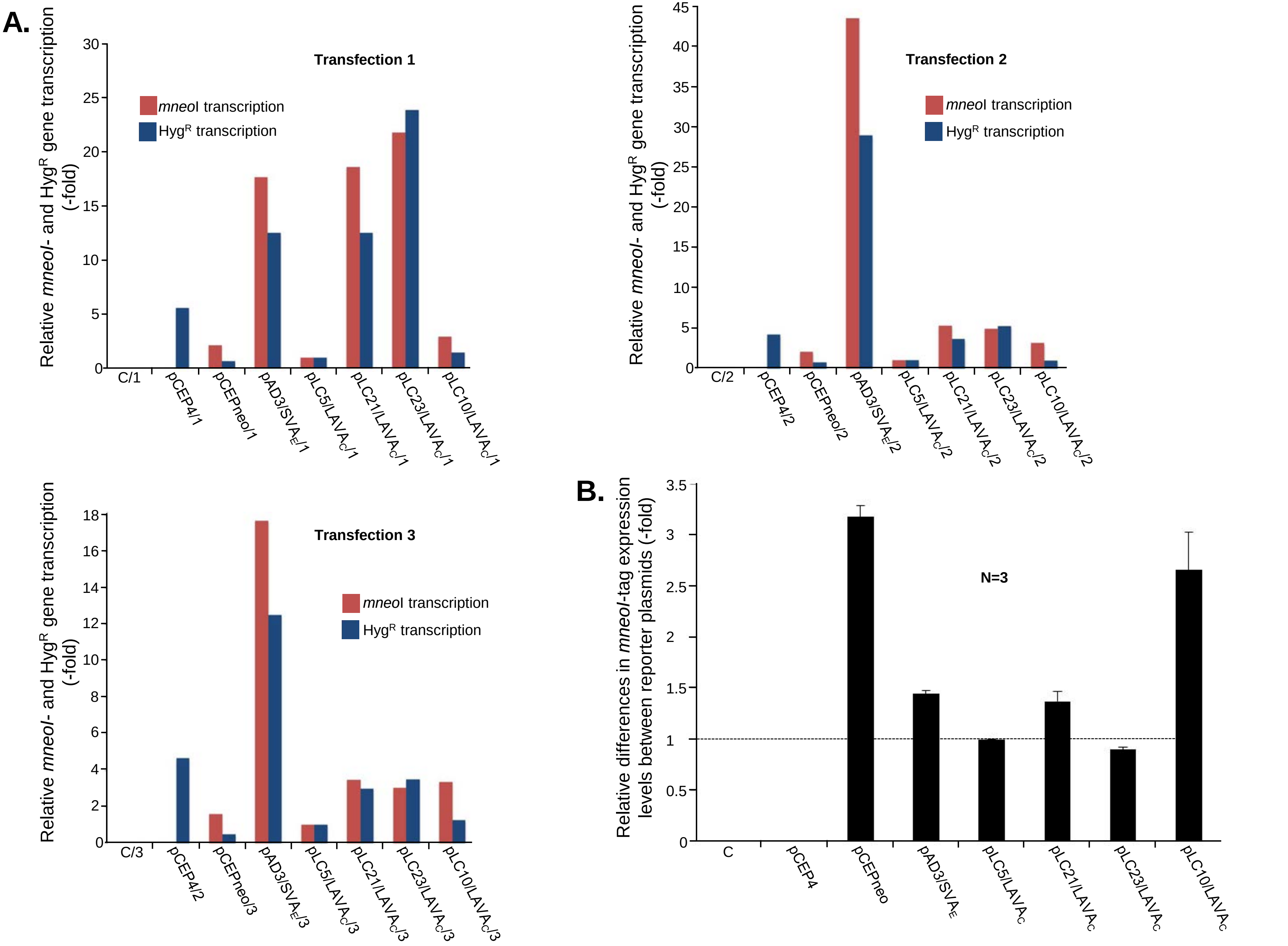
		Retrotransp. Frequency [x 10^{-6}][a]	Experimental range/ N=9 [clones per 1.8x10^{6} cells][b]	L1 *Cis* activity [% ± SD][c]	
🟩 pLC21/LAVA$_C$	pCEP4	0.31 ± 0.30	0 - 1	0.0010 ± 0.0020	
	+ L1$_{RP}$	89.14 ± 4.91	146 - 176	0.3177 ± 0.0186	
	ΔORF1	1.42 ± 0.40	2 - 4	0.0052 ± 0.0012	
⬛ pLC23/LAVA$_C$	pCEP4	0.06 ± 0.19	0 - 1	0.0002 ± 0.0006	
	+ L1$_{RP}$	11.08 ± 3.18	12 - 24	0.0396 ± 0.0113	
	ΔORF1	0.19 ± 0.28	0 - 1	0.0007 ± 0.0010	
🟦 pLC5/LAVA$_C$	pCEP4	1.98 ± 0.69	2 - 5	0.0074 ± 0.0023	
	+ L1$_{RP}$	1352.2 ± 138.5	1803 - 2598	4.8291 ± 0.4946	
	ΔORF1	2.41 ± 0.83	2 - 6	0.007 ± 0.0046	
🟥 pLC10/LAVA$_C$	pCEP4	0	0	0	
	+ L1$_{RP}$	700.68 ± 99.4	1021 - 1575	2.5024 ± 0.355	
	ΔORF1	0	0	0	
🟪 pAD3/SVA$_E$	pCEP4	0.86 ± 0.97	0 - 4	0.0031 ± 0.0035	
	+ L1$_{RP}$	195.6 ± 71.5	209 - 544 (1:1)	0.6984 ± 0.2551	
◼ pCEPneo	pCEP4	0	0	0	
	+ L1$_{RP}$	5.99 ± 1.38	8 - 16	0.0214 ± 0.0049	
□ pJM101/L1$_{RP}$	+ pCEP4	2.85 x 10^{4} ± 3600		100	1/1
		2757 ± 320	4587 - 5745 (1:10)		1/10
		278.3 ± 41.3	404 - 601 (1:100)		1/100

Supplementary Figure S4: Human L1-mediated retrotransposition of the LAVA$_C$ reporter element and its truncated versions. G418^R clones in T-75 flasks were Giemsa-stained and indicate marked LAVA$_C$ and SVA$_E$ retrotransposition events, and processed pseudogene fomation (pCEPneo). [a]number of G418^R foci per 10^6 cotransfected HeLa cells; [b]number of G418^R foci per T-75 flask on which 1.8x10^6 cells were plated for cotransfection; [c]LAVA and SVA retrotransposition frequency and processed pseudogene formation was measured relative to the *cis*-activity of L1 encoded by pJM101/L1$_{RP}$ which was set as 100%.



	Retrotransposition Frequency (x10^{-6})	L1 Cis activity (%)
pLC23/LAVA$_C$	11.08 ± 3.18	0.040
pLC21/LAVA$_C$	89.14 ± 4.91	0.318
pLC5/LAVA$_C$	1352.2 ± 138	4.829
pLC10/LAVA$_C$	700.7 ± 99.4	2.502
pAD3/SVA$_E$	195.62 ± 71.52	0.698
pCEPneo	5.99 ± 1.38	0.021
pJM101/L1$_{RP}$ + pCEP4	2.85 x 10^4 ± 3.6 x 10^3	100

Supplementary Figure S5: *Trans*-mobilization frequency of the full-length LAVA$_C$ element (pLC10/LAVA$_C$) comprises ~2.5% of the L1$_{RP}$ *cis* activity. LAVA$_C$ expression cassettes encoded by pLC10/LAVA$_C$ and pLC5/LAVA$_C$ and harboring the VNTR-U1-*Alu*Sz-U2 module, are *trans*-mobilized ~7.9-fold and ≥15.2-fold more efficiently than the 3'-truncated LAVA$_C$ reporter element of pLC21/LAVA$_C$ lacking this combination of LAVA modules. To better illustrate the significant differences in retrotransposition activities between the diverse LAVA$_C$ fragments, only those *trans*-mobilization frequencies presented in Figure 2B that were observed in the presence of both overexpressed L1$_{RP}$-encoded proteins (L1$_{RP}$) were displayed using a different scale in the x-axis of the bar diagram. The number of retrotransposition events per 10^6 co-transfected cells and retrotransposition frequencies relative to the *cis*-retrotransposition activity of the L1 reporter cassette of pJM101/L1$_{RP}$ (L1 Cis activity [%]) which was set as 100%, are listed.



Supplementary Figure S6: Normalization of *mneo*I expression against expression of episomal plasmid-encoded selection marker Hyg^R.

Supplementary Figure S6: Normalization of *mneo*I expression against expression of episomal plasmid-encoded selection marker gene Hyg^R.

A) Analysis of the relative expression of the reporter cassette -encoded *mneo*I gene and the eukaryotic selectable marker gene Hyg^R encoded by the plasmid backbone.

In order to accurately confine the impact of variations in transcription rates of the six different *mneo*I-tagged reporter cassettes from variable transfection efficiencies on the variability of overall spliced *mneoI* transcript levels in the differently transfected cells, we analyzed both reporter cassette and Hyg^R expression levels in HeLa cells transfected exclusively with a reporter plasmid in the absence of any driver plasmid and without hygromycin selection. Subsequently, we normalized qRT-PCR data on spliced *mneo*I transcripts with qRT-PCR data on Hyg^R expression (which is directly proprtional to the number of pCEP4-based reporter plasmids in the transfected cells).

A) Relative amounts of both spliced *mneo*I-encoded transcripts and Hyg^R transcripts isolated from cells that were exclusively transfected with the reporter plasmid pCEPneo, pAD3/SVA$_E$, pLC5/LAVA$_C$, pLC21/LAVA$_C$, pLC23/LAVA$_C$, or pLC10/LAVA$_C$, were quantified by qRT-PCR. Relative *mneo*I and Hyg^R transcript levels expressed in pLC5/LAVAC-transfected cells were set as 1. In each of the 3 transfection rounds (Transfection 1-3), each of the six reporter plasmids and the negative control construct pCEP4 were separately transfected in parallel. Three different plasmid DNA preparations of each reporter construct were generated, and for each of the 3 transfection rounds, a separate plasmid preparation of each construct was used. Total RNA from untransfected (C/1 to C/3) and pCEP4-transfected cells (pCEP4/1 to pCEP4/3) served as negative controls.

B) Arithmetic means of the relative differences in *mneo*I-transcript levels expressed from various reporter cassettes after normalization against Hyg^R expression which is a measure for differences in transfection efficiencies. Arithmetic means were calculated form qRT-PCR data obtained from the three transfection experiments presented in panel A. *mneo*I-expression in pLC5/LAVA$_C$-transfected cells was set as 1. C, untransfected cells; pCEP4, pCEP4-transfected cells. Error bars indicate standard errors of the mean (SEM).

Supplementary Table S1: Primers used for isolation and sub-cloning of LAVA reporter elements (lines 1-7) and for EPTS-LM PCR (lines 8-17).

Oligonucleotide	**Sequence 5’ → 3’**	**Remarks**
GS_LAVA19_F	TGCCTATTTTCATAGCCATTCCC	Amplification of LAVA19 fromBAC DNA
GS_LAVA19_R	CTCTGCCCTCGGATGCTAGG	Amplification of LAVA19 fromBAC DNA
GS.LAVA1	**GCGGTACC**TTT**AGAAAGGGACAC**CCTTC LAVA19 TSD	PCR amplification of LAVA fragments in pLC5/LAVA$_{\mathrm{F}}$, pLC10/LAVA$_{\mathrm{F}}$, and pLC21/LAVA$_{\mathrm{F}}$
GS.LAVA2	**CCGCTAGC**GCACACAAAAACAATAAACATTTTC	PCR amplification of LAVA fragments in pLC10/LAVA$_{\mathrm{F}}$ and pLC23/LAVA$_{\mathrm{F}}$.
GS.LAVA3	**CCGCTAGC**GGTTGGTTGGTTTGTCTGTTG	PCR amplification of LAVA fragment in pLC5/LAVA$_{\mathrm{F}}$,
GS.LAVA4	**CCGCTAGC**GAGGCGCTCCTCTATTCTTC	PCR amplification of LAVA fragment in pLC21/LAVA$_{\mathrm{F}}$,
GS.LAVA5	**GCGGTACC**TACCACAGAGGCCAGAAG	PCR amplification of LAVA fragment in pLC23/LAVA$_{\mathrm{F}}$,
GS86	GAAGAACTCGTCAAGAAGGCGATAGAAGG	Flanking Neo-Intron
GS87	GCCATTGAACAAGATGGATTGCACGCAGG	Flanking Neo-Intron
GS 177	bio- CCAGCCACGATAGCCGCGCTGCCTCGTCCTGAAGCTC	EPTS
LS2	AAACGACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG	EPTS linker
LS1	CCTAACTGCTGTGCCACTGAATTCAGATCTCCCGGGTC	EPTS linker
GS 90	TTCCACACCCTAACTGACACACATTCC	Post-EPTS amplification 5’
GS 94	GGACAGGTCGGTCTTGACAAAAAGAACCG	Post-EPTS amplification 5’
OC I	GACCCGGGAGATCTGAATTC	EPTS linker-specific
OC II	AGTGGCACAGCAGTTAGG	EPTS linker-specific
GS88	CCTTCTATCGCCTTCTTGACGAGTTCTTC	5’ end amplification downstream primer

Supplementary Table S2: Oligonucleotides used for amplification of full/empty sites, 5' junctions, and 3' junctions of LAVA de novo insertions. [Note: 3' junction of Clone 10 could not be amplified].

Locus	**Oligonucleotide Name**	**Sequence (5' → 3')**	**Used in amplification**
Clone 14	C14_5pFlank	GCCGGAAAACCTCTCACAGA	Full/empty site and 5' junction.
	C14_3pFlank	GTCGTTGGCCCCTTACTACC	Full/empty site.
	LAVA19_5pSVA1	TTCCCTCTGTGGCCCAGG	5' junction.
	GS90	TTCCACACCCTAACTGACACACATTCC	3' junction.
	LINK_2A	CTTAAGTCTAGAGGGCCCAGGGATTGACGACACGGTGA	3' junction.
Clone 6	C6_5pFlank_B	GCGTTTCCAGCCCGTTATTC	Full/empty site and 5' junction.
	C6_5pFlank_C	GCCCCTTCCACTAATGCCAT	Full/empty site.
	C6_3pFlank_B	CGAAGGAGTTCAGGCAGGTT	Full/empty site.
	LAVA19_5pSVA1	CCTGGGCCACAGAGGGAA	5' junction.
	GS90	TTCCACACCCTAACTGACACACATTCC	3' junction.
	LINK_2A	CTTAAGTCTAGAGGGCCCAGGGATTGACGACACGGTGA	3' junction.
Clone 10	C10_5pFlankb	GGAAGAAGACCTCTCCCCCT	Full/empty site.
	C10_5pFlank_D	CTCACCTGGCCATCGGTTTA	Full/empty site and 5' junction.
	LAVA19_5pU2A	CAACAGACAAACCAACCAACCA	5' junction.
	C10_3pFlank_B	TCGCCCAAGGTCACATAACC	3' junction.

Supplementary Table S3: Recovered LAVA insertions are listed by clone number (column 1), genomic location of the pre-insertion site (column 2), and if the insertion occurred into a gene (column 3). Column 4 indicates the orientation of the insertion relative to the reference genome (hg19). Column 5 lists the length of the entire retrotransposed sequence, including the spliced *mneo*I retrotransposition indicator cassette (1.7 kb) and LAVA sequences. The text in parentheses indicates whether the insertion is full-length or, if truncated, in which SVA module truncation occurred. Column 6 indicates the percent GC content for of +/- 5kb of flanking sequence around the pre-insertion site in the reference genome. TSDs and L1 EN cleavage sites on the bottom strand of the genomic DNA are listed in columns 7 and 8, respectively. Column 9 lists lengths of polyA tails in the number of nucleotides.

[a]Mean lengths of SVA de novo insertions with characterized 5' ends, TSDs and polyA tail lengths are shown. Genomic coordinates and gene annotations are according to hg19 (http://genome.ucsc.edu).

Clone	Genomic location	Gene	Strand	Length (bp), (module)	GC Content (+/- 5kb)	TSD (length)	L1 EN site (TTTT/a)	PolyA length
6	chr22:45563158-45563159	*NUP50* (Intron 1)	Antisense	~ 4,692 (full-length)	47.3%	AAAAAACCTTTCCA (14)	TTTT/gt	76
10	chr17:73299414-73299444	-	Antisense	~ 3,853 (5' truncated *Alu*Sz)	49.1%	Unknown	Unknown	Unknown
14	chr20:33600095-33600096	*TRPC4AP* (Intron 11)	Coding	~ 4,692 (full-length)	50.0%	AAAATCa/gGGACCTTTC (16)	TTTT/aa	13
Mean				~4,412	48.8%	15		44.5

Supplementary Table S4: Genomic target sequences of preexisting gibbon retrotransposons.

A. Target sequences of preexisting gibbon L1-PA elements.

TSD Sequence	TSD Length	EN Cleavage Site	TSD Sequence	TSD Length	EN Cleavage Site
AAAAAAAAAAAAAAGT	16	CTAAAA	AATAAAGAAAAT	12	TGAATA
AAAAAAAAAAAAAAGTTGAACCA	23	AAAAAA	AATAACTCTAAGAATA	16	CTAATA
AAAAAAAAGAAAG	13	TCAAAA	AATAATATTAAATGTG	16	GCAATA
AAAAAAACCAAAAG	14	CAAATA	AATAATTATAA	11	TTAATA
AAAAAAGAAGA	11	TTAAAA	AATACAAGTAAATATTTTTGGC	22	TTAAAA
AAAAAAGTATAAACAG	16	TTAAAA	AATACAGAAACCAGTAA	17	TTAAAA
AAAAAGATAAACATTGGTT	19	ACAAAA	ACAAGTTTA	9	TTATAA
AAAAAGTAGATATA	14	GTAAAA	AGAAAGAAACAA	12	TTAGAA
AAAAAGTCAGGAAAC	15	TTAAAA	AGAAATTCATGCA	13	TGACAA
AAAAATAATAATTAT	15	TTAATA	AGAATGGGAGAAAATA	16	ACAGAA
AAAAATAGGAATGAGGAG	18	ATAAAA	AGGGAAAGAAATTGATACTTTA	22	TGAAAA
AAAAATCAAA	10	CTAAAA	ATAAAAAAAAGTG	13	ATATAA
AAAAATCAGTTTG	13	TTAAAA	ATAAAAAATTAAA	13	CTATAA
AAAAATCCACAGTT	14	TTAAAA	ATAAAACTATTGTC	14	TTAGAA
AAAAATGGAAAGCAA	15	TTAAAA	ATAATAAAAGTAAATTTTA	19	TTAAAA
AAAAATTAAGGC	12	TTAAAA	ATAGAATCCCGAGAGGACCGAAAA	24	TTAAAG
AAAAGGAAATAAAAGTA	17	TTATAA	GAAAATTAACTAAACTTTGA	20	ATGAAA
AAAATAAGA	9	TCAAAA	GACAATAATAATCTAAAGA	19	TTAATA
AAAATATAATTCATCAGTCCATT	23	TTTATA	GTAAGAATACGTTTCA	16	ATGTAA
AAAATCTAAGACCTGAA	17	TTAAAA	TATCAAAGGCAG	12	ACTATC
AAACAAAAACTTGGG	15	TGAGAA	AATTTTAAAAGAAGAGAG	18	TTAAAA
AAACAAGAACAGCAGATGG	19	GCAAAA	AAAATATGA	9	ATAAAA
AAACAGTCCAAATGAG	16	TTAAAA	AAAAAAAAAAAAAAA	15	TTAAAA
AAAGAAAA	8	AAAAAG	GGAAGAAACAAAAACA	16	GTAAAA
AAAGAAAAGTAAGTG	15	TTAAAT	GAAAATAAGTGAATTA	16	TTGAAA
AAAGAAATAATTGA	14	TGAAAG	AAAGATAAATAAATA	15	TCAAAA
AAATCCTAACTAAAAA	16	GGAAAA	AAGAATTGGGATGTGGAG	18	TTAAGA
AACAATACATCACGCAGTTG	20	ATAACA	AAAATAATAATTAA	14	GCAAAA
AAGAAAAAAT	10	ATAAAA	ATAAGAAGAGACAATGAA	18	CTAAAA
AAGAAAATAATTTT	14	TAAATA	AAGAAAATA	9	TTAAAA
AAGAAATTAAAAAGTA	16	GTAAAA	TAAAATTAAATGCTTGAGGG	20	TCTAAA
AAGAATTAAAAGTTGA	16	TTAAGA	AAAAGTAAACAATTTAGT	18	TTAAAA
AAGAATTTAAAAAGTCGCA	19	TTAAGA	AAAAGATATTTCCTT	15	TTAAAA
AAGTGGAAAAA	11	TTAAGT	AAAATTAAGAA	11	TTAAAA
AAAAGATCCCAAGGGG	16	GTAAAA	ATAAAAATAAAATTAAAAGG	20	TCATAA
AATAAAAGAGA	11	GAAATA	AAAATAATAATAATA	15	TTAAAA
AAAATGCTTCTGCTGGGGT	19	TCAAAA	ACAATAAATTATTTTTG	17	GCAAAA
AAGAGATTATATA	13	TTAAGA	AGGAAGGAATAAAGAAATGGGGGGAGG	27	TTAAGA
AATCAGAGGTCCA	13	CAAATT	AAAACAATAAAAAGAGAAAA	20	GCAAAA
AAAATCAGCTACTT	14	TTAAAA	AAAAGAACAAGTCGGGG	17	GAAAAA
AAACAGCAAAATT	13	TTAAAT	AAAAAATTACTG	12	ATAAAA
CAAATATTAAGCTGGA	16	TCCAAA	AAAATGTAAAAAT	13	ATAAAA
AGAAATTTTGAAACACCTG	19	TTAGAA	AAAAAATAAA	10	TTAAAA
ACAAAAAACAAAAACAAAACACACACACA	29	TTAAAA	AGAAATATCTCTCATTTT	18	TTAGAA
AAAAAAAAAGCCCATGACC	19	TTAAAA	GGAAAGAATGGTTCACTT	18	TTAAAA
AAAAGCATGGGCTTTG	16	TTAAAA			

B. **Target sequences of preexisting gibbon LAVA elements (Carbone et al. 2012).**

BAC Acc. Number	Target Site Duplication	TSD Length	Target site
AC182776.2	Aa/gAAGCCTGTAGC	14	TTTT/AA
AC187943.2	AAGGCCTTCATTTGG	16	CCTT/AA
AC190010.2	AAACATTTGAACA	14	GTTT/AG
AC198100.1	AAGAAGGTAAAGC	14	TCTT/AA
AC198102.2	AGAAGGGGCAGGGT	15	TCTT/GA
AC198155.2	AAGATATAAATGTTTC	17	TCTT/AA
AC198875.2	AAAATGGATGAAAGA	16	TTTT/AA
AC199991.2	AAGTGGGATTAa/gCA	15	ACTT/AT
AC200373.2	AAAATGCAAATGCTC	16	TTTT/AA
AC200379.2	AAAAAGCTTATCCACCACA	20	TTTT/GA
AC200379.2	AAAAATTAGTCGGGCA	17	TTTT/GT
AC200383.2	AAGAGATGAAG	12	TCTT/GA
AC202760.1	AAAATGAAGTTGG	14	TTTT/TT
AC202765.2	AGAAAGGGACAC	13	TCTT/AA
AC203714.2	AAGAAGAAAATT-/tCCT	17	TCTT/AT
AC203720.2	AAGGATGTGG	11	TCTT/AT
AC203720.2	AAGAGCTGTAACACAC	16	CCTT/AA
AC203721.2	AAAATGTTGGGATTACA	18	TTTT/TG
AC204185.2	AAAAGAATAG	11	TTTT/AA
AC204446.3	AAAAAAAAAGAACTAGAGGC	21	TTTT/GA
AC204794.2	AAAATTATTTCTC	14	TTTT/AG
AC206079.2	AAGAATATCCTCAGGATT	19	TCTT/AA
AC208823.3	AAAGAATGGGCAAGG	16	CTTT/GT
AC208823.3	AAAGAATGGGCAAGG	16	CTTT/GT
AC208960.3	AGAAAAGACAGATCCC	17	TTCT/AA
AC209355.3		N/A	CTTT/CT
AC214603.3	AAGAAATGTAAATGGT	17	TCTT/GT
AC214650.2	AAAAATTAGTCGGGCA	17	TTTT/GA
AC216139.3	ACAAAATTAGCCAGGCATG	20	TTGT/AT
AC216158.2	AGAAATATTTGCTGAAGTC	20	TTCT/AT
AC217745.2	AAAAGTAAGAAATTG	16	TTTT/GT
CT954291.6	AAGACTGTCCCTGATCAA	19	TCTT/CA
CT954299.8	GAAAAAAGTGCCTTT	16	TTTC/AA
CT954309.16	AAAAATTAGTCGGGCA	17	TTTT/GT

Supplementary Methods:

qRT-PCR analysis for normalization of *mneo*I-tagged reporter transcription against Hyg^R expression

2x10^5 HA-HeLa cells were separately transfected with 1µg of pCEP4, pCEPneo or a LAVA or SVA retrotransposition reporter plasmid. Three days after transfection, total RNA was extracted from each of the transfected cell cultures and from untransfected HA-HeLa cells (C), using the RNeasy® Minikit (Quiagen) with On-Column DNase digestion following the manufacturer's instructions. Using the SuperScript®III First-Strand synthesis SuperMix for qRT-PCR (Invitrogen, Cat. No. 11752050), first-strand cDNA was synthesized from 1 µg of DNaseI-digested total RNA, according to the manufacturer's instructions. To quantify levels of spliced *mneo*I transcripts and Hyg^R transcripts expressed from the *mneo*I-tagged reporter elements and the corresponding plasmid backbone-encoded Hyg^R gene of pCEPneo, pAD3/SVA$_E$, pLC5/LAVA$_C$, pLC21/LAVA$_C$, pLC23/LAVA$_C$, pLC10/LAVA$_C$, respectively, real time PCR was performed for each transfection in triplicate using TaqMan® chemistry (Applied Biosystems) in an Applied Biosystems 7900 HT Fast Real-Time PCR system base unit. Each transfection was performed in three biological replicates from three independent plasmid preparations. Primer/probe combinations used to specifically recognize the spliced *mneo*I cassette and to quantify cellular 18srRNA levels are described in the paragraph 'Quantitative Real-Time RT-PCR to analyze *mneo*I -transcription after cotransfection of driver and reporter plasmids' in the 'Materials and Methods' section. To quantify Hyg^R-encoded transcripts , we used the following primer/probe combination: Hygrofor: 5'-CGCAAGGAATCGGTCAATACACTA-3'; Hygrorev: 5'-CACAGTTTGCCAGTGATACACATG-3'; 5'-FAM-CATGGCGTGATTTCAT; Both spliced *mneo*I transcript levels and Hyg^R transcript levels were normalized against cellular 18srRNA levels.

Cycling conditions were as follows: 95°C for 15 min (initial denaturation), 95°C for 15s and 60°C for 1 min (40 cycles). The software applied to analyze real-time and end point fluorescence was RQ manager 1.2. Relative quantification of RNA expression was carried out using the $\Delta\Delta C_t$ method (Livak and Schmittgen 2001).