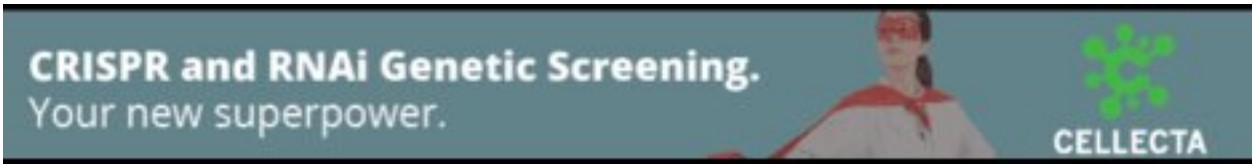
Somatic retrotransposition in the developing rhesus macaque brain

Victor Billon, Francisco J Sanchez-Luque, Jay Rasmussen, et al.

Genome Res. published online June 21, 2022
Access the most recent version at doi:10.1101/gr.276451.121



Somatic retrotransposition in the developing rhesus macaque brain

Victor Billon[1,2,11], Francisco J. Sanchez-Luque[3,4,5,11], Jay Rasmussen[1,11], Gabriela O. Bodea[1,6,11], Daniel J. Gerhardt[6], Patricia Gerdes[6], Seth W. Cheetham[6], Stephanie N. Schauer[6], Prabha Ajjikuttira[1], Thomas J. Meyer[7], Cora E. Layman[8], Kimberly A. Nevonen[8], Natasha Jansz[6], Jose L. Garcia-Perez[3,4], Sandra R. Richardson[6], Adam D. Ewing[6,*], Lucia Carbone[7,8,9,10,*], Geoffrey J. Faulkner[1,6,*]

[1]Queensland Brain Institute, University of Queensland, St. Lucia, QLD, 4067, Australia.
[2]Biology Department, École Normale Supérieure Paris-Saclay, 91190 Gif-sur-Yvette, France.
[3]GENYO. Pfizer-University of Granada-Andalusian Government Centre for Genomics and Oncological Research, PTS Granada 18016, Spain.
[4]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom.
[5]Institute of Parasitology and Biomedicine 'Lopez-Neyra' – Spanish National Research Council, PTS Granada, 18016, Spain.
[6]Mater Research Institute - University of Queensland, Woolloongabba, QLD, 4102, Australia.
[7]Division of Genetics, Oregon National Primate Research Center, Beaverton, Oregon, USA.
[8]Department of Medicine, Knight Cardiovascular Institute, Oregon Health & Science University, Portland, Oregon, USA.
[9]Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA.
[10]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA.
[11]These authors contributed equally to this work.

*Corresponding authors: adam.ewing@mater.uq.edu.au, carbone@ohsu.edu, faulknergj@gmail.com

Abstract

The retrotransposon LINE-1 (L1) is central to the recent evolutionary history of the human genome, and continues to drive genetic diversity and germline pathogenesis. However, the spatiotemporal extent and biological significance of somatic L1 activity is poorly defined, and is virtually unexplored in other primates. From a single L1 lineage active at the divergence of apes and Old World monkeys, successive L1 subfamilies have emerged in each descendant primate germline. As revealed by case studies, the presently-active human L1 subfamily can also mobilize during embryonic and brain development *in vivo*. It is unknown whether non-human primate L1s can similarly generate somatic insertions in the brain. Here we applied ~40× single-cell whole genome sequencing (scWGS), and retrotransposon capture sequencing (RC-seq), to 20 hippocampal neurons from two rhesus macaques (*Macaca mulatta*). In one animal, we detected and robustly PCR validated a somatic L1 insertion that generated target site duplications, carried a short 5′ transduction, and was present in ~7% of hippocampal neurons but absent from cerebellum and non-brain tissues. The corresponding donor L1 allele was exceptionally mobile *in vitro*, and was embedded in *PRDM4*, a gene expressed throughout development and in neural stem cells. Nanopore long-read methylome and RNA-seq transcriptome analyses indicated young retrotransposon subfamily activation in the early embryo, followed by repression in adult tissues. These data highlight endogenous macaque L1 retrotransposition potential, provide prototypical evidence of L1-mediated somatic mosaicism in a non-human primate, and allude to L1 mobility in the brain over the last 30 million years of human evolution.

Introduction

Neurons can acquire mutations during brain development and later life. Human single-cell and bulk tissue genomic analyses have revealed neuronal somatic copy number variants (McConnell et al. 2013; Chronister et al. 2019), single nucleotide variants (Chen et al. 2017; Xing et al. 2021; Abascal et al. 2021), and transposable element (TE) insertions (Sanchez-Luque et al. 2019; Evrony et al. 2015; Erwin et al. 2016). The estimated frequency of each of these mutations depends on the detection method and brain region assayed (Chronister et al. 2019; Xing et al. 2021; McConnell et al. 2017; Abascal et al. 2021). The spatial distribution of a given somatic variant is influenced by its type and genomic location, the cell in which it arose and, potentially, post-mutational selection. As a result, a mutation may be present throughout the brain, or in only one cell, and in either case contribute to a wider mosaic genome landscape. While somatic variants can drive neuronal pathogenesis, such as mTOR

mutations leading to focal cortical dysplasia (Nakashima et al. 2015; Lim et al. 2015; King et al. 2015), it is unclear whether they influence normal brain function or phenotype (McConnell et al. 2017; Faulkner and Garcia-Perez 2017).

The retrotransposon long interspersed element 1 (LINE-1, or L1) constitutes 17% of the human genome (International Human Genome Sequencing Consortium 2001). Despite their prevalence, only a small subset of L1 copies are mobile in the germline, or in neuronal lineage cells (Erwin et al. 2016; Macia et al. 2017; Muotri et al. 2005; Coufal et al. 2009; Evrony et al. 2015; Sanchez-Luque et al. 2019; Faulkner and Garcia-Perez 2017; Brouha et al. 2003; Beck et al. 2010). To retrotranspose, L1 generates a bicistronic mRNA encoding two proteins (ORF1p and ORF2p) that, amongst several key activities, catalyze genomic DNA nicking and reverse transcription of the L1 mRNA (Kazazian and Moran 2017; Moran et al. 1996; Scott and Devine 2017; Feng et al. 1996). New L1 insertions typically carry retrotransposition hallmarks, including target site duplications (TSDs) and a long 3′ poly(A) tract, and integrate at a degenerate 5′-TTTT/AA-3′ motif (Moran et al. 1996; Jurka 1997; Doucet et al. 2015). Numerous factors restrict L1 retrotransposition in somatic cells (Goodier 2016), including transcriptional silencing complexes recruited by DNA methylation (Castro-Diaz et al. 2014; de la Rica et al. 2016; Greenberg and Bourc'his 2019; Deniz et al. 2019; Ewing et al. 2020; Scott et al. 2016; Muotri et al. 2010; Thayer et al. 1993; Robbez-Masson et al. 2018). Embryonic L1 insertions can nonetheless arise prior to gastrulation (Richardson et al. 2017; van den Hurk et al. 2007; Feusier et al. 2019; Kano et al. 2009) and before the complete establishment of L1 methylation (Sanchez-Luque et al. 2019; Macia et al. 2017). Single-cell whole genome sequencing (scWGS) of post-mitotic neurons (Sanchez-Luque et al. 2019; Evrony et al. 2015) has uncovered somatic retrotransposition events traced to unusual donor (source) L1 loci that evade methylation in mature tissues (Sanchez-Luque et al. 2019; Ewing et al. 2020; Faulkner and Billon 2018). Despite relevant studies of endogenous retrotransposition in mouse and fly (Siudeja et al. 2021; Hazen et al. 2016; Richardson et al. 2017; Li et al. 2013; Keegan et al. 2021), scWGS has to date been applied only to human neurons generated *in vivo*, leaving the capacity of L1 to retrotranspose in non-human primate neuronal lineages an important open question. Starting with Haig Kazazian's first report of germline L1 mutagenesis (Kazazian et al. 1988), case studies have largely underpinned efforts to define the spatiotemporal extent of L1 mobility (Brouha et al. 2002; Scott et al. 2016; Sanchez-Luque et al. 2019; Evrony et al. 2015; de Boer et al. 2014; Miki et al. 1992; van den Hurk et al. 2007), founded on the principle that one robustly verified insertion is sufficient to show L1 retrotransposition can occur in a given context.

Because of its neuroanatomical, cognitive, social and genetic similarities with human, rhesus macaque (*Macaca mulatta*) is a cornerstone model organism for biomedical and neuroscience research (Feng et al. 2020; Phillips et al. 2014). L1 copies comprise 16% of the macaque reference genome, including ~44,000 annotated as rhesus-specific L1 (L1RS) sequences (Tang and Liang 2019; Warren et al. 2020). L1RS2 is the youngest and most active macaque L1 subfamily and incorporates 2,235 full-length (>6kbp) elements, far more than the equivalent human-specific L1 (L1HS) subfamily (329 full-length copies) (Warren et al. 2020). Several *Alu* short interspersed element (SINE) subfamilies, presumably retrotransposed *in trans* by L1 proteins (Dewannieux et al. 2003) are mobile in the macaque germline, as may be endogenous retroviruses (ERVs) (Han et al. 2007; Tang and Liang 2019; Warren et al. 2020; Liu et al. 2009). It is however unknown whether TEs retrotranspose in macaque neuronal lineage cells. Here, we exploited an updated macaque reference genome assembly, which greatly improved TE annotations (Warren et al. 2020), to profile retrotransposition in individual hippocampal neurons, and more broadly survey macaque TE transcription and repression *in vivo*.

Results

Genomic analyses of macaque retrotransposition

To explore TE mobilization in the macaque brain, we isolated RBFOX3$^+$ (NeuN$^+$) neuronal nuclei from the postmortem hippocampi of two animals (ON22212 and ON22213; both 7-year old male adults) and performed multiple displacement amplification (MDA) on each nucleus (Sanchez-Luque et al. 2019). We applied Illumina scWGS (average 40× genome-wide depth) to 20 neurons (7 from ON22212 and 13 from ON22213) that passed quality control, and performed Illumina WGS (average 41× depth) on matched bulk liver tissues (**Fig. 1A**; **Supplemental Table S1**). To further increase coverage of TE-genome junctions, we synthesized a new retrotransposon capture sequencing (RC-seq) (Baillie et al. 2011) probe pool targeting young macaque TE subfamilies (Warren et al. 2020; Han et al. 2007). Conceptually, this design most closely resembled one we previously developed in mouse (Richardson et al. 2017), and involved 80 densely-overlapping probes targeting the 5′ and 3′ termini of 8 TE subfamily consensus sequences (**Supplemental Table S1**). Barcoded Illumina libraries generated from each of the 20 MDA-amplified neuronal nuclei were hybridized to this probe pool, eluted, and then subjected to paired-end 2×150mer sequencing to maximize the number of RC-seq reads spanning a TE-genome junction (**Fig. 1A**).

Using the TEBreak computational pipeline (Carreira et al. 2016), we identified 194 L1, 3,348 *Alu* and no ERV non-reference insertions in the two liver WGS datasets (**Supplemental Table S2**). The *Alu* insertion count and frequency relative to L1 were each higher than what would be predicted based on prior human analyses, as expected (Tang and Liang 2019; Ewing et al. 2020). Of the 3,542 total events, 2,781 were present in one or the other animal, but not both. 189/194 (97.4%) non-reference L1s were annotated as belonging to the L1RS2 subfamily, and 44/194 (22.7%) were 5′ inverted (Ostertag and Kazazian 2001). L1 and *Alu* insertions were depleted from annotated protein-coding exons (**Fig. 1B**). Intronic L1 insertions were significantly ($p < 0.04$, binomial test) less abundant (61/194, 31.4%) than random expectation (43.6%) (**Fig. 1B**) and disproportionately few (23/61, 37.7%) of these were sense oriented to their host gene. Given modest genome-wide L1 integration site preferences, which mainly reflect the underlying distribution of AT-rich sequences, these patterns were likely dominated by post-integration selection, and are concordant with prior results obtained by human analyses (Sultana et al. 2019; Smits et al. 2021; Flasch et al. 2019; Ewing and Kazazian 2010; Attig et al. 2018; Smit 1999). Consistent with L1-mediated retrotransposition in humans and other mammals (Jurka 1997; Tang and Liang 2019; Moran et al. 1996; Richardson et al. 2017; Ewing et al. 2020; Smits et al. 2021), the L1 and *Alu* insertions generated TSDs with a median length of 15bp (**Fig. 1C**) and integrated at a motif strongly resembling the preferred L1 endonuclease motif (**Fig. 1C**). These analyses highlighted TE-driven genetic polymorphism amongst macaques, as well as the capacity of TEBreak to identify and characterize non-reference TE insertions in this model organism.

Next, we used the reference and non-reference L1RS2 insertions found in the bulk liver WGS datasets to estimate the detection sensitivity for potential somatic L1 insertions present in the 20 MDA-amplified neuronal genomes. For the reference analysis, we joined adjacent L1RS annotations, the majority of which represent 5′ inverted L1s that are often annotated as two oppositely oriented elements sharing a breakpoint, reducing the number of L1RS2 copies from 6,492 to 5,221. Of these, 3,200 (61.3%) and 3,113 (60.0%) were present in ON22212 and ON22213, respectively. Sensitivity was then recorded as a range, with the lower, more stringent, bound based on insertions being found by 5 or more reads at each of their 5′ and 3′ genome TE-genome junctions in a given neuron, and the upper bound only requiring at least one read at either junction. On average, 19.0%-63.6% of the reference L1RS2 copies were detected in the corresponding MDA-amplified neurons, including 10.4%-50.8% of heterozygous insertions. Of the aforementioned 189 non-reference L1RS2 insertions, 12.0%-44.9% were on average identified in the MDA-amplified neurons,

including 11.8%-42.7% of heterozygous elements (**Supplemental Table S2**). These results anticipated the false negative rate of our single-cell genomic analysis when applied to the discovery of somatic TE insertions.

A somatic L1RS insertion arising during brain development

Candidate TE insertions detected by scWGS or RC-seq and called stringently by TEBreak in at least one neuron from only one animal, and absent from the liver WGS, were considered provisional somatic events (**Fig. 1A**). With these parameters, we identified an intergenic somatic L1RS insertion, which we called $L1RS_{somatic}$, on Chromosome 4 of animal ON22213 neuron #15 (**Fig. 1D**; **Supplemental Table S2**). PCR followed by capillary sequencing recovered the 5′ and 3′ junctions of $L1RS_{somatic}$ in two of the ON22213 neurons analyzed by scWGS (#15 and #39), and in the matched bulk hippocampus (**Fig. 1E**). $L1RS_{somatic}$ was not detected by PCR in the liver of either ON22212 or ON22213. Amongst 59 additional MDA-amplified $RBFOX3^{+}$ neuronal nuclei, junction-specific PCRs identified $L1RS_{somatic}$ in neurons #29, #55 and #57 (**Supplemental Fig. S1A,B**). Hence, 5/72 (6.9%) of the tested ON22213 neurons, which were isolated from 3 separate hippocampus samples (**Supplemental Fig. S1A,B**), harbored $L1RS_{somatic}$. Junction PCRs amplified $L1RS_{somatic}$ in 2/4 additional ON22213 bulk hippocampus samples, and not in skeletal muscle, sciatic nerve, spinal cord or cerebellum (**Supplemental Fig. S2**). $L1RS_{somatic}$ therefore arose during central nervous system development, most likely in a neural progenitor cell located in the ventricular zone of the anterior (rostral) neural tube.

$L1RS_{somatic}$ was full-length, belonged to the L1RS2 subfamily, integrated at a 5′-TATT/AT-3′ motif, and was flanked by a 16bp TSD (**Fig. 1F**). These features were consistent with a *bona fide* L1 retrotransposition event (Jurka 1997; Moran et al. 1996). Capillary sequencing of the 3′ junction PCR products revealed a very long (>170nt) poly(A) tract at the 3′ end of $L1RS_{somatic}$ in neuron #15 (**Supplemental Fig. S1C**). As observed for somatic L1 insertions found by scWGS of human neurons (Evrony et al. 2015; Sanchez-Luque et al. 2019), the poly(A) tract of $L1RS_{somatic}$ varied substantially in length (~110bp to ~170bp) from neuron to neuron (**Fig. 1E**; **Supplemental Fig. S1B**). $L1RS_{somatic}$ was preceded by a 5′ untemplated guanine, as potentially associated with reverse transcription of an mRNA 5′ cap structure (Lavie et al. 2004; Gilbert et al. 2005), as well as a 4bp 5′ transduction (**Fig. 1F**).

We traced the 5′ transduced sequence (AGAG) to a putative L1RS2 donor element positioned in sense to intron 10 of the *PRDM4* gene on Chromosome 11 (**Fig. 1F**). We

termed this element L1RS$_{PRDM4}$. To characterize L1RS$_{somatic}$ and L1RS$_{PRDM4}$, we PCR amplified and fully capillary sequenced each element using template DNA from animal ON22213 (**Fig. 2A**; **Supplemental Table S2**). L1RS$_{somatic}$ and L1RS$_{PRDM4}$ were identical, apart from the much shorter 3′ poly(A) tract carried by L1RS$_{PRDM4}$ (**Fig. 2B**). Another candidate donor L1 (Chr4:107,868,275-107,874,430) closely matched L1RS$_{somatic}$ but lacked the adjacent 5′ AGAG sequence. Moreover, visual inspection of the aligned bulk liver WGS data indicated this element on Chromosome 4 was absent from ON22213. None of the non-reference full-length L1s detected by the ON22213 liver WGS (**Supplemental Table S2**) were preceded by a 5′ AGAG. These analyses very strongly linked L1RS$_{somatic}$ to L1RS$_{PRDM4}$ or, with lower probability, to a closely related but undetected non-reference donor L1 located perhaps in an unassembled genomic region (Zhou et al. 2020; Ewing et al. 2020).

WGS-based genotyping indicated that, as expected, L1RS$_{somatic}$ was heterozygous in ON22213 neuron #15, whereas L1RS$_{PRDM4}$ was homozygous in ON22213 liver. L1RS$_{PRDM4}$ was heterozygous in ON22212 liver. An analysis of primate reference genome assemblies indicated L1RS$_{PRDM4}$ was present in the closely related crab-eating macaque (*Macaca fascicularis*) and absent from the more evolutionarily distant southern pig-tailed macaque (*Macaca nemestrina*) and green monkey (*Chlorocebus sabaeus*), suggesting L1RS$_{PRDM4}$ entered the macaque germline 3-5 million years ago (Kumar et al. 2017; Kent et al. 2002; Springer et al. 2012). The two L1RS$_{PRDM4}$ alleles carried by ON22213 were identical and deviated from the macaque reference genome L1RS$_{PRDM4}$ element at a single 5′UTR position (A413G) and two ORF2 nucleotide positions: (A)2312A, which introduced to the reference sequence a premature ORF2p stop codon, and G2891A, a non-synonymous mutation of unclear significance for ORF2p activity (**Fig. 2B**). These analyses confirmed L1RS$_{PRDM4}$ and L1RS$_{somatic}$ encoded intact ORFs, while the reference L1RS$_{PRDM4}$ sequence did not, indicating L1RS$_{PRDM4}$ may be present and retrotransposition-competent in only some macaques.

Exceptional L1RS$_{PRDM4}$ retrotransposition in cultured cells

To assess the retrotransposition efficiency of L1RS$_{PRDM4}$, and therefore L1RS$_{somatic}$, we employed a quantitative cell culture-based retrotransposition assay (Moran et al. 1996; Kopera et al. 2016) where an L1 is tagged with an antibiotic selectable marker cassette only activated upon retrotransposition (**Fig. 2C**). Using this assay in HeLa cells, we found L1RS$_{PRDM4}$ mobilized >3-fold more efficiently than a highly active human L1HS element (L1.3) carrying the same marker cassette and employed as a positive control (Sassaman et al.

1997; Dombroski et al. 1993). No retrotransposition was detected for a negative control L1.3 disabled by an ORF2p reverse transcriptase mutation (D702A) (Moran et al. 1996).

We next tested L1RS$_{PRDM4}$ in cultured HEK293T cells using a related assay where, instead of an antibiotic selectable marker cassette, retrotransposition activates an enhanced green fluorescent protein (EGFP) marker, and L1 mobility is quantified via flow cytometry (Kopera et al. 2016; Ostertag et al. 2000). In HEK293T cells, L1RS$_{PRDM4}$ reproducibly mobilized >8-fold more efficiently than L1.3, while the L1.3 ORF2p reverse transcriptase mutant did not retrotranspose (**Fig. 2D**). An alignment of L1RS2 and L1HS consensus sequences (**Supplemental Fig. S3A**) indicated amino acid substitutions in both ORFs (**Supplemental Fig. S3B**), and particularly ORF1, as noted previously (Khazina and Weichenrieder 2018). To explore the disparate retrotransposition efficiencies of L1RS$_{PRDM4}$ and L1.3, we generated and tested a series of chimeric L1.3-L1RS$_{PRDM4}$ vectors in the HeLa- and HEK293T-based experimental assays. While interchanging either the 5′ or 3′ UTR of L1.3 and L1RS$_{PRDM4}$ in each system minimally impacted their mobility, replacing either L1RS$_{PRDM4}$ ORF with the corresponding L1.3 ORF severely reduced retrotransposition, compared to L1RS$_{PRDM4}$ (**Fig. 2C,D**). Next, we employed the antibiotic resistance cassette-based retrotransposition assay to test L1RS$_{PRDM4}$ and L1.3 activity in Chinese hamster V79B cells, where the expression of each L1 was ensured by a cytomegalovirus promoter (CMVp) element (**Fig. 2E**). L1RS$_{PRDM4}$ mobilized >2.2-fold more efficiently than L1.3. These results alluded to a functional interplay between L1RS$_{PRDM4}$ ORF1p and ORF2p that may be less relevant to human L1s, as L1RS$_{PRDM4}$ retrotransposed most efficiently when both of its native ORFs were present, and was far more mobile than L1.3 in human cells and in the more evolutionarily distant context of a rodent cell line. The retrotransposition competence of L1RS$_{PRDM4}$ was also consistent with its mobilization *in vivo*.

L1$_{PRDM4}$ methylation in adult tissues

The developmental origins of somatic mutations, including L1 insertions, can be inferred from their spatial distribution in adult tissues (Sanchez-Luque et al. 2019; Richardson et al. 2017; Evrony et al. 2015). Detection of L1RS$_{somatic}$ in bulk ON22213 hippocampus and a substantial fraction (~7%) of MDA-amplified hippocampal neurons, but not in other tissue samples, pointed to its integration at the outset of brain development but after formation of the neural tube (Stiles and Jernigan 2010). DNA methylation mediates L1 transcriptional silencing (Greenberg and Bourc'his 2019; Deniz et al. 2019; Thayer et al. 1993) and is relaxed amongst L1 promoters during early embryogenesis (Sanchez-Luque et al. 2019;

Coufal et al. 2009; Macia et al. 2017). Notably, neuronal and non-neuronal L1 insertions occurring later in human development have been traced to donor L1s escaping methylation even in mature somatic cells (Ewing et al. 2020; Sanchez-Luque et al. 2019; Scott et al. 2016; Evrony et al. 2015). On this basis we hypothesized L1RS$_{PRDM4}$ was aberrantly demethylated in the hippocampus. To test this possibility, and evaluate TE methylation genome-wide, we applied Oxford Nanopore Technologies (ONT) long-read sequencing to ON22213 bulk hippocampus and liver tissue (**Supplemental Table S1**). Examining the *PRDM4* locus, we found the *PRDM4* promoter was fully unmethylated, whereas the promoter and body of L1RS$_{PRDM4}$ were near-completely methylated (**Fig. 3A**). We confirmed these results with locus-specific bisulfite sequencing (**Fig. 3B,C**), and concluded L1RS$_{PRDM4}$ did not escape methylation in adult tissues.

The expression of an intronic donor L1 may be influenced by the activity of its host gene (Philippe et al. 2016), as per an L1HS element located in the human *TTC28* gene that is highly mobile in epithelial cancers (Sanchez-Luque et al. 2019; Tubio et al. 2014). *PRDM4* is strongly expressed in mammalian embryonic cells, and later downregulated as a catalyst for neuronal differentiation (Bogani et al. 2013; Chittka et al. 2012). We therefore compiled published RNA-seq transcriptome profiling data from various stages of early macaque development, including metaphase I and II oocytes, six stages of pre-implantation embryogenesis, and adult hippocampus (Wang et al. 2017; Yin et al. 2020). This analysis indicated high *PRDM4* expression was maintained until the 8-cell stage and was followed by an 85% reduction at the morula (16-cell) stage (**Fig. 3D**). *PRDM4* nonetheless was expressed in the hippocampus (**Fig. 3D,E**), and in neural stem cells generated *in vitro* (**Fig. 3E**) (Zhao et al. 2014). We concluded L1RS$_{PRDM4}$ was positioned in a genomic locus likely transcribed throughout embryogenesis and brain development, when L1RS$_{somatic}$ arose, despite near-complete methylation of the L1RS$_{PRDM4}$ promoter in the mature hippocampus.

Dynamic TE expression during macaque development

Human pluripotent cells support endogenous L1 demethylation, transcription and mobilization (Sanchez-Luque et al. 2019; Macia et al. 2017; Klawitter et al. 2016; Garcia-Perez et al. 2007). Although accurate TE locus-specific measurement of transcription with short-read RNA-seq is extremely challenging (Lanciano and Cristofari 2020), this approach can be used to quantify expression of TE subfamilies genome-wide (Faulkner et al. 2008; Hashimoto et al. 2009; Faulkner et al. 2009). We therefore used the RNA-seq datasets described above to temporally profile the transcript abundance of a focused cohort of TE

subfamilies, selected to represent the LINE, SINE and ERV superfamilies. These were: L1RS2, L1PA5 (mobile in the last common macaque-human ancestor, and now immobile), *Alu*YRa1 (the most numerous macaque *Alu*Y element) and MacERV1 (a young, horizontally transferred macaque ERV) (Han et al. 2007; Warren et al. 2020). As controls, we re-analyzed published human (Zhang et al. 2019) and mouse (Macfarlan et al. 2012) RNA-seq datasets and faithfully recapitulated the associated conclusions of abundant human endogenous retrovirus-H (HERVH) and murine endogenous retrovirus-L (MERVL) expression, respectively, in pre-implantation embryonic cells (Macfarlan et al. 2012; Peaston et al. 2004; Svoboda et al. 2004; Zhang et al. 2019; Grow et al. 2015) (**Supplemental Fig. S4**). Examining macaque TE subfamily expression with two computational pipelines (Hashimoto et al. 2009; Faulkner et al. 2008; Jin et al. 2015), we noted L1RS2 was consistently more highly expressed than L1PA5 throughout development, particularly at the 8-cell and morula stages, whereas *Alu*YRa1 expression lagged slightly behind, peaking at the morula and blastocyst stages (**Fig. 4A**; **Supplemental Fig. S5**). By contrast, MacERV1 displayed a 17-fold increase in expression between metaphase II oocytes and the 2-cell stage, as seen for MERVL in mouse (Macfarlan et al. 2012), and was lowly expressed from the morula stage onwards (**Fig. 4A**). The widespread occurrence of TEs within introns and immediately downstream of protein-coding genes can make readthrough and independent TE transcription difficult to distinguish with short-read RNA-seq (Lanciano and Cristofari 2020). However, closely examining individual L1RS2 and *Alu*YRa1 loci, we found the most strongly expressed elements tended to be intergenic (**Supplemental Fig. S6A**) or, if adjacent to a protein-coding gene, exhibit more temporally-restricted expression than that gene (**Supplemental Fig. S6B,C**). These RNA-seq analyses altogether highlighted L1RS2 and *Alu*YRa1 transcriptional activation in the macaque embryo, distinct to that of other TEs, and consistent with the *in vivo* timing of endogenous retrotransposition events traced elsewhere to human and mouse embryogenesis (van den Hurk et al. 2007; Feusier et al. 2019; Richardson et al. 2017).

Macaque TE methylome landscapes

Exceptions to DNA methylation at specific donor L1 loci appear to facilitate somatic retrotransposition in humans (Faulkner and Billon 2018). However, it was unclear whether similar “escapee” L1s reside in macaque methylomes, especially as L1RS$_{PRDM4}$ was here heavily methylated in the hippocampus. We therefore analyzed our macaque ONT sequencing data with MethylArtist (Cheetham et al. 2022) to survey TE subfamily

methylation genome-wide and at individual TE loci. We observed median CpG methylation values of 83.3% and 75.0% for L1RS2 copies in hippocampus and liver, respectively, with these values substantially lower than those for *Alu*YRa1 (94.0% and 92.9%) (**Fig. 4B**). L1RS2 was modestly (~3%) more methylated than L1PA5 in each tissue. Of the TE subfamilies analyzed, MacERV1 elements were the most variably methylated (**Fig. 4B**). These trends largely aligned with those observed for the approximately equivalent TE subfamilies in human tissues, where TE methylation in the hippocampus was also generally higher than in liver (Ewing et al. 2020). Profiling methylation across full-length (>6kbp) L1RS2 copies, we observed a trough within the 5′UTR (**Fig. 4C**). This trough was however less pronounced than the one identified for the human L1HS subfamily (Ewing et al. 2020), perhaps owing to the differing 5′UTR CpG densities of L1RS2 (2.7 CpGs per 100bp) and L1HS (4.3 CpGs per 100bp). We found 88 L1RS2, 22 L1PA5, 2 MacERV1 and 176 *Alu*YRa1 copies differentially methylated ($p < 0.05$, Fisher's exact test with Bonferroni correction) in hippocampus compared to liver, with most being less methylated in the latter tissue (**Fig. 4B,D**; **Supplemental Fig. S7A**; **Supplemental Table S3**). As well, 7 L1RS2, 2 L1PA5, 1 MacERV1 and 76 *Alu*YRa1 copies were less than 50% methylated in both hippocampus and liver (**Supplemental Fig. S7B**; **Supplemental Table S3**). These results indicated that, while the vast majority of young TEs were repressed, a handful were unmethylated in a subset of macaque brain or liver cells.

Discussion

Endogenous L1 retrotransposition requires a complex series of molecular steps to be completed amidst the host genome defenses maintained by somatic and germ cells (Scott and Devine 2017; Goodier 2016). We show here that the cellular circumstances leading to L1 mobility can nonetheless come about during macaque brain development, as in human (Sanchez-Luque et al. 2019; Evrony et al. 2015; Erwin et al. 2016). That this mechanism may be evolutionarily conserved is notable, given these species diverged nearly 30 million years ago (Kumar et al. 2017), as did their mobile L1 subfamilies and host defense pathways. We speculate the L1PA5 common ancestor of the youngest macaque (L1RS2) and human (L1HS) subfamilies (Warren et al. 2020) was similarly able to retrotranspose in the neuronal lineage, and this potential was inherited by other primate L1 subfamilies.

$L1RS_{somatic}$ bore a striking resemblance to the three somatic L1 insertions characterized to date via scWGS of human neurons (Sanchez-Luque et al. 2019; Evrony et al. 2015). Each of these four events generated TSDs of 13-20bp, incorporated 3′ poly(A) tracts

of ~90bp to ~170bp, integrated at a degenerate L1 endonuclease motif, and, via transductions, were traced to mobile donor L1s. These sequence features are congruent with the mechanistic model of L1 target-primed reverse transcription (Moran et al. 1996; Jurka 1997). As per the three human insertions, $L1RS_{somatic}$ was detected in multiple post-mitotic neurons, where its poly(A) tract varied considerably in length, consistent with asymmetric poly(A) microsatellite shortening in the clonal lineages giving rise to the hippocampus (Grandi et al. 2013; Evrony et al. 2015; Sanchez-Luque et al. 2019). The detection of multiple $L1RS_{PRDM4}$ alleles reinforces prior findings relating to retrotransposition-competent L1 alleles in the human brain (Sanchez-Luque et al. 2019) and germline (Lutz et al. 2003; Seleme et al. 2006). These data indicate neurodevelopmentally-active primate donor L1s can be polymorphic, and include both mobile and immobile alleles.

5′ and 3′ transductions are carried by <1% and <10%, respectively, of new human germline L1 insertions (Ewing et al. 2020; International Human Genome Sequencing Consortium 2001; Gardner et al. 2017). By contrast, all four human and macaque somatic L1s identified to date with scWGS incorporated a 5′ (2) or 3′ (2) transduction. The reasons for this apparent bias however remain unresolved. The untemplated guanine preceding the 4bp $L1RS_{somatic}$ 5′ transduction (Lavie et al. 2004), and the presence of a pyrimidine/purine initiator dinucleotide (Sandelin et al. 2007) at the corresponding transcription start site upstream of $L1RS_{PRDM4}$, together indicate the mRNA template for $L1RS_{somatic}$ could have been transcribed at the direction of the canonical $L1RS_{PRDM4}$ 5′UTR promoter, then capped and reverse transcribed without 5′ truncation. The $L1RS_{PRDM4}$ promoter nonetheless provides the main difference with the three human insertions; these were each traced (Sanchez-Luque et al. 2019) to a donor L1, or an upstream promoter, demethylated in brain tissue, while $L1RS_{PRDM4}$ was near-completely methylated. We propose the following scenarios, in order of decreasing likelihood, to explain $L1RS_{somatic}$ in the face of $L1RS_{PRDM4}$ promoter methylation in hippocampus: (*i*) $L1RS_{PRDM4}$ was hypomethylated and transcribed in the neural progenitor cell giving rise to $L1RS_{somatic}$, (*ii*) the requisite $L1RS_{PRDM4}$ mRNA was carried forward from earlier embryonic development, (*iii*) $L1RS_{PRDM4}$ was transcribed as part of a chimeric mRNA initiated by the demethylated *PRDM4* promoter and 5′ truncated to remove almost all of the upstream *PRDM4* exons, or (*iv*) DNA methylation does not as strongly underpin macaque L1RS2 transcriptional repression as it does human L1HS repression.

$L1RS_{PRDM4}$ is the first endogenously mobile non-human primate L1, to our knowledge, to be tested in a cultured cell retrotransposition assay. Its natural mobility *in vitro* was high: >3-fold, >8-fold and >2.2-fold more than the positive control L1.3 in HeLa,

HEK293T and V79B cells, respectively. Adaptive evolution involving strong positive selection of amino acid substitutions has been observed amongst primate L1 protein domains (Khan et al. 2006; Boissinot and Furano 2001; Furano et al. 2020; Wagstaff et al. 2013; Khazina and Weichenrieder 2018). However, as opposed to its individual ORF1p or ORF2p activities or 5′UTR promoter strength, the efficiency of L1RS$_{PRDM4}$ appeared due to ORF1p-ORF2p synergy. We speculate that evolutionary changes in ORF1p-ORF2p epistatic interactions (Wagstaff et al. 2011) occurred after the divergence of human and macaque and, as a result, increased L1RS2 retrotransposition efficiency. Another possible explanation is that these interactions were supported by the ancestral L1PA5 subfamily and weakened, or were lost, during the later evolution of the L1HS lineage. Either rationale is supported by the relative retrotransposition efficiencies of L1RS$_{PRDM4}$ and L1.3 in cultured rodent V79B cells, where host factor interactions specific to the RNA or proteins of either element are presumably absent. Of the young macaque L1 subfamilies, L1RS2 has the lowest average sequence divergence (1.1%) and the highest proportion of full-length (>6kbp) elements (Warren et al. 2020). Nearly 7-fold more full-length L1RS2 copies than L1HS copies are annotated in the respective reference genomes and these also make up a higher proportion of the elements in that subfamily (L1RS2: 34.4%, L1HS: 19.3%), despite L1RS2 modestly predating the emergence of L1HS (Khan et al. 2006; Warren et al. 2020). These differences, as well as the exceptional *in vitro* mobility of L1RS$_{PRDM4}$, imply endogenous L1 retrotransposition potential may presently be higher in macaque than in human.

Single-cell analyses now provide conclusive evidence of L1-mediated somatic mosaicism in the macaque and human brain. Endogenous retrotransposition is also likely encountered in mouse and fly neuronal lineages (Siudeja et al. 2021; Hazen et al. 2016; Li et al. 2013; Keegan et al. 2021; Coufal et al. 2009; Muotri et al. 2005). A major limitation of scWGS is the generation of MDA and Illumina library preparation false positives (Treiber and Waddell 2017; Faulkner and Garcia-Perez 2017; Abascal et al. 2021). The stringent approach adopted here and elsewhere (Evrony et al. 2015; Sanchez-Luque et al. 2019), requiring complete resolution of L1RS$_{somatic}$ and its associated retrotransposition hallmarks, excludes false positives with near certainty, but also raises the prospect of false negatives. On average, only 11.8% of the heterozygous non-reference L1RS2 insertions present in our bulk liver WGS datasets were found, using robust thresholds, in the corresponding neuronal genomes. While L1RS$_{somatic}$ was detected by PCR at its 5′ or 3′ genomic junction in 5/72 MDA-amplified neurons from ON22213, and in bulk hippocampus, the complete insertion could only be PCR amplified in neuron #15. Furthermore, we analyzed a pan-neuronal

(RBFOX3$^+$) hippocampal population, which could obscure potential sublineage-specific L1 activity (Faulkner and Garcia-Perez 2017; Bodea et al. 2022). These considerations in our view prohibit an accurate calculation of L1 mobilization frequency in the macaque brain, extrapolated from one *bona fide* somatic L1RS2 insertion. We did not identify any somatic *Alu* insertions, concordant with prior human neuron scWGS analyses (Sanchez-Luque et al. 2019; Evrony et al. 2015). We have also not included an analysis of somatic single nucleotide variants, because of the potential difficulties in distinguishing these from MDA artifacts present in scWGS data (Abascal et al. 2021). Finally, while the impact of somatic retrotransposition upon brain development remains hypothesized (Erwin et al. 2014; Muotri and Gage 2006), its apparent occurrence amongst multiple primate species, and likely other animals, may inform future studies testing the association of L1 mobility or expression with cellular or cognitive functions.

Methods

Macaque samples

Snap frozen hippocampus and liver tissue from two post-mortem macaques (identifiers ON22212 and ON22213) without evidence of pathology was provided by the Monkey Alcohol Tissue Research Resource (MATRR) biobank (https://gleek.ecs.baylor.edu/) to L.C. with ethical approval to be used as described previously (Daunais et al. 2014). ON22212 and ON22213 were 7-year old male adults, bred in the same animal research facility, and without parents or grandparents in common.

Isolation and whole genome amplification of neuronal nuclei

Single neurons were isolated from hippocampal tissue and genomic DNA amplified via multiple displacement amplification (MDA) as previously described (Sanchez-Luque et al. 2019). Reagents were pre-chilled and the entire procedure performed on ice. Frozen hippocampus samples were first gently Dounce homogenized for 2min in 2mL cold nucleus extraction buffer composed of 10mM Tris (pH 7.4), 10mM NaCl, 3mM $MgCl_2$, and 0.1% IGEPAL CA-630 (Sigma-Aldrich). Tissue homogenates were then filtered into a 5mL tube with a 40µm cell-strainer cap and centrifuged at 500g for 2min at 4°C. Following centrifugation, pellets were resuspended in a wash buffer of 1% bovine serum albumin (BSA, Sigma-Aldrich, A2153) in PBS. To tag neuronal nuclei, anti-RBFOX3 (Merck-Millipore, MAB377X) antibodies and DAPI (Sigma-Aldrich, D9542) were added to the solution and incubated for 15min at 4°C. Nuclei were spun down as above and re-suspended in 1× PBS.

DAPI^{+}/RBFOX3^{+} nuclei were sorted using a BD FACSAria Cell Sorter (Becton Dickinson) in a block buffer (10% goat serum and 5% BSA). Purified nuclei were then picked using an Olympus IX71 inverted microscope, with an Eppendorf TransferMan 2 micromanipulator and Eppendorf CellTram. During picking, single nuclei were washed in PBS, and transferred to individual UV sterilized 0.2mL PCR tubes. MDA was then performed upon each nucleus using a REPLI-g Single Cell Kit (Qiagen, 150345). First, nuclei were incubated at 65°C for 10min in 3µL buffer D2 and then placed on ice with 3µL Stop Solution. DNA was amplified at 30°C for 8hrs with 1× sc Reaction Buffer, phi29 DNA polymerase, and nuclease-free UV-treated water, for a final volume of 40µL. The polymerase was then inactivated at 65°C for 3min. MDA-amplified DNA clean-up was performed with 1:1.3 (v/v) ratio AMPure XP beads (Beckman Coulter, A63881) immediately before the de-branching step. To screen nuclei for those with the most even genome-wide amplification, multiplexed PCR primers were designed for a panel of 12 non-repetitive loci (**Supplemental Table S1**). For each nucleus, three reactions were undertaken on a DNA Engine Tetrad 2 Thermal Cycler (Bio-Rad), with MyTaq HS DNA polymerase (Bioline, BIO-21105), 5× MyTaq Reaction Buffer, 10µL of a 25mM primer mix containing 4 primer pairs, 12ng template DNA and 0.25U of enzyme in a 25µL final volume. PCR cycling conditions were as follows: (95°C, 1min)×1; (95°C, 15s; 58°C, 15s; 72°C, 15s)×35; (72°C, 5min; 4°C, hold)×1. Amplicons were visualized via GelDoc (Bio-Rad) on a 1.5% agarose gel stained with SYBR Safe (Invitrogen, S33102). Twenty nuclei (7 from ON22212 and 13 from ON22213) where at least 9/12 genomic loci amplified were selected for further analysis.

Illumina sequencing

Genomic DNA was extracted from ON22212 and ON22213 bulk liver tissue. Libraries were prepared for these samples, as well the material from the 20 MDA-amplified neuronal nuclei, using a TruSeq DNA PCR-Free Kit (Illumina, 20015963) and an insert size of 550bp. Each library was subjected separately to paired-end 2×150mer WGS using an Illumina HiSeq X platform (Macrogen, South Korea).

For RC-seq, DNA from the MDA-amplified neuronal nuclei was used to prepare barcoded libraries with a TruSeq Nano DNA Kit (Illumina, FC-121-4001/2). Briefly, 4µg of MDA-amplified DNA was diluted to 130µL final volume and sheared in a Covaris M220 Focused-Ultrasonicator (peak power 50, duty factor 20, pulses per burst 200) for 110s in MicroTube AFA Snap-Cap tubes (Covaris, 520045). DNA was purified via AMPure XP bead clean-up using a 1:1 volume of beads, and eluting in 60µL of resuspension buffer. The

TruSeq Nano protocol was then followed as indicated by the manufacturer until the tandem clean-up after the adaptor ligation step. At this stage, samples were instead suspended in 20µL of resuspension buffer and loaded on a 2% high-resolution agarose gel (Sigma-Aldrich) for visualization via electrophoresis. Size selection was achieved by purifying gel cuts of 600-650bp size, which were eluted using a MinElute Gel Extraction Kit (Qiagen, 28604). QG buffer was added at a ratio of 600µL per 0.1mg of gel cut and the agarose dissolved at room temperature. Elution was performed using 12.5µL of 60°C pre-heated EB buffer twice, for a final 25µL elution volume. Library amplification was performed using 1× Phusion High-Fidelity PCR Master Mix (New England Biolabs) with 100pmol of each Illumina primer in a 100µL final volume. Cycling conditions were as follows: (98°C, 45s)×1; (98°C, 15s; 60°C, 30s; 72°C, 30s)×7; (72°C, 5min; 4°C, hold)×1. Samples were purified by AMPure XP beads clean up using 1:1 ratio of DNA to beads, eluted in 30µL of molecular grade water and quantified using a Bioanalyzer DNA 1000 chip (Agilent Technologies) according to the manufacturer's instructions. Libraries were pooled in equimolar amounts and the pool hybridized as per a prior study (Richardson et al. 2017) to a custom sequence capture probe pool (Roche NimbleGen) targeting the L1RS2, L1RS37, *Alu*YRa4, *Alu*YRb4, *Alu*YRc2, *Alu*YRd4, LTR14 (HERVK14) and LTR4 (MacERV1) subfamilies (**Supplemental Table S1**). The post-hybridization library pool was then sequenced with paired-end 2×150mer reads, using two flow cells of an Illumina HiSeq X platform (Macrogen, South Korea). This was intended to achieve a similar post-enrichment sequencing depth per targeted element as the WGS, while minimizing read duplicates caused by saturating the comparatively narrower TE-junction target windows generated by the TE enrichment probes.

Non-reference TE insertion detection

Bulk liver WGS data were aligned to the rheMac10 reference genome with BWA-MEM (Li 2013) (optional parameters -M -Y). Alignments were then analyzed with TEBreak (Carreira et al. 2016) using default parameters, where the Repbase consensus sequences for the L1RS2, L1RS37, L1PA5, *Alu*YRa4, *Alu*YRb4, *Alu*YRc2, *Alu*YRd4 and LTR4 (MacERV1) subfamilies were used to annotated potential insertions caused by young TEs. The TEBreak output table was then parsed to retain only putative non-reference insertions detected by at least 5 reads spanning each of their (5′ and 3′) TE-genome junctions, and remove insertions that were outside of canonical assembled macaque chromosomes (Chr1-20, X, Y), or were 3′ truncated or, for *Alu*, were 5′ inverted or 5′ truncated by more than one nucleotide. Insertions were further stratified as homozygous (variant allele fraction ≥ 0.8) or heterozygous (variant

allele fraction < 0.8) with the number of reads spanning the annotated (empty) insertion point providing the denominator. To identify putative somatic TE insertions, the neuron scWGS and RC-seq datasets were similarly aligned, added to and processed together with the bulk liver WGS with TEBreak. The TEBreak output table was filtered as before, except with the additional requirement that any events called in either liver sample, or in neurons from more than one animal, were removed. The resulting filtered tables listed 3,543 non-reference TE insertions, including $L1RS_{somatic}$, and are presented as **Supplemental Table S2**.

$L1RS_{somatic}$ junction PCR validation experiments

We designed PCR primers to amplify the 5′ and 3′ L1-genome junctions of $L1RS_{somatic}$ (**Supplemental Table S2**) with Primer3 (Untergasser et al. 2012). Reactions were undertaken on a DNA Engine Tetrad 2 Thermal Cycler (Bio-Rad), with MyTaq HS DNA polymerase, 5× MyTaq Reaction Buffer, 10pmol of each primer, 5ng of template DNA, and 2.5U of enzyme, in a 25μL final volume. 5′ junction PCR cycling conditions were as follows: (95°C, 1min)×1; (95°C, 15s; 59°C, 15s; 72°C, 15s)×40; (72°C, 5min; 4°C, hold)×1. 3′ junction nested PCR involved two reactions, the first with cycling conditions of (95°C, 1min)×1; (95°C, 15s; 59°C, 15s; 72°C, 15s)×15; (72°C, 5min; 4°C, hold)×1, followed by sample treatment with ExoSAP-IT PCR Product Cleanup (Thermo Fisher Scientific, 75001.1.ML) (37°C, 15min; 80°C, 15min) and a second reaction with cycling conditions of (95°C, 1min)×1; (95°C, 15s; 59°C, 15s; 72°C, 15s)×30; (72°C, 5min; 4°C, hold)×1. All PCRs were performed with non-template control (NTC), each MDA-amplified DNA from animal ON22213 hippocampal neurons, as well as DNA extracted from animals ON22212 and ON22213 bulk tissues. Amplicons were visualized via GelDoc (Bio-Rad) on a 1.5% agarose gel stained with SYBR Safe. GeneRuler 1kb plus (Thermo Fisher Scientific, SM1331) was used as the ladder. Amplicons of the correct size were gel extracted using a Qiagen MinElute Gel Extraction Kit and cloned with a pGEM-T Easy Vector system (Promega, A1360) using One Shot TOP10 chemically competent *E. coli* cells (Thermo Fisher Scientific, C404010) prior to capillary sequencing by Macrogen. For each $L1RS_{somatic}$ 3′ junction amplicon, the L1 poly(A) size was estimated by taking the average of the pure poly(A) (or polyT) tract lengths observed by capillary sequencing in both directions, as per Supplemental Fig. S1C. Note that this approach provided a lower bound estimate for bulk or pooled input DNA, owing to polymerase slippage and the varying template poly(A) tract lengths amongst the neurons carrying $L1RS_{somatic}$.

PCR amplification of L1RS$_{somatic}$ and L1RS$_{PRDM4}$

To test the retrotransposition efficiency of L1RS$_{PRDM4}$, we designed PCR primers specific to L1RS$_{somatic}$ (**Supplemental Table S2**) by placing a forward primer, incorporating a NotI restriction site to facilitate later cloning, across the 5′ L1-genome junction and a reverse primer in the 3′ L1-genome flanking region. PCR was performed using an Expand Long Range dNTPack kit (Sigma Aldrich, 11681834001) in 1× reaction buffer with $MgCl_2$, 0.5mM of each dNTP, 3% DMSO, 10pmol of each primer, 1.75U of enzyme, and 10ng of ON22213 hippocampal neuron #15 MDA-amplified DNA template in a 25μL final volume. PCR cycling conditions were as follows: (92°C, 2min)×1; (92°C, 10s; 58°C, 15s; 68°C, 6min)×10; (92°C, 10s; 58°C, 15s; 68°C, 6min+20s/cycle)×30; (68°C, 10min; 4°C hold)×1. The L1RS$_{PRDM4}$ donor element for L1RS$_{somatic}$ was amplified by forward (5′-GGACAGTAGGCGGAGTTGAG-3′) and reverse (5′-CCACCATGCCCAGTCTACTT-3′) primers placed in the 5′ and 3′ genomic flanks of L1RS$_{PRDM4}$, respectively, with the same reaction conditions as used to amplify L1RS$_{somatic}$, except using 10ng of animal ON22213 liver DNA template. PCR products were resolved by electrophoresis on a 1% agarose gel and imaged with a Typhoon FLA 9500 Scanner (GE Healthcare Life Sciences). PCR bands of the appropriate size were excised and purified via conventional phenol:chloroform DNA extraction followed by ethanol precipitation. PCR products containing L1RS$_{somatic}$ and L1RS$_{PRDM4}$ were cloned in a TOPO XL PCR cloning kit (Life Technologies, K8050-10) using One Shot TOP10 Electrocomp *E. coli* cells (Thermo Fisher Scientific, C404050). PCR products and TOPO XL clones were capillary sequenced using stepping primers to resolve the complete sequence of each L1, and identify potential allelic variants within the two L1RS$_{PRDM4}$ alleles, using a previous approach (Sanchez-Luque et al. 2019). The two alleles of L1RS$_{PRDM4}$ and the sequence of L1RS$_{somatic}$ were found to be identical (**Supplemental Table S2**).

Cultured cell retrotransposition assays

The L1RS$_{somatic}$ sequence was cloned into three pCEP4-derived vectors to assay its retrotransposition efficiency, based on a prior strategy applied to human L1s (Sanchez-Luque et al. 2019). The first and second vectors contained a neomycin resistance cassette (*mneoI*) driven by a simian virus 40 early promoter (SV40p) and terminated by a herpes simplex virus (HSV)-thymidine kinase polyadenylation signal, and positioned downstream of, and in reverse orientation to, the L1 but was interrupted by an intron in the same orientation as the L1, meaning the cassette was only activated by retrotransposition (Moran et al. 1996). The

second vector included a CMVp upstream of the L1 to ensure its transcription (Moran et al. 1996). The third vector was similar in structure but lacked the upstream CMVp and, instead of the *mneoI* cassette, contained an enhanced green fluorescent protein (EGFP) retrotransposition reporter cassette (*mEGFPI*) driven by CMVp. Also in this vector the original pCEP4 hygromycin resistance marker was replaced by a puromycin resistance gene for selecting transfected cells (Ostertag et al. 2000). Each of these vectors was originally designed to clone L1 sequences between NotI and BstZ17I restriction sites, lacking a small fragment downstream of the BstZ17I site in the L1 3′UTR (Ostertag et al. 2000; Moran et al. 1996). We previously restored the L1 3′UTR, only deleting the thymine base within the natural polyadenylation signal to still allow the retrotransposition cassette to be incorporated into the L1 mRNA (Sanchez-Luque et al. 2019). The cloning strategy here broadly involved rebuilding the $L1RS_{somatic}$ sequence into the vector from several TOPO XL clone segments, avoiding clone-specific PCR mutations, and similarly altering the L1 polyadenylation signal as for the human L1.3 controls. We took advantage of the BstZ17I site in the $L1RS_{somatic}$ 3′UTR (conserved with human L1HS) to first engineer the 3′ end of $L1RS_{somatic}$ downstream of the BstZ17I site into the vector, without the polyadenylation signal. We amplified the ~80bp fragment downstream of the BstZ17I site from a TOPO XL clone without PCR mutations by using the primers (5′-GGAAGATCTCTAGCGGCCGCATGTATACATATGTAACAAACCTGCACGTTATGCACA-3′) and (5′-GAGATTTAAATTTTTTTTTTTTTTATACTTTAAGTTGTAGGGTACATG-3′). This reaction generated a 104bp amplicon containing a NotI restriction site upstream of the BstZ17I site and a SwaI restriction site downstream of a 15bp polyadenine tract, and lacking the polyadenylation signal. PCR was performed using the Q5 High-Fidelity DNA Polymerase (New England Biolabs, M0491S) in a reaction containing 1×Q5 Reaction Buffer, 0.2mM of each dNTP, 20pmol of each primer, 1U of enzyme and ~50ng of input DNA in a 25µL final volume. Cycling conditions were as follows: (98°C, 30s)×1; (98°C, 10s; 58°C, 15s; 72°C, 30s)×30; (72°C, 2min; 4°C hold)×1. The resulting fragment was digested with NotI and SwaI restriction enzymes and cloned into the NotI and BstZ17I sites of the three aforementioned original vectors lacking the human L1 3′UTR downstream of the BstZ17I site (Ostertag et al. 2000; Moran et al. 1996). The remaining $L1RS_{somatic}$ sequence was reconstructed between the NotI site and the new BstZ17I site. Plasmid DNA vectors were produced using a Qiagen Plasmid Midi Kit (12143).

Engineered L1 retrotransposition experiments were performed in HeLa, HEK293T and V79B Chinese hamster lung fibroblast cells, following previously described guidelines

(Sanchez-Luque et al. 2019). For the HeLa neomycin resistance cassette-based reporter assay (Moran et al. 1996), HeLa JVM cells were seeded into 6-well plates at a density of 5×10^3 cells/well in 2mL of Dulbecco's Modified Eagle Medium (DMEM, Life Technologies, 11965092), 10% Fetal Bovine Serum (FBS, Life Technologies, 10082147), 2mM L-Glutamine (Life Technologies, 25030081) and 100U/mL Penicillin-Streptomycin solution (Pen-Strep, Life Technologies, 15140122) per well. Cells were incubated at 37°C, 5% CO_2 and ~95% humidity for the course of the experiment. Transfection was performed ~14hr after seeding by adding 100µL of transfection mix to each well, which contained 1µg of plasmid DNA, 96µL of Opti-MEM (Life Technologies, 31985070) and 4µL of FuGENE-HD (Promega, E2311). Plates were shaken gently to homogenize the transfection mix. Technical replicates were plated from the same cell suspension and transfected with the same transfection master mix. Media were replaced with 2mL of complete media 24hr after transfection, and then replaced by complete media supplemented with 400µg/mL of G418 sulphate (Geneticin Selective Antibiotic, Life Technologies, 10131035) every 48hr for a total of 12 days. On day 14, media were aspirated and each well washed with 1-2mL of Dulbecco's Phosphate Buffered Saline (DPBS, Life Technologies, 14190144). After aspirating the plates, cells were fixed by adding 1mL of 1× DPBS, 0.2% glutaraldehyde and 2% formaldehyde solution, and incubating at room temperature for 20min. The fixing solution was discarded and the wells carefully washed with reverse osmosis-purified (RO) H_2O. Cell colonies were stained by adding 1mL of 0.1% crystal violet solution to each well and incubating at room temperature for 10min. The dying solution was discarded and the plates washed with RO H_2O and air-dried before scanning. Plates were imaged using a Canon EOS Rebel T3 camera and a white light transilluminator.

The neomycin resistance cassette-based reporter assay in V79B cells was adapted from conditions used for similar experiments using Chinese hamster ovary (CHO) cells (Morrish et al. 2007). This was performed as described above for HeLa cells but using a seeding density of 2×10^4 cells/well and Dulbecco's Modified Eagle Medium with 1g/L glucose (DMEM low glucose, Life Technologies, 11885084), 10% FBS (Life Technologies, 10082147), 2mM L-Glutamine (Life Technologies, 25030081), 1× Non-essential Amino Acids (100× NEAA, Life Technologies, 11140050) and 100U/mL Penicillin-Streptomycin solution (Pen-Strep, Life Technologies, 15140122) as culture media for the course of the experiment. Media changes and antibiotic selection, respectively, were performed with the timing and G418 sulphate concentration (400µg/mL) used for HeLa cells.

For the enhanced green fluorescent protein (EGFP) cassette-based reporter assay

(Ostertag et al. 2000), experiments were performed using HEK293T cells. In 6-well plates, 2×10^5 cells were seeded per well in media composed of 2mL DMEM, 10% FBS, 2mM L-Glutamine and 100U/mL Pen-Strep solution. Transfection was performed ~14hr after seeding by adding a similar transfection mix as for HeLa cells above. Again, technical replicates were seeded from the same cell suspension and transfected with the same transfection master mix. 24hr post-transfection, media were replaced and supplemented with 0.5µg/mL puromycin (Puromycin Dihydrochloride, Life Technologies, A1113803). This was then repeated daily for 4 more days, but with media supplemented with 1µg/mL of puromycin. On day 6, cells were washed with DPBS and incubated with 0.5mL of Trypsin-EDTA 0.25% at 37°C for 5min. Trypsinization was stopped by adding 1mL of DPBS with 10% FBS to each well. Cells were resuspended by pipetting, transferred to a 1.5mL tube and centrifuged at 4°C, 450g for 5min. Supernatant was aspirated and cell pellets resuspended in 300µL of 4°C 1× DPBS. Cells were analyzed in an Accuri Flow Cytometer (Becton Dickinson) with the assistance of the Institute of Genetics and Cancer flow cytometry facility (Edinburgh, United Kingdom).

Plasmid transfection efficiencies were calculated by co-transfecting with pCEP-EGFP into each cell line (Alisch et al. 2006). Briefly, 2×10^4 of HeLa, HEK293T or V79B cells were seeded in 2mL of the corresponding media in 6-well plates. Cells were then transfected 14hr after seeding, as described above except with the addition of 0.5µg pCEP-EGFP (Garcia-Perez et al. 2007; Alisch et al. 2006) alongside 0.5µg of each L1 plasmid to each well. Media were replaced 24hr post-transfection without antibiotic supplementation and analyzed by flow cytometry on day 5 post-transfection. Untransfected cells were used to set the boundary in flow cytometry between $EGFP^-$ and $EGFP^+$ events. Transfection efficiency assays for HeLa cells were performed in technical duplicates and used to normalize colony counts by the corresponding transfection efficiency. For HEK293T cells, transfected cells were selected through supplementing media with puromycin and, therefore, no correction by transfection efficiency was necessary. Thus, transfection efficiency analysis was performed only as a quality check.

For the retrotransposition assays, untransfected HeLa and V79B cells were selected with G418 as a negative control to confirm that neomycin resistant colonies were due to retrotransposition events. Untransfected HEK293T cells were selected with puromycin as a control to ensure the $EGFP^+$ cell percentages for each tested construct were obtained from wells with no untransfected cells. No HeLa, HEK293T or V79B cells survived antibiotic treatment. For the HEK293T cell-based assay, untransfected cells untreated with puromycin were used to set the $EGFP^-$ signal level in flow cytometry.

Bisulfite sequencing methylation analysis

Locus-specific bisulfite sequencing was performed as previously described for individual human L1 copies and protein-coding genes (Sanchez-Luque et al. 2019; Schauer et al. 2018; Nguyen et al. 2018; Salvador-Palomeque et al. 2019). Briefly, this involved first treating genomic DNA with an EZ DNA Methylation Lightning kit (Zymo Research, D5030). Primers were then designed against a CpG island flanking the *PRDM4* transcription start site (5′-TGTTATGAAGATTGAAATTTTGAG-3′ and 5′-CAACCCACCTAACAACTAC-3′) and the L1RS$_{PRDM4}$ 5′ end (5′-TGATAGTAAAGGTTTTGTAGAG-3′ and 5′-ACTACTATAAACTCCACCCAAT-3′). PCR reactions involved MyTaq HS DNA Polymerase (Bioline) and the following cycling conditions for the *PRDM4* assay: (95°C, 2min)×1; (95°C, 30s; 55°C, 30s; 72°C, 30s)×40; (72°C, 5min; 4°C, hold)×1. The same conditions were used for the L1RS$_{PRDM4}$ assay, apart from an annealing temperature of 52°C. Amplicons from each sample were then pooled and prepared for sequencing with a NEBNext Ultra II DNA Library Prep kit (New England Biolabs, E7645S). Paired-end 2×300mer sequencing was performed on a MiSeq platform (Illumina). Paired-end reads were assembled into contigs via FLASH (Magoč and Salzberg 2011) and assessed against target amplicons as described (Sanchez-Luque et al. 2019). Methylation cartoons were then generated for 50 randomly chosen reads for each amplicon and sample via the Quantification tool for Methylation Analysis (QUMA) (Kumaki et al. 2008) with default parameters, plus requiring strict CpG recognition and excluding identical bisulfite sequences.

RNA-seq analyses

To quantify TE subfamily expression during macaque development, we assembled published RNA-seq data generated from single oocytes and preimplantation embryos (Wang et al. 2017) and hippocampus tissue (Yin et al. 2020). Oocyte and embryo sequencing data were obtained from the Sequence Read Archive (SRA) under project identifier SRP089891 and encompassed germinal vesicle (GV) oocyte (n=3), metaphase II stage oocyte (n=3), pronucleus (1-cell embryo), (n=3), 2-cell embryo (n=3), 4-cell embryo (n=2), 8-cell embryo (n=5), morula (n=3) and blastocyst (n=4) stages. Bulk hippocampus tissue RNA-seq data (SRP188855) were included from the dentate gyrus, CA1, and CA3 regions of 8 animals, making a total of 24 samples. For each library, we aligned reads to the rheMac10 genome assembly with STAR (Dobin et al. 2013) version 2.6 (parameters --twopassMode Basic --outSAMprimaryFlag AllBestScore --winAnchorMultimapNmax 1000 --

outFilterMultimapNmax 1000) and marked duplicate reads with Picard MarkDuplicates (http://broadinstitute.github.io/picard). To profile protein-coding gene expression, we considered only uniquely mapped reads overlapping RefSeq exon coordinates and built .wig plots (such as for *PRDM4*) in the Integrative Genomics Viewer (Robinson et al. 2011).

High copy number sequences with limited divergence, such as young TE subfamilies, present a significant mappability issue where genuine signal is lost due to reads mapping to multiple genomic loci (multi-map reads) (Lanciano and Cristofari 2020; Faulkner et al. 2008). We therefore followed an existing strategy to, where possible, assign multi-map reads a weighting at each position based on the relative abundance of uniquely mapping reads nearby (Faulkner et al. 2008, 2009; Hashimoto et al. 2009). Specifically, for each multi-map read we counted the number of uniquely mapped reads within 100bp of the aligned multi-map read at each of its potential best map genomic locations. We then assigned a weighting to each position in proportion to the fraction of uniquely mapped reads found at that position out of the total number of uniquely mapped reads found at any position for the given multi-map read. If no uniquely mapped reads were found at any of the *n* multi-map positions, each position was assigned a weighting of 1/*n*. Uniquely mapped reads were assigned a weighting of 1. To produce estimates of transcript abundance for TE subfamilies, we intersected weighted alignments with RepeatMasker (Smit et al. 1996) coordinates and produced totals for each individual TE, and then summed these to produce a value for each TE subfamily genome-wide. Values were normalized by the total number of weighted mapped reads (tags per million). For display in histograms, L1RS2 was represented by the "L1_RS2" RepeatMasker subfamily, *Alu*YRa1 and L1PA5 were eponymous and MacERV1 was quantified as the sum of "MacERV1_int-int" and "MacERV1_LTR4" values.

As an orthogonal computational approach, we quantified transcript abundance across TE subfamilies using the TEtranscripts package (Jin et al. 2015). Again we mapped the RNA-seq data described above to the rheMac10 reference genome using STAR (Dobin et al. 2013) as recommended (parameters --winAnchorMultimapNmax 100 --outFilterMultimapNmax 100), and marked duplicate reads with Picard MarkDuplicates. TEtranscripts version 2.2.1 was then used to generate read counts for protein-coding genes and repetitive elements, using annotations sourced from the ncbiRefSeq gene model tables of the UCSC Genome Browser (Kent et al. 2002), and a custom GTF file generated using RepeatMasker (Smit et al. 1996) provided with TEtranscripts. The fpm function of DESeq2 (Love et al. 2014) version 1.30.1 was used to normalize read counts (tags-per-million) for display histograms.

As positive controls, we employed the same approaches to analyze RNA-seq data

from prior studies reporting specific expression of MERVL (Macfarlan et al. 2012) and HERVH (Zhang et al. 2019) retrotransposons, respectively, in mouse 2-cell embryo and human embryonic stem cell (hESC) samples. Mouse samples included triplicate 2-cell embryo (SRP009468) and oocyte (SRP009469) experiments. MERVL expression was quantified as the sum of "MERVL-int" and "MT2_Mm" values, representing MERVL and its flanking LTR sequences. Human samples included duplicate hESC cardiomyocyte differentiation time courses (SRP152979), sampled at day 0 (hESC), day 2 (mesoderm), day 5 (cardiac mesoderm), day 7 (cardiac progenitor), day 15 (primitive cardiomyocyte) and day 80 (ventricular cardiomyocyte). HERVH expression was taken as the sum of "HERVH" and "LTR7" values. Reference genome assemblies mm10 and hg38, and their associated genome annotations, were used for mouse and human analyses, respectively.

ONT sequencing and methylation analysis

High molecular weight DNA was extracted from animal ON22213 hippocampus and liver tissue using a Nanobind CBB Big DNA Kit (Circulomics, NB-900-001-01) and sheared to a ~10kb average size to improve sequencing yield. ONT sequencing libraries were prepared for each DNA sample using a Ligation Sequencing Kit (Oxford Nanopore Technologies, SQK-LSK109), pooled, and sequenced on an ONT PromethION platform (Kinghorn Centre for Clinical Genomics, Australia). Bases were called with Guppy 4.0.11 (Oxford Nanopore Technologies) and reads aligned to the rheMac10 reference genome build using minimap2 version 2.20 (Li 2018) and SAMtools version 1.12 (Li et al. 2009). Reads were indexed and per-CpG methylation calls generated using nanopolish version 0.13.2 (Simpson et al. 2017). Methylation likelihood data were sorted by position and indexed using tabix version 1.12 (Li 2011). Methylation statistics for the genome divided into 6kbp bins, and reference TEs defined by RepeatMasker coordinates (http://www.repeatmasker.org/), were generated using MethylArtist version 1.0.4 (Cheetham et al. 2022), using commands db-nanopolish, segmeth and segplot with default parameters. Methylation profiles for individual loci were generated using the MethylArtist command locus, where parameters specified a 30bp sliding window with a 2bp step, and smoothed with a window size of 8 for the Hann function. The L1RS subfamily methylation profiles shown in **Fig. 4C** were generated for elements >6kbp with the MethylArtist composite command. To identify individual TEs exhibiting differential methylation in the comparison of ON22213 hippocampus and liver ONT data (**Supplemental Table S3**), we required elements to have at least 4 reads and 20 methylation calls in each sample. Comparisons were carried out via Fisher's exact test using methylated and non-

methylated call counts, with significance defined as a Bonferroni corrected p value of less than 0.05. The significance of observed versus expected intronic L1 insertions was calculated with a binomial test.

Data access

The Illumina and ONT sequencing data generated in this study are available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under BioProject accession number PRJEB37719. Sanger trace files and unprocessed gel images for this study can be found in **Supplemental File S1** and at Mendeley Data (https://dx.doi.org/10.17632/wpnv9ktv7p.2).

Competing interest statement

The authors declare that they have no competing interests.

Acknowledgements

The authors thank John V. Moran for sharing L1.3 plasmids and the HeLa-JVM cell line, Margaret Z. Zdzienicka for sharing the V79B cell line, Jeffrey A. Jeddeloh for assistance with RC-seq probe design, and the QBI, TRI and IGC flow cytometry facilities for technical advice. This study was funded by: Australian NHMRC Investigator Grants (GNT1161832 to S.W.C., GNT1176574 to N.J., GNT1173476 to S.R.R., GNT1173711 to G.J.F.), an NHMRC-ARC Dementia Research Development Fellowship (GNT1108258 to G.O.B.), an Australian Government Research Training Program Scholarship awarded to P.G., the Australian Department of Health Medical Frontiers Future Fund (MRFF) (MRF1175457 to A.D.E.), the Australian Research Council (DP200102919 to S.R.R. and G.J.F.), MINECO-FEDER (SAF2017-89745-R) and European Research Council (ERC-STG-2012-309433) funding and a private donation from Ms Francisca Serrano (Trading y Bolsa para Torpes, Granada, Spain) to J.L.G-P., a National Institute of Health (NIH) Office of Directors P51 Grant (OD011092) to the Oregon National Primate Research Center to support L.C., an Andalusian Government EMERGIA grant (20_00225) to F.J.S-L., a CSL Centenary Fellowship to G.J.F., and the Mater Foundation. Rhesus macaque tissues were obtained from the Monkey Alcohol Tissue Research Resource (MATRR) biobank, supported by NIH Grant 2R24 AA019431.

Authors' contributions

V.B., F.J.S-L., J.R., and G.O.B. are equal co-authors. V.B., F.J.S-L., J.R., G.O.B., D.J.G., P.G., S.N.S., P.A., C.E.L. and K.A.N. performed experiments. V.B., F.J.S-L., J.R., S.W.C., T.J.M., N.J., S.R.R., A.D.E. and G.J.F. analyzed the data. G.O.B., J.L.G-P., S.R.R., L.C. and G.J.F. provided resources. V.B., F.J.S-L., A.D.E. and G.J.F. prepared figures. V.B., L.C. and G.J.F. conceived and designed the project. G.J.F. wrote the manuscript. All authors read and approved the final manuscript.

References

Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, et al. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**: 405–410.

Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. 2006. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* **20**: 210–224.

Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**: 1067–1081.e17.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–537.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159–1170.

Bodea GO, Ferreiro ME, Sanchez-Luque FJ, Botto JM, Rasmussen J, Rahman MA, Fenlon LR, Gubert C, Gerdes P, Bodea L-G, et al. 2022. LINE-1 retrotransposon activation intrinsic to interneuron development. *bioRxiv*. doi: 10.1101/2022.03.20.485017.

Bogani D, Morgan MAJ, Nelson AC, Costello I, McGouran JF, Kessler BM, Robertson EJ, Bikoff EK. 2013. The PR/SET domain zinc finger protein Prdm4 regulates gene expression in embryonic stem cells but plays a nonessential role in the developing mouse embryo. *Mol Cell Biol* **33**: 3936–3950.

Boissinot S, Furano AV. 2001. Adaptive evolution in LINE-1 retrotransposons. *Mol Biol Evol* **18**: 2186–2194.

Brouha B, Meischl C, Ostertag E, de Boer M, Zhang Y, Neijens H, Roos D, Kazazian HH Jr. 2002. Evidence consistent with human L1 retrotransposition in maternal meiosis I. *Am J Hum Genet* **71**: 327–336.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**: 5280–5285.

Carreira PE, Ewing AD, Li G, Schauer SN, Upton KR, Fagg AC, Morell S, Kindlova M,

Gerdes P, Richardson SR, et al. 2016. Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mob DNA* **7**: 21.

Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, Duc J, Jang SM, Turelli P, Trono D. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev* **28**: 1397–1409.

Cheetham SW, Kindlova M, Ewing AD. 2022. Methylartist: Tools for Visualising Modified Bases from Nanopore Sequence Data. *Bioinformatics* **38**: 3109–3112.

Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, Xie XS. 2017. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**: 189–194.

Chittka A, Nitarska J, Grazini U, Richardson WD. 2012. Transcription factor positive regulatory domain 4 (PRDM4) recruits protein arginine methyltransferase 5 (PRMT5) to mediate histone arginine methylation and control neural stem cell proliferation and differentiation. *J Biol Chem* **287**: 42995–43006.

Chronister WD, Burbulis IE, Wierman MB, Wolpert MJ, Haakenson MF, Smith ACB, Kleinman JE, Hyde TM, Weinberger DR, Bekiranov S, et al. 2019. Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep* **26**: 825–835.e7.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**: 1127–1131.

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

Daunais JB, Davenport AT, Helms CM, Gonzales SW, Hemby SE, Friedman DP, Farro JP, Baker EJ, Grant KA. 2014. Monkey alcohol tissue research resource: banking tissues for alcohol research. *Alcohol Clin Exp Res* **38**: 1973–1981.

de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, Warris A, Roos D. 2014. Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. *Hum Mutat* **35**: 486–496.

de la Rica L, Deniz Ö, Cheng KCL, Todd CD, Cruz C, Houseley J, Branco MR. 2016. TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells. *Genome Biol* **17**: 234.

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. *Nat Rev Genet* **20**: 417–431.

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

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

Dombroski BA, Scott AF, Kazazian HH Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc Natl Acad Sci U S A* **90**: 6513–6517.

Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. 2015. A 3′ Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell* **60**: 728–741.

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15**: 497–506.

Erwin JA, Paquola ACM, Singer T, Gallina I, Novotny M, Quayle C, Bedrosian TA, Alves FIA, Butcher CR, Herdy JR, et al. 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* **19**: 1583–1591.

Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, Yang L, Haseley P, Lehmann HS, Park PJ, et al. 2015. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**: 49–59.

Ewing AD, Kazazian HH Jr. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**: 1262–1270.

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell* **80**: 915–928.e5.

Faulkner GJ, Billon V. 2018. L1 retrotransposition in the soma: a field jumping ahead. *Mob DNA* **9**: 22.

Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM. 2008. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**: 281–288.

Faulkner GJ, Garcia-Perez JL. 2017. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends Genet* **33**: 802–816.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563–571.

Feng G, Jensen FE, Greely HT, Okano H, Treue S, Roberts AC, Fox JG, Caddick S, Poo M-M, Newsome WT, et al. 2020. Opportunities and limitations of genetically modified nonhuman primate models for neuroscience research. *Proc Natl Acad Sci U S A* **117**: 24022–24031.

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–916.

Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**: 1567–1577.

Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, Wilson TE,

Moran JV. 2019. Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* **177**: 837–851.

Furano AV, Jones CE, Periwal V, Callahan KE, Walser J-C, Cook PR. 2020. Cryptic genetic variation enhances primate L1 retrotransposon survival by enlarging the functional coiled coil sequence space of ORF1p. *PLoS Genet* **16**: e1008991.

Garcia-Perez JL, Marchetto MCN, Muotri AR, Coufal NG, Gage FH, O'Shea KS, Moran JV. 2007. LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* **16**: 1569–1577.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 Genomes Project Consortium, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**: 1916–1929.

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.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA* **7**: 16.

Grandi FC, Rosser JM, An W. 2013. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol Biol Evol* **30**: 503–512.

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**: 590–607.

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**: 221–225.

Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, Huang CT, Sandifer E, Hebert K, Barnes EW, et al. 2007. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. *Science* **316**: 238–240.

Hashimoto T, de Hoon MJL, Grimmond SM, Daub CO, Hayashizaki Y, Faulkner GJ. 2009. Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* **25**: 2613–2614.

Hazen JL, Faust GG, Rodriguez AR, Ferguson WC, Shumilina S, Clark RA, Boland MJ, Martin G, Chubukov P, Tsunemoto RK, et al. 2016. The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. *Neuron* **89**: 1223–1236.

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

Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**: 3593–3599.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94**: 1872–1877.

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev* **23**: 1303–1312.

Kazazian HH Jr, Moran JV. 2017. Mobile DNA in Health and Disease. *N Engl J Med* **377**: 361–370.

Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164–166.

Keegan RM, Talbot LR, Chang Y-H, Metzger MJ, Dubnau J. 2021. Intercellular viral spread and intracellular transposition of Drosophila gypsy. *PLoS Genet* **17**: e1009535.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16**: 78–87.

Khazina E, Weichenrieder O. 2018. Human LINE-1 retrotransposition requires a metastable coiled coil and a positively charged N-terminus in L1ORF1p. *Elife* **7**: e34960.

King DA, Jones WD, Crow YJ, Dominiczak AF, Foster NA, Gaunt TR, Harris J, Hellens SW, Homfray T, Innes J, et al. 2015. Mosaic structural variation in children with developmental disorders. *Hum Mol Genet* **24**: 2733–2745.

Klawitter S, Fuchs NV, Upton KR, Muñoz-Lopez M, Shukla R, Wang J, Garcia-Cañadas M, Lopez-Ruiz C, Gerhardt DJ, Sebe A, et al. 2016. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat Commun* **7**: e10286.

Kopera HC, Larson PA, Moldovan JB, Richardson SR, Liu Y, Moran JV. 2016. LINE-1 cultured cell retrotransposition assay. *Methods Mol Biol* **1400**: 139–156.

Kumaki Y, Oda M, Okano M. 2008. QUMA: quantification tool for methylation analysis. *Nucleic Acids Res* **36**: W170–5.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**: 1812–1819.

Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. *Nat Rev Genet* **21**: 721–736.

Lavie L, Maldener E, Brouha B, Meese EU, Mayer J. 2004. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* **14**: 2253–2260.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-

MEM. *arXiv [q-bioGN]* arXiv:1303.3997.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**: 718–719.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lim JS, Kim W-I, Kang H-C, Kim SH, Park AH, Park EK, Cho Y-W, Kim S, Kim HM, Kim JA, et al. 2015. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med* **21**: 395–400.

Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE. 2009. Comparative analysis of Alu repeats in primate genomes. *Genome Res* **19**: 876–885.

Li W, Prazak L, Chatterjee N, Grüninger S, Krug L, Theodorou D, Dubnau J. 2013. Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat Neurosci* **16**: 529–531.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lutz SM, Vincent BJ, Kazazian HH Jr, Batzer MA, Moran JV. 2003. Allelic heterogeneity in LINE-1 retrotransposition activity. *Am J Hum Genet* **73**: 1431–1437.

Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21**: 423–425.

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**: 57–63.

Macia A, Widmann TJ, Heras SR, Ayllon V, Sanchez L, Benkaddour-Boumzaouad M, Muñoz-Lopez M, Rubio A, Amador-Cubero S, Blanco-Jimenez E, et al. 2017. Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Res* **27**: 335–348.

Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken RS, Vermeesch JR, Hall IM, et al. 2013. Mosaic copy number variation in human neurons. *Science* **342**: 632–637.

McConnell MJ, Moran JV, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, Erwin JA, Fasching L, Flasch DA, Freed D, et al. 2017. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**:

eaal1641.

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**: 643–645.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917–927.

Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV. 2007. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**: 208–212.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903–910.

Muotri AR, Gage FH. 2006. Generation of neuronal variability and complexity. *Nature* **441**: 1087–1093.

Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH. 2010. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**: 443–446.

Nakashima M, Saitsu H, Takei N, Tohyama J, Kato M, Kitaura H, Shiina M, Shirozu H, Masuda H, Watanabe K, et al. 2015. Somatic Mutations in the MTOR gene cause focal cortical dysplasia type IIb. *Ann Neurol* **78**: 375–386.

Nguyen THM, Carreira PE, Sanchez-Luque FJ, Schauer SN, Fagg AC, Richardson SR, Davies CM, Jesuadian JS, Kempen M-JHC, Troskie R-L, et al. 2018. L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution. *Cell Rep* **23**: 3730–3740.

Ostertag EM, Kazazian HH Jr. 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**: 2059–2065.

Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH Jr. 2000. Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res* **28**: 1418–1423.

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7**: 597–606.

Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G. 2016. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife* **5**: e13926.

Phillips KA, Bales KL, Capitanio JP, Conley A, Czoty PW, 't Hart BA, Hopkins WD, Hu S-L, Miller LA, Nader MA, et al. 2014. Why primate models matter. *Am J Primatol* **76**: 801–827.

Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea G-O, Muñoz-Lopez M, Jesuadian JS, Kempen M-JHC, Carreira PE, Jeddeloh JA, et al. 2017. Heritable L1

retrotransposition in the mouse primordial germline and early embryo. *Genome Res* **27**: 1395–1405.

Robbez-Masson L, Tie CHC, Conde L, Tunbak H, Husovsky C, Tchasovnikarova IA, Timms RT, Herrero J, Lehner PJ, Rowe HM. 2018. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. *Genome Res* **28**: 836–845.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Salvador-Palomeque C, Sanchez-Luque FJ, Fortuna PRJ, Ewing AD, Wolvetang EJ, Richardson SR, Faulkner GJ. 2019. Dynamic Methylation of an L1 Transduction Family during Reprogramming and Neurodifferentiation. *Mol Cell Biol* **39**: e00499.

Sanchez-Luque FJ, Kempen M-JHC, Gerdes P, Vargas-Landin DB, Richardson SR, Troskie R-L, Jesuadian JS, Cheetham SW, Carreira PE, Salvador-Palomeque C, et al. 2019. LINE-1 Evasion of Epigenetic Repression in Humans. *Mol Cell* **75**: 590–604.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**: 424–436.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr. 1997. Many human L1 elements are capable of retrotransposition. *Nat Genet* **16**: 37–43.

Schauer SN, Carreira PE, Shukla R, Gerhardt DJ, Gerdes P, Sanchez-Luque FJ, Nicoli P, Kindlova M, Ghisletti S, Santos AD, et al. 2018. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Res* **28**: 639–653.

Scott EC, Devine SE. 2017. The Role of Somatic L1 Retrotransposition in Human Cancers. *Viruses* **9**: 131.

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* **26**: 745–755.

Seleme M del C, Vetter MR, Cordaux R, Bastone L, Batzer MA, Kazazian HH Jr. 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc Natl Acad Sci U S A* **103**: 6611–6616.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.

Siudeja K, van den Beek M, Riddiford N, Boumard B, Wurmser A, Stefanutti M, Lameiras S, Bardin AJ. 2021. Unraveling the features of somatic transposition in the Drosophila intestine. *EMBO J* **40**: e106388.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**: 657–663.

Smit A, Hubley R, Green P, Smit HA. 1996. RepeatMasker Open-3.0.

https://www.scienceopen.com/document?vid=4d11d946-bf9a-4fca-aee8-1153c753b386.

Smits N, Rasmussen J, Bodea GO, Amarilla AA, Gerdes P, Sanchez-Luque FJ, Ajjikuttira P, Modhiran N, Liang B, Faivre J, et al. 2021. No evidence of human genome integration of SARS-CoV-2 found by long-read DNA sequencing. *Cell Rep* **36**: 109530.

Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C, Ryder OA, Janečka JE, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLoS One* **7**: e49521.

Stiles J, Jernigan TL. 2010. The basics of brain development. *Neuropsychol Rev* **20**: 327–348.

Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L, Nigumann P, Saccani S, Andrau J-C, et al. 2019. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol Cell* **74**: 555–570.

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol* **269**: 276–285.

Tang W, Liang P. 2019. Comparative Genomics Analysis Reveals High Levels of Differential Retrotransposition among Primates from the Hominidae and the Cercopithecidae Families. *Genome Biol Evol* **11**: 3309–3325.

Taylor MS, LaCava J, Mita P, Molloy KR, Huang CRL, Li D, Adney EM, Jiang H, Burns KH, Chait BT, et al. 2013. Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155**: 1034–1048.

Thayer RE, Singer MF, Fanning TG. 1993. Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. *Gene* **133**: 273–277.

Treiber CD, Waddell S. 2017. Resolving the prevalence of somatic transposition in Drosophila. *Elife* **6**: e28297.

Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, et al. 2014. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**: 1251343.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**: e115.

van den Hurk JAJM, Meij IC, Seleme M del C, Kano H, Nikopoulos K, Hoefsloot LH, Sistermans EA, de Wijs IJ, Mukhopadhyay A, Plomp AS, et al. 2007. L1 retrotransposition can occur early in human embryonic development. *Hum Mol Genet* **16**: 1587–1592.

Wagstaff BJ, Barnerssoi M, Roy-Engel AM. 2011. Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. *PLoS One* **6**: e19672.

Wagstaff BJ, Kroutter EN, Derbes RS, Belancio VP, Roy-Engel AM. 2013. Molecular reconstruction of extinct LINE-1 elements and their interaction with nonautonomous elements. *Mol Biol Evol* **30**: 88–99.

Wang X, Liu D, He D, Suo S, Xia X, He X, Han J-DJ, Zheng P. 2017. Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for DNA double-strand break repair in primate oocytes and early embryos. *Genome Res* **27**: 567–579.

Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LW, et al. 2020. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**: eabc6617.

Xing D, Tan L, Chang C-H, Li H, Xie XS. 2021. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc Natl Acad Sci U S A* **118**: e2013106118.

Yin S, Lu K, Tan T, Tang J, Wei J, Liu X, Hu X, Wan H, Huang W, Fan Y, et al. 2020. Transcriptomic and open chromatin atlas of high-resolution anatomical regions in the rhesus macaque brain. *Nat Commun* **11**: 474.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**: 1380–1388.

Zhao Y, Ji S, Wang J, Huang J, Zheng P. 2014. mRNA-Seq and microRNA-Seq whole-transcriptome analyses of rhesus monkey embryonic stem cell neural differentiation revealed the potential regulators of rosette neural stem cells. *DNA Res* **21**: 541–554.

Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. 2020. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res* **48**: 1146–1163.

Figure legends

Figure 1. Characterization of germline and somatic macaque TE insertions.

(*A*) Genomics experimental design. Individual hippocampal neuron (RBFOX3^{+}) nuclei from two rhesus macaques (ON22212 and ON22213) were subjected to whole genome amplification (WGA), followed by Illumina scWGS and RC-seq, to identify somatic TE insertions. Bulk liver DNA was analyzed with Illumina WGS to discriminate germline and somatic variants.

(*B*) Percentages of exonic, intronic and intergenic non-reference L1 (top left) and *Alu* (top right) insertions. Genomic features were annotated according to RefSeq coordinates, with the underlying proportions of each feature (random expectation) shown at bottom.

(*C*) Target site duplication (TSD) size distributions for non-reference L1 (left) and *Alu* (right) insertions, as annotated by TEBreak. Inset sequence logos (Crooks et al. 2004) display the observed integration site nucleotide composition for each TE family. These resembled the L1 endonuclease motif.

(*D*) A somatic L1RS2 insertion ($L1RS_{somatic}$) was detected on Chromosome 4 of animal ON22213 hippocampal neuron #15. Reads spanning the 5′ or 3′ L1-genome junctions of this event are shown, as is the corresponding TSD.

(*E*) PCR validation of $L1RS_{somatic}$. Primer (symbols α, ε, δ, γ, β and Φ) positions relative to the L1 insertion are indicated in the schematic provided at top. The 5′ L1-genome junction was amplified by combining primers α and γ, whereas nested PCR (ε+Φ then δ+β) was employed to amplify the 3′ L1-genome junction. Reaction input in each case consisted of non-template control (NTC), 13 ON22213 hippocampal neurons analyzed with scWGS and RC-seq, bulk ON22213 hippocampus and liver DNA, and bulk ON22212 liver. Red arrowheads and crosses indicate amplicons confirmed as on-target and off-target, respectively, by capillary sequencing. Numbers next to confirmed 3′ L1-genome junction bands indicate the L1 poly(A) tract length for that amplicon.

(*F*) Complete sequence characterization of $L1RS_{somatic}$. TSD nucleotides are highlighted in red. The intergenic L1 was full-length (L1RS2 subfamily consensus start position 0), carried a 4bp 5′ transduction (pink rectangle) with an untemplated guanine (underlined G), and was followed by a long, pure 3′ poly(A) tract. The transduction indicated a putative donor L1 intronic to the *PRDM4* gene on Chromosome 11 ($L1RS_{PRDM4}$).

Figure 2. An endogenous L1 mobile in the macaque brain and *in vitro*.

(*A*) The complete sequence of $L1RS_{somatic}$ and its homozygous donor element, $L1RS_{PRDM4}$, were amplified by PCR reactions (primers α+β) with input template DNA from ON22213 neuron #15 and bulk liver, respectively. Note: primer α spanned the 5′ junction of $L1RS_{somatic}$ to more efficiently amplify the L1 allele. Red arrowheads indicate amplicons confirmed as on-target by capillary sequencing.

(*B*) $L1RS_{somatic}$ and $L1RS_{PRDM4}$ were cloned and completely capillary sequenced. Nucleotide variants amongst the reference (REF) genome $L1RS_{PRDM4}$ sequence, the two identical $L1RS_{PRDM4}$ alleles carried by animal ON22213, and the $L1RS_{somatic}$ sequence are shown. Non-synonymous mutations are highlighted in red. The 4bp 5′ transduction (AGAG) carried by $L1RS_{somatic}$ is colored in pink.

(*C*) Engineered L1 retrotransposition efficiency measured in cultured HeLa cells (Moran et al. 1996). The assay design (top) shows either $L1RS_{PRDM4}$ (brown) or L1.3 (purple), a highly mobile human L1 (Dombroski et al. 1993), tagged with a neomycin (G418) resistance cassette activated only upon retrotransposition (S, seeding; T, transfection; M, change of media; R, result analysis; filled lollipop, polyadenylation signal; numbers represent days of treatment with G418). AA(T)AAA indicates where a thymine base was removed to ablate the natural $L1RS_{PRDM4}$ and L1.3 polyadenylation signals. Tested elements (bottom) included, in order, positive (L1.3) and negative (L1.3 RT^-, D702A mutant) controls (Moran et al. 1996; Sassaman et al. 1997), $L1RS_{PRDM4}$, a set of 3 chimeric elements where L1.3 was fused to $L1RS_{PRDM4}$ at the 3′ end of the L1.3 5′UTR, ORF1 and ORF2, and a set of 3 reciprocal elements where $L1RS_{PRDM4}$ and L1.3 were joined at the 3′ end of the $L1RS_{PRDM4}$ 5′UTR, ORF1 and ORF2 sequences. L1 expression was driven by native promoters only. Chimeric element fusion points are marked by inverted triangles. Representative well pictures are shown. Histogram values are normalized to L1.3 (100%). Data consist of three technical replicates, and their mean ± SD, obtained from one representative experiment of three independent biological replicates.

(*D*) As per (*C*), except assayed in HEK293T cells using an EGFP-based L1 reporter system (Ostertag et al., 2000) where cells are selected for puromycin resistance and retrotransposition efficiency is measured as the percentage of $EGFP^+$ sorted cells.

(*E*) As per (*C*), except with the inclusion of a cytomegalovirus promoter (CMVp) to additionally drive $L1RS_{PRDM4}$ and L1.3 expression, and tested in the Chinese hamster fibroblast V79B cell line.

Figure 3. Regulation and embryonic expression of the *PRDM4* locus.

(*A*) Methylation profile of the *PRDM4* locus obtained from ONT long-read sequencing (Ewing et al. 2020; Cheetham et al. 2022). The first panel shows $L1RS_{PRDM4}$ orientated in sense to intron 10 of *PRDM4*, with genomic coordinates (rheMac10) provided, as well as a magnified view of the $L1RS_{PRDM4}$ 5′UTR displaying CpG dinucleotides (orange lines) forming a CpG island (pink bar). The positions of primers used to assess $L1RS_{PRDM4}$ methylation via locus-specific bisulfite sequencing in panel (*C*) are shown. The second panel displays animal ON22213 ONT read alignments, with unmethylated CpGs colored in blue (hippocampus) and orange (liver), and methylated CpGs colored black. The third panel indicates the relationship between CpG positions in genome space and CpG space, including those corresponding to the *PRDM4* CpG island (shaded light green) and the $L1RS_{PRDM4}$ 5′UTR and body (shaded light and dark brown, respectively). The fourth panel indicates the fraction of methylated CpGs for each tissue across CpG space.

(*B*) Targeted bisulfite sequencing of the *PRDM4* CpG island, as indicated in panel (*A*), in animal ON22213 hippocampus and liver tissue. Each cartoon displays 50 non-identical randomly selected sequences, where methylated CpGs (mCpGs) and unmethylated CpGs are represented by black and white circles, respectively, as well as the overall mCpG percentage.

(*C*) As per (*B*), except for the $L1RS_{PRDM4}$ 5′UTR CpG island.

(*D*) *PRDM4* expression (blue circles) measured in RNA-seq tags-per-million (TPM), compared to that of the housekeeping gene *ACTB* (purple squares). Data were obtained from prior analyses of germinal vesicle (GV) and metaphase II (MII) oocytes, pre-implantation embryo development stages (Wang et al. 2017) and adult hippocampus (Yin et al. 2020). Horizontal bars represent the mean of biological replicates.

(*E*) Examples of *PRDM4* expression during rhesus macaque development, showing .wig coverage tracks generated from published 8-cell embryo, neural stem cell and hippocampus RNA-seq datasets (Wang et al. 2017; Yin et al. 2020; Zhao et al. 2014).

Figure 4. Genome-wide analyses of young TE subfamily transcription and methylation.

(*A*) Subfamily-wide TE expression measured by RNA-seq (TPM, tags-per-million) and an existing strategy to account for multi-mapping reads (Faulkner et al. 2008; Hashimoto et al. 2009; Faulkner et al. 2009). Data were obtained from prior studies (Wang et al. 2017; Yin et al. 2020) and encompassed germinal vesicle (GV) and metaphase II (MII) oocytes, early embryonic development, and adult hippocampus. Horizontal bars represent the mean of biological replicates.

(*B*) Violin plots showing CpG methylation ascertained by ONT sequencing upon animal ON22213 hippocampus and liver. Results are shown for the whole genome (6kbp windows), L1RS2 and L1PA5 copies >6kbp, *Alu*YRa1 copies >300bp and MacERV1 long terminal repeats >300bp.

(*C*) Composite L1RS subfamily methylation profiles. Each graph displays 100 profiles. A schematic of the L1RS2 consensus sequence is provided at top, with CpG positions indicated by pink bars.

(*D*) Exemplar methylation profile of an L1RS2 element located on Chromosome 5 and hypomethylated in the liver. The panel is composed as described for Fig. 3A.